Chapter 1

Introduction[1]

1.1 Introduction

This chapter discusses important background issues of the thesis, such as the problem statement, research questions and hypothesis. The setting of Linguistic Information Systems[2] within the philosophy of science is also highlighted. The relation of this project to other Biblical information systems is also touched upon. The history of the creation of the thesis is discussed because it is important to understand the unique structure and set-up, which does not conform to the traditional thesis format. Following the discussion of the research approach, methodology and plan, an outline of the various chapters is given. The contribution to the field of Information and Communication Technologies (ICT), which the candidate hopes to make, is indicated. Finally, a short list of working definitions of technical terms is provided.

1.2 Background: forty years of Biblical Hebrew information systems

To write a complete historical overview of the development of Biblical information systems in a section of an introductory chapter is an impossible task. Twenty years ago a whole book was already written on the early years of this new discipline (see Hughes, 1987). More recent updates are available in Poswick (2004) and Tov (2003 & 2006). The author will, therefore, rather highlight the major trends and research topics that received attention in the past. Deficiencies in existing products will also be

[1] Sections on the philosophical nature and place of Linguistic Information Systems have been published as an article ("Linguistic information [systems] - a humanistic endeavour") in *Innovate* (see Kroeze, 2007a).

[2] When referring to software, *information system(s)* is spelled with small letters. The name of the scientific discipline of *Information Systems* (or *Informatics* as the discipline is often called in South Africa) is spelled with initial capital letters.

indicated because they reveal the opportunities for future research that could build on the amazing body of work that has already been done.

1.2.1 Levels of analysis

Biblical Hebrew grammar can be and has been studied from many different angles. Over the past forty years much of this knowledge has been captured in various computer software systems and databases. The most basic levels are the digital representation of the text in Hebrew characters and the transliteration level, which is an exact representation in the Roman alphabet, if needed. The transliteration may be used to rebuild the text in the Hebrew alphabet, while a separate phonological transcription could embody the pronunciation of the Hebrew text, but the letters or signs used do not correspond exactly to the Hebrew spelling. These basic levels are followed by the morphological, morpho-syntactic and syntactic levels. The more advanced linguistic modules, such as the semantic and pragmatic levels, have received less attention, and one can only hope that knowledge databases and expert systems that deal with these levels will become more readily available. Bothma (1992a) indicates various levels of grammatical information that should, ideally, be available in Biblical information systems, i.e. "[d]escriptions of phonetic, morphological, morphosyntactic, syntactic and semantic phenomena".

There are currently fourteen software tools containing the Codex Leningradensis version of the Hebrew Bible and one tool containing the Aleppo Codex (Tov, 2006: 343). Eleven of these tools offer morphological analyses (Tov, 2006: 356). Only a select few contain syntactic data, for example the database of the Werkgroep Informatica of the Free University in Amsterdam (WIVU). In addition to these grammatically oriented software, ancient and modern translations are available, as well as critical apparatuses and other tools such as dictionaries and even one reference grammar (Tov, 2006).

According to Tov (2006: 346) "[t]he most sophisticated programs ... allow for the search of morphological features ... and also the search for combinations of lexical and grammatical information". Five main groups of morphological analysis exist (these analyses are subjective and do not agree entirely), i.e.:

- Westminster Hebrew Old Testament Morphology (Groves-Wheeler)
- Werkgroep Informatica (Talstra)
- Bar-Ilan analysis
- Academy of the Hebrew Language
- Additional commercial and private morphological analyses (ibid.)

The latest version of the Hebrew Old Testament linguistic database, developed over the past three decades by the Werkgroep Informatica at the Free University in Amsterdam (WIVU), has been included in the Stuttgart Electronic Study Bible (SESB) which has been published on the Libronix/Logos platform (Deutsche Bibelgesellschaft: www.SESB-ONLINE.com, 2001; Stuttgart Electronic Study Bible (SESB) Version 2.0 Logos, 2002-2008). This tool allows researchers to perform advanced syntactic queries, for example, to find examples of clauses having a conjunction and proper name as subject preceding an imperfect verb (Talstra, 2007: 93). According to Talstra (2007:96), "[t]he search for syntactic data offers one way to get a better handle on the function of ordinary and extraordinary constructions in a literary composition" in order to discover grammatical tensions purposefully built into a literary text. The search engine operates mainly on formalistic characteristics and the researcher needs to "translate"/break down his/her query into these terms (cf. Talstra, 2007: 91, 93, 95), but the SESB provides a user interface allowing users to use buttons and checkboxes to select various "combinations of syntactic or grammatical features and functional categories", either on word, phrase or clause level (Kummerow, 2005: 2-3). Although Gómez (2004) regards the search function as "one of the jewels of the crown" of the SESB, and although plain morphological searches are simple to execute, "there is a steep learning curve to overcome" for more complex queries. The search engine was developed in cooperation with prof. C. Hardmeier of the University of Greifswald and prof A. Groves of Philadelphia, Texas (Ernst-Moritz-Arndt-Universität Greifswald, Forschungsschwerpunkt, Computergestützte Philologie und Bild-Erschliessung, 2001). The WIVU-team is currently working on a more advanced version of the database that captures pragmatic information. The researchers who are involved in this project claim that the tool will be "a powerful linguistic instrument for the research into the languages of the Old Testament" (www.th.vu.nl/~wiweb/const/quest2.htm). The ability of the tool to produce phrase parsing, clause level parsing and clause hierarchies distinguishes it from other Biblical Hebrew information systems. There is, however, not a clear distinction between the various linguistic modules in the tagging system.

Another project that offers analysis on an advanced syntactic level is the private database developed by Andersen & Forbes. Their project analyses the elements of each clause up to the most atomic elements (morphemes). Syntactic information and structures are presented as horisontal trees. Andersen & Forbes (2003) trust that their proposal will make a contribution to the field of Biblical Hebrew linguistic information systems by moving beyond the limits of single clauses. Although their work is without any doubt very useful for the study of syntax, it neither differentiates clearly between the various linguistic modules nor does it facilitate multi-modular linguistic studies. In addition, one also needs extensive knowledge of the symbols used to make sense of the myriad of labels used to tag the nodes and leaves of their representations. Their tagging of semantics is limited to word level.

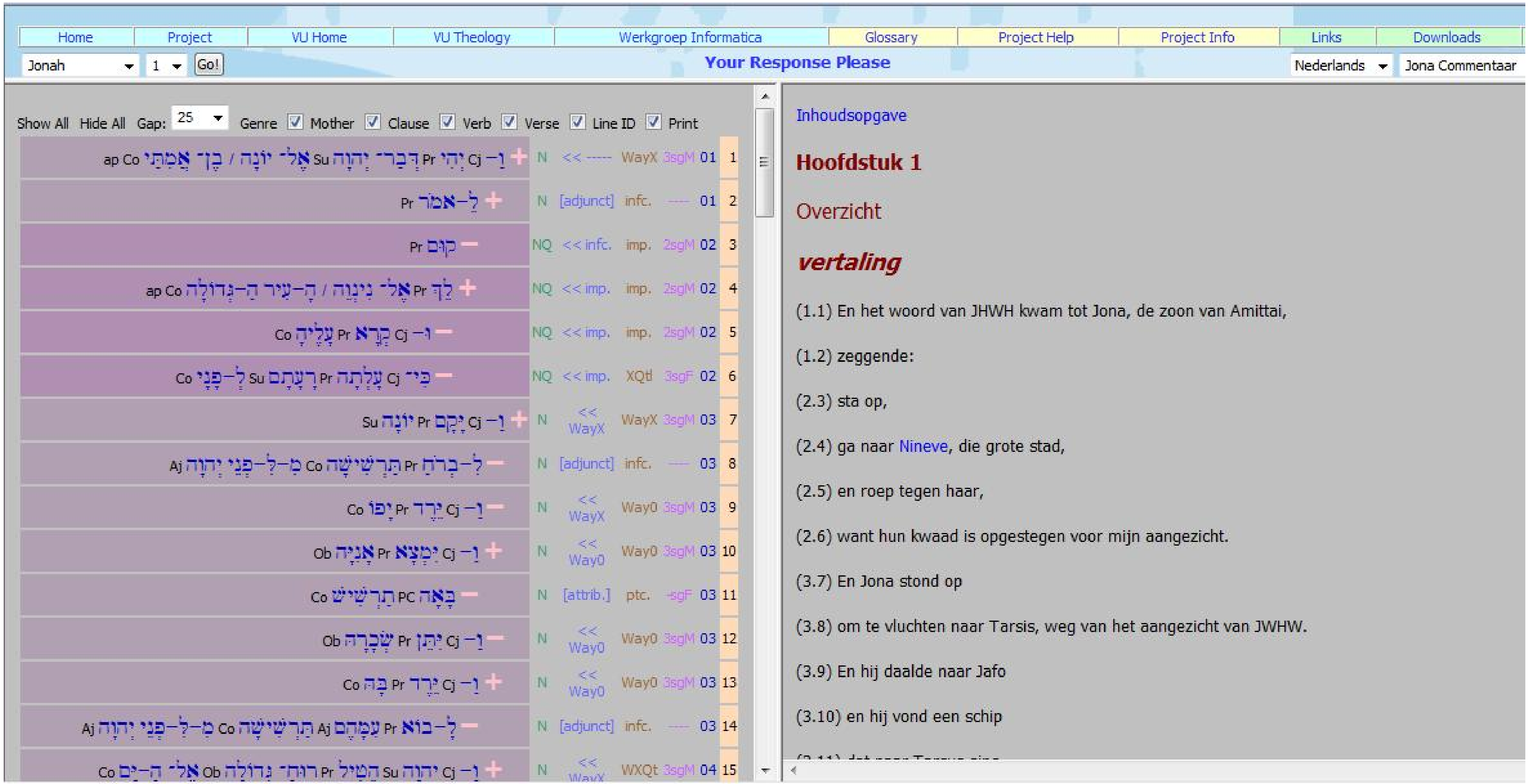
1.2.2 Underutilization of existing tools

Poswick (2004) gives an overview of Biblical information system projects of the period 1985-2004. His impression is that, although various tools provide morphological analyses and even other levels of analysis, "classical Biblical exegesis would not appear to be benefiting as yet from the results of this type of analysis". Tov (2006: 337) agrees with Poswick that Biblical scholars still do not make optimal use of these tools. This may be due to the fact that scholars have been so focused on creating and improving the tools themselves that they have not yet maximised in-depth exploration of the huge amounts of data that have been made available by these tools (Tov, 2006: 338).

However, one has to note that many exegetical articles have been produced as a result of the Werkgroep Informatica's databases (cf. the (outdated, incomplete) bibliography available at http://www.th.vu.nl/~wiweb/cons/publicaitions.htm). A recent example of the use of their database of syntactic hierarchies may be found in Talstra (2006: 231-232) to investigate the use of *yiqtol* verbs in narrative prose found in Exodus, shedding new light on the exegesis of sentences where these verbal forms follow *wayyiqtol* forms. In her review of the SESB, which includes this database, Conybeare (2005) states that "[t]he student who was most excited by the possibilities of the SESB was the one most closely engaged already with biblical exegesis". In order to fully exploit this tool, a user would probably have to make a careful study of the "fascinating essay" in the manual that explains "how the Hebrew text was analyzed to facilitate more complex syntactical searches" (ibid.).

A feeling of information overload may be another reason for the underutilization of Biblical information systems. Claassen & Bothma (1988: 83) highlight the problem of (electronic) information overload that already existed in Biblical research twenty years ago. According to Bothma (1992b) hypermedia may be used to minimise problems of information overload because the network of hyperlinks allows the user to access only relevant information.

A third reason for the underutilization of these systems may be the lack of ease of use. According to Tov (2006:338) Biblical research software is under-used because "[t]here remains a wide gap between the knowledge of the experts creating the tools and that of the scholars for whom the tools are intended". Indeed, many of these tools are not easy to use and are in desperate need of user-friendly interfaces. The Werkgroep Informatica has already started a pilot project to make available the contents of their database in a familiar web format (www.th.vu.nl/~wiweb/const/index.htm) – see Figure 1.1 below for an example. Regrettably, only five chapters of the Old Testament are available in this format, providing information on morphological and syntactic levels. Subordinate clauses are indicated by means of indentation. The presentation of the data is done by means of Unicode and HTML (www.th.vu.nl/~wiweb/html/learn.htm).



Show All Hide All Gap: 25 Genre ☑ Mother ☑ Clause ☑ Verb ☑ Verse ☑ Line ID ☑ Print

+ וַ– Cj יְהִי Pr דְּבַר־ יְהוָה Su אֶל־ יוֹנָה / בֶּן־ אֲמִתַּי Co ap	N	<< -----	WayX	3sgM	01	1
+ לֵ–אמֹר Pr	N	[adjunct]	infc.	----	01	2
− קוּם Pr	NQ	<< infc.	imp.	2sgM	02	3
+ לֵךְ Pr אֶל־ נִינְוֵה / הָ–עִיר הַ–גְּדוֹלָה Co ap	NQ	<< imp.	imp.	2sgM	02	4
− וּ– Cj קְרָא Pr עָלֶיהָ Co	NQ	<< imp.	imp.	2sgM	02	5
− כִּי־ Cj עָלְתָה Pr רָעָתָם Su לְ–פָנָי Co	NQ	<< imp.	XQtl	3sgF	02	6
+ וַ– Cj יָּקָם Pr יוֹנָה Su	N	<< WayX	WayX	3sgM	03	7
− לִ–בְרֹחַ Pr תַּרְשִׁישָׁה Co מִ–לִּ–פְנֵי יְהוָה Aj	N	[adjunct]	infc.	----	03	8
− וַ– Cj יֵּרֶד Pr יָפוֹ Co	N	<< WayX	Way0	3sgM	03	9
+ וַ– Cj יִּמְצָא Pr אֳנִיָּה Ob	N	<< Way0	Way0	3sgM	03	10
− בָּאָה PC תַרְשִׁישׁ Co	N	[attrib.]	ptc.	-sgF	03	11
− וַ– Cj יִּתֵּן Pr שְׂכָרָהּ Ob	N	<< Way0	Way0	3sgM	03	12
+ וַ– Cj יֵּרֶד Pr בָּהּ Co	N	<< Way0	Way0	3sgM	03	13
− לָ–בוֹא Pr עִמָּהֶם Aj תַּרְשִׁישָׁה Co מִ–לִּ–פְנֵי יְהוָה Aj	N	[adjunct]	infc.	----	03	14
+ וַ– Cj יהוָה Su הֵטִיל Pr רוּחַ־ גְּדוֹלָה Ob אֶל־ הַ–יָּם Co	N	<< WayX	WXQt	3sgM	04	15

Inhoudsopgave

Hoofdstuk 1

Overzicht

vertaling

(1.1) En het woord van JHWH kwam tot Jona, de zoon van Amittai,

(1.2) zeggende:

(2.3) sta op,

(2.4) ga naar Nineve, die grote stad,

(2.5) en roep tegen haar,

(2.6) want hun kwaad is opgestegen voor mijn aangezicht.

(3.7) En Jona stond op

(3.8) om te vluchten naar Tarsis, weg van het aangezicht van JHWH.

(3.9) En hij daalde naar Jafo

(3.10) en hij vond een schip

(3.11) dat naar Tarsis ging

Figure 1.1. An example of WIVU's morphological and syntactic information, which is interactively available on the web (www.th.vu.nl/~wiweb/const/index.htm).

This thesis attempts to address the problem of underutilization by giving pointers for the integration and advanced processing and mining of sets of linguistic data. The author believes that visual presentations of the data and patterns in the data that are easy to understand could enhance the usage of Biblical information systems.

1.2.3 Integration as a solution to enhance utilisation

Having various electronic aids for the study of the Hebrew Bible is wonderful, but also overwhelming and even frustrating, due to the fact that various tools have to be used to study different levels and to get various perspectives. Bothma (1990) proposes the use of integrated Biblical information systems to enhance the process of computer-based education and to solve the problem of Biblical languages that are often studied in isolation. These systems should integrate introductory grammars, reference grammars, sources on the cultural background of the Bible and research databases. Various levels of granularity of data should be available for users with different levels of knowledge and requirements. Poswick (2004) also indicated the use of hypermedia to take Biblical research to a new level, "from the accumulation of electronic texts to the construction of hyper-textual links between them with all the cultural data which permit their interpretation".

Systems have been suggested and at least one has already been developed to display multi-level analyses of Hebrew clauses, integrating the various dimensions of clausal analysis in an interlinear table format on one screen, for example the Lexham Hebrew-English interlinear Bible (Van der Merwe, 2005). These tables resemble those found in relational databases, and this gives birth to the wish of being able to do *ad hoc* queries on the stored data. However, these tables cannot simply be transformed into relational database tables, because there is a separate table for each record (or clause) and the rows do not represent unique records. A closer inspection reveals that the rows actually represent various dimensions or levels of data-analysis that are strongly linked to the elements in the upper row. This type of interlinear table is in fact a twodimensional representation of three- (or multi-) dimensional linguistic data structures. Bothma (1992a) proposed and successfully tested the use of SGML, of which XML is a derivative subset, to provide a platform-independent databank of the linguistic and other related data.

This thesis addresses the need, expressed by Bothma (1992b), for syntactic and semantic databases of Biblical Hebrew. Such databases may enhance grammatical research because "manual searching for complex syntactic examples is extremely difficult and inadequate in that retrieved information is very often incomplete due to the size of the corpora of texts" (Bothma, 1992b: 340). Although various syntactic databases have become available, the author is not aware of any databases containing a separate module of semantic functions that may help users to understand the logical relations between the constituents of clauses and sentences.

The XML data structure suggested in this thesis may make a contribution to find ways to find an "appropriate information model for presenting Biblical information in an electronic form", with reference to integrating and storing information from various linguistic modules (cf. Bothma, 1992b: 345).

The advantages of XML, however, are not limited to the creation of a database structure. According to Van der Merwe (1995: 419) the purpose of an electronic reference grammar "plays a major role in determining its structure and content". The extensibility and adaptability of advanced mark-up languages such as XML make them ideal to implement a custom-made macro-structure, which should, for example, fulfil the following requirements: "An electronic BH [reference grammar – JHK] should serve as a cheap up-to-date, as well as updateable, source of easily retrievable information on BH for readers of the BH text of the OT. These readers may have various degrees of receptive competency of BH" (Van der Merwe, 1995: 420).

The use of XML as mark-up language to tag the data in a bank of Biblical data may also enable learners to move between teaching and reference textbooks and to emulate deductive grammars, according to Bothma (1992a). Furthermore, it could also facilitate the move in Biblical research focus from textual aspects to communicative aspects (Poswick, 2004).

The combination of hypermedia, such as XML, and database concepts forms a strong and promising alliance of techniques, which facilitates solutions to cater for a diversity of domains, users and applications, including integrated Biblical information systems (Claassen & Bothma, 1988: 84). This thesis may be a step in the right direction to solve the problem of new requirements that may be laid down by the "shift of paradigm from exegesis based on a philological approach, to hermeneutic based on a linguistic and socio-linguistic approach" (*sic*) (Poswick, 2004), since the use of an extensible, multidimensional data structure could facilitate the accommodation of other types of linguistic and non-linguistic data.

1.2.4 Visualisation and flexibility

Adding visualisation techniques to the mixture of XML and databases could provide even more exciting possibilities. Claassen & Bothma (1988: 88-89) suggest the use of visualisation to direct users in finding their way through the convoluted sets of paths in hyperspace. Advanced processing and visualisation techniques may also make a contribution towards the development of user-friendly interfaces (Bothma, 1992b: 348). This thesis aims to contribute to the attainment of this goal by proposing a macro-structure for the integration and packaging of Biblical Hebrew linguistic information and by experimenting with some visualisation techniques to render captured data in innovative ways.

According to Andersen & Forbes (2003: 44) one of the requirements of a proper rendering of syntactic structures of Biblical Hebrew is that it should be pictorial, that is "clearly and concisely diagrammed". They use graphs and trees to visualise ("represent") the hierarchical syntactic structures of Biblical Hebrew clauses.

Scalability is a serious issue that needs to be addressed if one would like to represent aggregate, linguistic information on a lateral level across the single units of the textual corpus. According to Andersen & Forbes (2003: 45) the text of the Hebrew Bible approximately consists of 59 000 main clauses and 13 000 embedded clauses. Although this thesis will propose ways to compile such aggregate information, it is still limited to one chapter only. Visualising lateral information of larger sections, books or the whole Hebrew Bible will surely create new and difficult challenges for researchers.

Tov (2006: 337) differentiates between non-flexible and flexible Biblical Hebrew software. Non-flexible tools reflect only the result of computer-assisted software in textual format, and the reader does not have access to the original data or tool itself; these may even have become obsolete. Flexible tools, however, allows interactive use of the tool and data. These tools, especially the flexible versions, may be used as "an extension of our own thinking" and to "improve and expand the areas of our research". The tools that are already available may be categorised according to their intended purpose, i.e. to serve as aids in authorship studies, analyses of stylistics and linguistics, as well as statistical and text-critical studies (Tov, 2006: 338-342). Making use of interactive visualisation tools could pave the way to more flexible Biblical Hebrew linguistic software.

In addition to the representation of linguistic data, a "comprehensive Biblical information system" should, according to Bothma (1995), include images of textual-critical material and cultural-historical objects in order to facilitate the preservation, publication and research of the ancient manuscripts. Multiple disciplines are involved in such a system which necessitates team work since no researcher could have all the skills needed to construct the various building blocks. Bothma (1992b: 348) highlights the necessity of cooperation between linguists, theologians and IT specialists that is needed to build well-designed Biblical information systems. Although there might not be many researchers who have an in-depth command of all of these disciplines, the members of the team should have a basic understanding of the complex nature of each other's abilities and fields.

Although some of the projects, discussed above, do facilitate rather advanced searches, they do not clearly differentiate between the linguistic levels of syntax and semantics; neither do they facilitate comparative studies on and between these levels.

1.3 Problem statement

As indicated above, there is still a need for more language-oriented, multidimensional Biblical information systems, in which the linguistic characteristics of the Hebrew text of the Old Testament are embedded, to enable researchers to do advanced *ad hoc* queries. For example, a researcher may want to do a specific search in order to find good examples of a certain syntactic structure, to study semantic role frameworks in Biblical Hebrew, or to explore the mapping of syntactic functions onto semantic functions.

One of the core foci of Information Systems is the study of databases. Among other approaches, such as object-oriented database systems, relational database management systems remain the technology most often used and taught, usually within the realm of business management. This widely used and standardised technology is, however, not of its own accord the most applicable for text-based databanks.

Standard (relational) database management systems such as MS Access are, in fact, not ideal for the storage and retrieval of linguistic data, because they require groups of similar records having highly structured (usually twodimensional) data. Free text, however, is covertly structured. If the linguistic elements of a clause, for example, are captured in a standard database, the word order of the clause's elements are lost, and it results in tables with an unacceptably high number of columns, many of which contain null values, because every sentence's structure may differ considerably from the former one. According to Bourret (2003) the structure of sentences "varies enough that mapping it to a relational database results in either a large number of columns with null values (which wastes space) or a large number of tables (which is inefficient)".

Furthermore, to capture information of various linguistic modules, the relational database should contain various tables with columns for each of the syntactic and semantic functions, word groups, etc., causing even more overhead. Therefore, storing text-related data in conventional databases is not ideal. It "artificially creates lots of tuples/objects for even medium-sized documents" (Xyleme, 2001: 3) and "requires multiple index lookups and multiple disk reads" (Bourret, 2003). Complex format mappings would be needed to convert the data into a table-based relational database management system. Many joins would be needed to reconstruct the original document when queries are run (Vakali et al., 2005: 62).

Since traditional, relational databases seem to be problematic, one has to look for an efficient solution elsewhere. XML has been identified by various computational linguists as viable technology for linguistic databases and computing (cf. Witt, 2002; Witt, 2005; Bayerl et al., 2003; Burnard, 2004). Due to its flexibility, XML may be used to create either a less structured, text-oriented file or a more structured, database-oriented document.

There are, of course, exceptions to the rule that XML is the best option for language-related database projects. According to Bradley (2005: 133) some projects do begin with text, but hidden below the surface is data-oriented material that better suits a relational database approach. For the foreseeable future he proposes a complementary approach that uses both relational database management systems for such projects and XML-based systems for materials that are saying "more 'subtle' things" (ibid.: 134, 141). Since XML also provides opportunities for the integration of existing databases (Golfarelli et al., *s.a.*), this thesis will investigate its usability to capture and explore data from various linguistic modules. The proposed structure of the XML database could be used in future as a model for the integration of information extracted from existing but divergent Biblical Hebrew databanks.

1.4 Research questions

1.4.1 Main research question

The main research question that is addressed in this thesis, may be formulated as follows: *How can XML be used to build an exploitable linguistic database of the text of the Hebrew Bible?*

The scope of the empirical study will be limited to Genesis 1:1-2:3, the first pericope of the Hebrew Bible. Scalability issues to cover more extensive texts, such as the whole Hebrew Bible, will not be dealt with.

1.4.2 Secondary research questions

The main focus of the thesis falls on the use of XML for the permanent, platform-independent, storage of linguistic data, which may be used by various algorithms, programming languages and visualisation techniques for in-depth processing. Flowing from the main research question, the following secondary research questions will be addressed in the various chapters of the thesis:

- Chapter 2: *How can multidimensional Biblical Hebrew linguistic data be captured and stored in the computer's temporary memory using a programming language such as Visual Basic?*
- Chapter 3: *How can multidimensional Biblical Hebrew linguistic data be processed with a programming language such as Visual Basic?*
- Chapter 4: *How can multidimensional Biblical Hebrew linguistic data be stored permanently to allow a stable environment for editing and processing?*
    - *Why should XML be explored as an option to store an exploitable database of linguistic data?*
    - *How can XML represent inherently multidimensional data?*
    - *How can XML represent the phonological representation and translation, as well as the morpho-syntactic, syntactic and semantic analyses of the Biblical Hebrew text?*
- Chapter 5: *How can linguistic data be recovered from and saved to a permanent storage device (such as an XML database)?*
- Chapter 6: *How can linguistic data be explored to unveil hidden patterns in and between the various language modules?*
- Chapter 7: *How can visualisation be used to enhance text-mining of multidimensional linguistic data?*

1.5 Hypothesis

The central theoretical statement of the thesis is: *Taking advantage of the flexibility and inherent hierarchical nature of XML provides a suitable technology to transform free text and its covert linguistic characteristics into a platform-independent database that may be explored and mined to uncover hidden linguistic patterns in the Hebrew Bible text.*

The candidate expects to prove the following propositions:

- *A clause's word order can be kept intact while other features such as syntactic and semantic functions are marked up as related elements.*
- *The inherent hierarchical nature of XML is ideal to design a well-structured database containing linguistic data from various modules in one, centralised, data structure that is suitable as a permanent storage device.*
- *The elements from the XML database can be accessed, processed and visualised by a third-generation programming language, such as Visual Basic and Java.*
- *A threedimensional array is an effective programming tool to process and mine the data.*

To test the hypothesis and expected findings, a multi-modular analysis of Genesis 1:1-2:3 will be implemented using XML, while Visual Basic 6 will be used for the online analytical processing (OLAP) on the data. A visualisation tool, created in Java, will be used to investigate visualisation of the linguistic data as a data-mining tool.

1.6 Positioning of Linguistic Information Systems within a research discipline

Science is traditionally regarded as having two major branches, natural sciences and human sciences. The natural sciences study subjects such as Chemistry and Physics and use mainly empiricist methods.[3] Human sciences consist of the social sciences (such as Economics and Psychology) and the humanities (Arts, Language and Philosophy). Although empiricist methods are used in some sciences, others in the humanities do not necessarily share this methodological approach. Most of the humanities use a rationalist approach that is not empiricist.

A non-empiricist point of departure for research is applicable when phenomena are studied which cannot be proven by hard, concrete facts and empirical observation. Cilliers (2005) pleads for alternative scientific methodologies regarding complex systems, which are modest and provisional, acknowledging that our understanding is limited and changing. Restricting the term *research* to empiricism is anachronistic, given the contemporary insight that knowledge is never final and beyond dispute. Modest claims about knowledge, however, invite knowledge workers to persevere in an ongoing search for meaning and generation of understanding. These "softer" research goals are the unique foci of especially the humanities.

Although it is tempting to view Information and Communication Technology (ICT) as "the epitome of rational expression", empiricist research methods are only appropriate for the study of engineering and algorithmic issues related to it. Therefore, the study of ICT is divided into three branches, i.e. Computer Science (the natural science branch), Information Science (the humanities branch) and Information Systems (the social sciences branch)[4]. However neat this division may seem, many phenomena often require a mixture of these perspectives for

[3] Empiricism is an approach based on the idea that scientific knowledge can only be valid if it is based on empirical observation and measurement.

[4] Information Systems is primarily regarded as a social science focusing on the influence of computer technology on human society.

comprehensive research. For example, Information Systems (IS) is primarily regarded as a social science, because it investigates socially constructed issues such as the influence of ICT on organisations. But it also sometimes studies harder (only factual) phenomena, typical of the natural sciences, such as programming techniques and algorithms, in order to build efficient software solutions for organisations and industry. Furthermore, it also has links with the humanities when it focuses on the use and application of ICT in education, health care and other humanistic[5] focus areas.

Information Systems should, therefore, be regarded as an interdisciplinary science. It should not only aim to add value to other disciplines, but also borrow from other contributing ICT (and non-ICT) disciplines in order to strengthen their alliances. "The *power* and not the weakness of IS research models is precisely that they situate IS constructs within constructs that other disciplines study" (Agarwal & Lucas, 2005: 390). For example, in one of the research foci of Information Systems, namely Human Computer Interaction, there are elements of all three branches of ICT: it studies the behaviour of computer system users, the use of professional algorithms to produce human-oriented output, and friendly design of interfaces by means of inputs from graphical design and multi-media.

As a science with strong links to the human sciences, there has indeed been a growing acceptance in Information Systems that empiricist research is not the only valid scientific methodology that could be used to produce good research. Avgerou (2005:105), for example, argues for critical research using interpretive methods in Information Systems to complement empirical and formal cognitive methods. She regards critical research as a process that aims to make sense of the investigated scenario, a radical procedure in which researchers' human capacities such as tacit knowledge and moral values are involved. "I see research as the art of putting together research questions with a critical content, multiple theories and epistemological awareness to develop claims of truth. This art cannot place

[5] The term *humanistic* is used in this thesis as an adjective of the noun *humanities*. It does not refer to the philosophy of Humanism. (Compare TheFreeDictionary, 2008: "**humanistic** - pertaining to or concerned with the humanities". Also see Aarseth, *s.a.*)

confidence for producing valid knowledge on adhering to a testable theory or research practice" (ibid.*:* 108). Although the knowledge claims contributed by interpretive case studies should be regarded as soft facts, they are still valid and should be generalised in clear formulations aimed at identified target audiences (Barret & Walsham, 2004: 298, 310).

Bondarouk & Ruël (2004) argues for the use of discourse analysis to enable a hermeneutic approach in the analysis of information systems documents. Discourse analysis is another non-empiricist scientific method. It is essentially interpretive and constructivist. It tries to "give a meaning to a text within a framework of the interpreter's experience, knowledge, time, epoch, culture, and history". It believes that understanding is an open, continuous process and that there is no final, authoritative interpretation.

Other non-empiricist approaches in Information Systems, which will not be discussed here in detail, are (cf. Carlsson, 2003; Du Plooy, 1998: 53-68):

- action research (the researcher collaborates with members of the organisation to experiment with possible solutions for a problem)
- actor network theory (the researcher studies the technical and social aspects of IT as a unity because values are believed to be built into software)[6]
- critical realism and adaptive theory (the researcher attempts to combine and synthesise empiricism and interpretivism)
- ethnography (the researcher participates in activities of the organisation that is studied)
- grounded theory (the researcher derives theory by means of qualitative data analysis)

[6] Like adaptive theory which is epistemologically neither positivist nor interpretivist (Carlsson, 2003), actor network theory (ANT) is positioned between deterministic and constructivist theories (Cordella & Shaikh, 2003). It studies the reciprocal influence of technology and society, the interaction between the human and non-human actors that constitute a network. Reality is believed to come into existence through this interplay.

- structuration theory (the researcher regards human agency and social structure as an inseparable duality)

From the discussion above it should already be clear that it has become acceptable to use other, non-empiricist and interpretive methods in Information Systems research. Humanities Computing may be another tool that could introduce a "softer" view and use of computers that would be more applicable in the human sciences than the "harder" approaches that are typical of the natural sciences. The representation of socially related data is one of the basic ventures of Humanities Computing (Neyt, 2006: 2-5). This approach has a broader scope than the mere use of computers only to empirically confirm or reject hypotheses, which constrains meaning (Ramsay, 2003). According to Aarseth (*s.a.*) Humanistic Informatics is the discipline that studies "the changing role of technology in the Humanities, as in society in general".

The study of databases also forms part of the discipline of Information Systems (IS). According to Vessey et al. (2002: 167) database management is one of the topics "at the heart of the IS discipline in that they are central to IS curricula and therefore to IS careers". The creation of knowledge databases and the exploration of these electronic repositories are thus part and parcel of Information Systems research, even if the encoded data come from other disciplines.

Mark-up languages, traditionally, fall more within the research focus of Information Science. Unsworth (2001) regards the marking up of texts as a form of knowledge representation within the field of Humanities Computing: "For humanities computing, knowledge representation is a compelling, revelatory and productive way of doing humanities research--and in many ways, it is what humanities computing has been doing, implicitly, for years." The study of mark-up and ontologies (taxonomies) is closely related to the study of language and semantics, which forms part of the broader discipline of humanities (the study of the arts, language and literature, philosophy and human culture). Research on the structure and use of mark-up languages to indicate metadata, therefore, operates within the grey area between Linguistics and Information Science.

Due to the flexibility of XML, an extensible mark-up language, it may also be used to create text-based databases. As argued above, the study of databases forms part of the discipline of Information Systems. The study of XML databases therefore falls within the field of Information Systems, with strong ties to Information Science. Aarseth (*s.a.*), for example, regards research on digital document representation and hypertext uses as part and parcel of *Humanistic Informatics* (also called *Alfa-informatica*): "In particular, text mark-up systems such as SGML, and the potential and limits of exploratory data analysis, can and should provide extremely interesting subjects for the field." The humanistic informatician fulfils an interstitial role between two (or more) different disciplines and needs to be acquainted with the methods and problems of all of these (ibid.).

The transformation of language and texts, via mark-up, into linguistic databases may, therefore, be viewed as a subfield of Humanistic Information Systems. This subfield, which is also the broader focus of this study, may be called *Linguistic Information Systems* or *Natural Language Information Systems*.

Natural Language Processing (NLP) provides another perspective on the study of the relations between ICT and language. This interdisciplinary field focuses on the simulation of language understanding and production by means of computer algorithms. This field combines Linguistics and Computer Science, a natural science. For an example of Natural Language Processing of Biblical Hebrew, compare Petersen (2004a), who wrote a program to automatically create conceptual graphs of the semantics of Genesis 1:1-3 using data from an Old Testament database.[7] The combination of Linguistic Information Systems and Natural Language Processing is called *Computational Linguistics*.

The creation of linguistic databases is, of course, not an end in itself. Besides being used to prove hypotheses, it may also be used to suggest new ideas and theories. Ramsay (2003) suggests that computing humanists should rather use software to

[7] Petersen used XML as input and output format for his lexicon and ontology. This part of his study may be regarded as more information system-oriented.

discover a multiplicity of meanings in literary sources. Such an approach will deepen the subjectivity that is essential for the creation of critical insight. Researchers and software creators should therefore work towards alternatives for the traditional, statistics-based "forensic semiotics" in the processing of texts in order to change the computer into a tool that supports interpretive processes: "[R]ather than to extol the computer as a scientific tool that can supposedly help prove particular facts about a text, we would do better to focus on its ability to help read, explore, experiment, and play with a text" (Sinclair, 2003: 176).

This thesis focuses on the use of an XML database to facilitate storage, exploration and visualisation of multidimensional linguistic data. It involves a combination of research on linguistics, database management and mark-up, as well as the visualisation of the patterns entrenched in tagged data. Since it focuses on the use of a database for the study of language it may, therefore, be regarded as a research endeavour in Linguistic Information Systems.

In a study with a similar approach to this one, T. Sasaki (2004) proposed design principles for XML documents to facilitate lexicographical and grammatical studies of Modern Hebrew. He reserves a place for the creation and use of annotated text databanks within the field of Computational Linguistics, which he defines as the "interface between Hebrew linguistics and computer science" (2004: 17).[8] Even if the construction of such an annotated databank could "seem rather naïve to NLP-oriented computational linguists" (T. Sasaki, 2004: 22), it is an essential task for the computational linguist who is more interested in the use of existing information technology to enhance his/her study of language than in the creation of algorithms that simulate human language understanding or production.

An XML schema defines the structure and content of the databank containing the XML mark-up tags (Clark et al., 2003). A schema is preferred to a DTD (document type definition) since it is more advanced and "more closely maps to database

[8] According to the exposition above it should rather be the interface between Hebrew Linguistics and Information Systems.

terminology and features", allowing the definition of variable types and valid values for the elements (Rob & Coronel, 2007: 579). Tags that are used to digitalise text "are not merely structural delineations, but patterns of potential meaning woven through a text by a human interpreter" (Ramsay, 2003: 171).

The discovery of meaningful patterns in numerical data (data mining) and in textual data (text mining) is another essential part of Information Systems research. It should be noted that this discovery of new information is an intentional process. Any new knowledge produced by it, is not simply discovered, but created (cf. Du Plooy 1998: 54, 59). The patterns that the candidate wants to unveil in the Genesis 1:1-2:3 XML document are covertly embedded within other visible patterns, i.e. the overt patterns specified by the schema. The creative process of knowledge discovery should be a stimulating, but "careful and responsible development of the imagination" (Cilliers, 2005: 264). This could eventually improve linguists' understanding of language as a complex social system, because "our understanding of complex systems cannot be reduced to calculation" (ibid.).

The thesis also contains elements of Computer Science and Information Science. Algorithms are developed to process and mine the linguistic data in the XML database, and visualisation is discussed and illustrated as a means to augment the exploratory searches for patterns hidden in the data. The study aims to show how a graphical visualisation tool could be used to stimulate imaginative knowledge-creation processes in a responsible way. In order to reach this goal the researcher must be enabled to perform experimental, trial-and-error investigations of the text that could reveal exciting new patterns built on re-orderings of marked-up text.

1.7 Research style and methodology

The study will entail various literature studies and empirical programming experiments to investigate the various research questions. Although the topic and the concepts for the various subsections of the project were planned as a whole (see the main and secondary research questions above), each chapter was approached as a unit so that it could be submitted as a conference paper or journal article during the course of the PhD study. Therefore, each chapter has a literature study and practical component. After completion of all the papers or articles, it was decided not to rewrite the whole thesis, consolidating all literature review sections into one chapter, since the information is closely related and interwoven with the application in each chapter. Keeping the basic structure of the thesis as a collection of independent, but related papers, each building on the preceding one, will also enable readers of the (unpublished) thesis to read each chapter as a unit. However, there is a logical flow from the second to the seventh chapter. Likewise, the practical programming concepts and examples also progress from very basic concepts to rather advanced data-mining and visualisation applications. The study first explores ways to store linguistic data in the computer's temporary memory and the various presentations of subsets facilitated by these data structures. Next, permanent storage and conversion betweeen temporary and permanent storage is discussed. Once a stable platform for storage and editing has been established, advanced processing and graphical interfaces for text mining are investigated.

While the suitability of threedimensional array processing, XML and visualisation concepts are presented and evaluated by the literature studies, the empirical and experimental components are used to test these ideas against the test data provided by Genesis 1:1-2:3. Therefore, the overall, dominant, research style of the thesis is constructive, i.e. developing a new framework and pursuing technical developments (cf. Cornford & Smithson, 1996: 43). Although the basic philosophical point of departure is positivistic, a large element of interpretation is built into the data-analyses used as data. This part of the study may be regarded as anti-positivistic, qualitative and interpretative.

The Visual Basic programming code presented in this thesis could still be optimised and improved to a large extent. Since the main focus of this study is not on the algorithms themselves, but on their usability to store and mine linguistic data, it has been left with a view to follow-up work to make the code more efficient and elegant. This could be done, for example, by migrating the project to a fully object-oriented language, such as Visual Basic 2005, and by implementing more procedures, functions, classes and objects.

1.8 Research plan

Although the six content chapters of the thesis (Chapter 2-7) had all initially been planned to constitute a coherent body of work (see 1.4 above), they were written and submitted separately as conference papers, some of which appeared in published conference proceedings, scholarly publications or on the internet after the comments in peer reviews have been taken into account and processed in the final versions. Other papers were revised, using the valuable comments of international and local reviewers, and have eventually been published in accredited journals. Parts of the introductory chapter (Chapter 1) were published in the *Innovate* magazine, published by the EBIT Faculty of the University of Pretoria. An overview article, which forms part of the conclusive chapter (Chapter 8), has been accepted as a paper by an international conference (Conf-IRM 2008). The feedback provided by many colleagues over the past five years has resulted in an organical development of the ideas in the thesis and is highly appreciated. See the table below for more specific details on the mapping between the various chapters of the thesis and the related research outputs.

As indicated above the literature review does not form a separate chapter, but is integrated into the content chapters, as well as the introductory chapter. The literature studies are interleaved with interpretative and constructive research experiments, namely the analysis of a Hebrew text to be used as test data and the creation of Visual Basic and Java programs to transform, store, process and data-mine the linguistic data. The Hebrew text of Genesis 1:1-2:3 has been analysed linguistically on various levels and this data was used as test data throughout Chapters 2 to 7. The text of Leviticus 1 was also analysed and coded in the same XML structure to test the plug-in scalability of the programs. Although this test was successful, this XML document will not be included or discussed in this thesis.

Although the original papers and articles have been revised, some extensively, others only to a lesser extent, in order to constitute a coherent and consistent text, the history of the origin is still present in the underlying structure of each chapter.

Therefore, some information is repeated, summarised or extended in the various chapters, partly due to the fact that the information was interwoven in the original research outputs, but also having the added benefit that each chapter may still be read as single, independent unit.

The following chapters correspond to the research outputs referred to above:

Chapter 1	The philosophical section(s) on Linguistic Information Systems is a revised version of: • KROEZE, J.H. 2007. Linguistic information [systems] - a humanistic endeavour. *Innovate,* 02: 38-39. (Published by the University of Pretoria, EBIT.)
Chapter 2	• KROEZE, J.H. 2004. Towards a multidimensional linguistic database of Biblical Hebrew clauses. *Journal of Northwest Semitic Languages (JNSL)*, *vol. 30, no. 2, pp. 99-120.* • Revised version of 2004 *AIBI VII* paper, Leuven, Belgium: Processing Hebrew clauses using threedimensional arrays.
Chapter 3	• KROEZE, J.H., BOTHMA, T.J.D. & MATTHEE, M.C. 2008? Slicing and dicing a linguistic data cube. Accepted for publication in *Handbook of Research on Text and Web Mining Technologies* (Nov. 2007), edited by M. Song. • Revised version of 2004 *SASNES* paper, RAU, Johannesburg: Slicing and dicing cyber cubes of Biblical Hebrew clauses.
Chapter 4	• KROEZE, J.H. 2006. Building and displaying a Biblical Hebrew linguistics data cube using XML. Paper read at *Israeli Seminar on Computational Linguistics (ISCOL),* Haifa, Israel, 29 June 2006. Available: http://mila.cs.technion.ac.il/english/events/ISCOL2006/ISCOL20060629_KroezeJH_XML_Paper.pdf.
Chapter 5	• KROEZE, J.H. 2007. Round-tripping Biblical Hebrew linguistic data. In *Managing Worldwide Operations and Communications with Information Technology* (Proceedings

	of 2007 Information Resources Management Association, International Conference, Vancouver, British Columbia, Canada, May 19-23, 2007), edited by M. Khosrow-Pour, pp. 1010-1012. Published in book format and on CD by IGI Publishing, Hershey, PA.
Chapter 6	• KROEZE, J.H. 2007. A computer-assisted exploration of the semantic role frameworks in Genesis 1:1-2:3. *Journal of Northwest Semitic Languages (JNSL)*, *vol. 33, no. 1, pp. 55-76.* • Revised version of 2006 *SASNES* paper, Unisa, Pretoria: Semantic role frameworks extracted from a multidimensional database of Gen. 1.
Chapter 7	• KROEZE, J.H., BOTHMA, T.J.D., MATTHEE, M.C. & KROEZE, J.C.W. 2008. Visualizing mappings of semantic and syntactic functions. *Proceedings of the Sixth International Conference on Informatics and Systems (INFOS2008),* Cairo, Egypt, 27-29 March 2008. (On CD. ISBN: 977-403-290-X.) Available online (http://www.fci-cu.edu.eg/infos2008/infos/MM_10_P061-072.pdf). • KROEZE, J.H., BOTHMA, T.J.D., MATTHEE, M.C., KROEZE, J.C.W. & KRUGER, O.C. 2008? Designing an interactive network graph of modular linguistic data in an XML database of Biblical Hebrew. Paper abstract accepted for AIBI VIII (2008). (The empirical part of this paper does not form part of the thesis, but should be regarded as a post-doctoral project building on the thesis.)
Chapter 8 (conclusion and summary)	• KROEZE, J.H., BOTHMA, T.J.D. & MATTHEE, M.C. 2008. From tags to topic maps: using marked-up Hebrew text to discover linguistic patterns. Paper accepted by *Conf-IRM 2008*. • Beta version read as guest lecture, University of Leiden, 31 May 2007.

1.9 Structure of the thesis

The thesis is structured as follows:

Front matter

The front matter consists of the cover page, title page, preface, table of contents, abstract and English and Afrikaans summaries.

Chapter 1: Introduction

This chapter discusses the background and goals of the thesis (see 1.1).

Chapter 2: Towards a multidimensional linguistic database of Biblical Hebrew

This chapter discusses the use of threedimensional arrays in Visual Basic 6 to build a data cube in the computer's RAM (random access memory). The theory is applied to the Hebrew text of Genesis 1:1-2:3, using the Visual Basic 6 programming language. The linguistic data is declared and initialised in a module of the program. The linguistic data cube is called a "clause cube". The clause cube provides a structure that may be used to integrate data from various linguistic modules. Data integration is a typical warehousing problem (Xyleme, 2001: 2). More data warehousing concepts are discussed in Chapter 3.

Chapter 3: Slicing and dicing the clause cube

This chapter discusses and illustrates the application of typical online analytical processing techniques and data warehousing operations, like rotation, drilling down, and slicing and dicing on the clause cube while it resides in the RAM.

Chapter 4: Building and displaying the clause cube using XML

This chapter discusses the use of the mark-up language XML to build a database for permanent storage of the linguistic data in the clause cube. An XML database is designed and programmatically built to store the test data.

Chapter 5: Conversion of the Genesis 1:1-2:3 linguistic data between the XML database and the array in Visual Basic

This chapter discusses the process of reading the linguistic data from the XML database (permanent storage) into the threedimensional array (temporary storage) for processing, as well as saving the updated data back to permanent storage. A program is written and used to "round-trip" (convert) and edit the test data.

Chapter 6: Advanced exploration of the clause cube

This chapter discusses data mining on the linguistic data using algorithms coded in Visual Basic. The logic of the program is discussed and the source code is provided. The results of two experiments are provided in textual format and compared to existing linguistic knowledge and hypotheses. The subsets of the threedimensional array, especially those that contain processed, aggregated information, may be regarded as temporary datamarts of the XML clause cube. These datamarts may be implemented as one-, two- or threedimensional arrays. The slicing and dicing and other operations done on the data during the data-mining process may be regarded as XOLAP, OLAP done on XML databases (Wang & Dong, 2001: 50).

Chapter 7: Visualisation of the Biblical Hebrew linguistic data in the XML cube

This chapter discusses theoretical aspects of visualisation and evaluates the results of a graphical text-mining tool that maps the syntactic and semantic functions in the test data.

Chapter 8: Conclusion

This chapter gives an overview of the content of the whole thesis and recapitulates the most important conclusions.

Bibliography

All the sources referenced in the thesis are listed in the bibliography, using the Harvard method, based on the guidelines found in Botha & Du Toit (1999) and Van der Walt (2002).

Addenda

The fourteen addenda contain additional material, such as the transcription and "ontologies" (taxonomies) of word groups, syntactic functions and semantic functions, used in the tagging of the linguistic database of Genesis 1:1-2:3. The XML database, source code and executable program files are also included as addendas. Some of the addenda have two versions, an executable program as well as a textual version of the source code. Due to the extensive size of this collection it is only provided on CD. Essential parts, however, are reproduced in the chapters where the material is discussed.

1.10 Contribution to the field of ICT

This study illustrates that a threedimensional data structure can be used to represent inherently multidimensional linguistic data regarding Biblical Hebrew clauses. Knowledge from various linguistic modules is captured and integrated in a single clausal data cube (clause cube), which is regarded as an efficient way of using a threedimensional database, implemented both in an array and XML structure. The structure of the database may in future be used as a model for the integration of linguistic data that have been captured in various other computer software systems. The chapter on round-tripping shows how integrated linguistic data may be converted between data structures in permanent and temporary storage.

The captured data can be viewed and manipulated in various ways, for example to create stacks of twodimensional interlinear tables showing required aspects of the data of clauses. In this way the threedimensional data cube facilitates actions that are typical of online analytical data processing and data warehousing. Such software can facilitate the linguistic analysis with which any exegetical process should commence, which in turn can benefit a multidimensional approach to biblical exegesis (cf. Van der Merwe, 2002: 94). It also facilitates a format in which the biblical text is presented for readers "succinctly enough to be handled by the short-term memory", thus enhancing the success of the communication process (ibid.).

The thesis also illustrates how text data mining may be performed on the linguistic information of an ancient language. Another contribution is the application of visualisation concepts to enhance text mining procedures by using *i.a.* graphical topic maps.

To sum up, the thesis makes a contribution to the field of Information and Communication Technology by demonstrating how software tools and concepts borrowed from Information Systems, Information Science and Computer Science may be used and adapted in Linguistic Information Systems for knowledge representation and processing.

1.11 Definition of terms

The definitions of the terms that are provided below are not official definitions quoted from other literature, but working definitions to indicate how these concepts are understood and used in this thesis:

Clause cube: A threedimensional data structure, either in temporary or permanent storage, used to capture inherently multidimensional linguistic data on the clausal level.

Data cube: A threedimensional data structure used to store related aspects of data in a single structure for efficient processing.

Data mart: A data mart is a subset of a data warehouse and contains extracted and summarised data related to a specific, required perspective on the data.

Data mining: The uncovering of hidden patterns and trends in (usually numerical) data.

Data warehouse: A collection of basic and aggregated data used to discover business intelligence. The clause cube in this thesis only contains detailed data, but the slices and aggregated data mined from it may be regarded as data marts, while the collection of all of these data structures may be seen as a linguistic data warehouse.

Dicing: In this thesis dicing is used to refer to the extraction of specific data nuggets within the clause cube or its slices. (Dicing is also often used as a synonym for rotation.)

Drilling down: Moving from aggregated data to the underlying detailed data in the data cube.

Hyper cube: A multidimensional data cube, having four or more dimensions.

MOLAP: Multidimensional online analytical processing, including processing on a threedimensional data cube and clause cube.

OLAP: Online analytical processing, i.e. the discovery of (usually business) intelligence using software to explore databases, data marts and data warehouses to trace (business) trends. In this thesis OLAP is performed on linguistic data captured in a clause cube.

Rotation: Getting various perspectives on the data by "spinning" the cube and looking at the different external planes.

Round-tripping: The conversion of data in both directions between a permanent data storage facility (such as an XML file) and temporary computer memory (RAM).

Slicing: Filtering the data in a data cube to reveal a specific subset or the data needed for business decisions or to test academic hypotheses.

Text mining (= text data mining): The discovery of information hidden in textual data.

Threedimensional array: A collection of related variables used to store various aspects of interconnected data, which can be used to implement a data cube.

Visualisation: 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.

XML: *eXtensible Markup Language*, similar to HTML but containing semantic value in the tags, which enhances its functionality and facilitates the creation and rendering of databases.

XOLAP: MOLAP done on a three- or multidimensional XML data structure.

1.12 Conclusion

This chapter stated the point of departure for the thesis, namely the proposition that threedimensional data structures can be used to capture integrated multidimensional linguistic data. The background, aims and outline of the thesis have been discussed. The use of threedimensional arrays and XML structures to implement and test this assumption has been indicated as the core part around which the rest of the thesis is built. Advanced data exploration and visualisation, discussed in later parts of the thesis, make use of these underlying data structures and data warehousing processes for knowledge creation. Although the various divisions have been structured and organised to form a coherent and logically flowing work when the thesis is read as a whole, enough information has been repeated so that each chapter may also be read as an independent unit. Readers are also informed and referred to the various conference proceedings or journals where large sections of this thesis have been or will be read and published.