# An Investigation into Teaching Description and Retrieval for Constructed Languages

A thesis presented in partial fulfilment of the requirements for the degree of Master of Science in Computer Science at Massey University

Son Hoang
2004

# Abstract

The research presented in this thesis focuses on an investigation on teaching concepts for constructed languages, and the development of a teaching tool, called VISL, for teaching a specific constructed language.

Constructed languages have been developed for integration with computer systems to overcome ambiguities and complexities existing in natural language in information description and retrieval. Understanding and using properly these languages is one of the keys for successful use of these computer systems. Unfortunately, current teaching approaches are not suitable for users to learn features of those languages easily.

There are different types of constructed languages. Each has *specific features* adapted for specific uses but they have in common explicitly *constructed* grammar. In addition, a constructed language commonly embeds a powerful *query engine* that makes it easy for computer systems to search for correct information from descriptions following the conditions of the queries. This suggests new teaching principles that should be easily adaptable to teach any specific structured language's structures and its specific query engine.

In this research, teaching concepts were developed that offer a multi-modal approach to teach constructed languages and their specific query engines. These concepts are developed based on the efficiencies of language structure diagrams over the cumbersome and non-transparent nature of textual explanations, and advantages of active learning strategies in enhancing language understanding. These teaching concepts then were applied successfully for a constructed language, FSCL, as an example. The research also explains how the concepts developed can be adapted for other constructed languages.

Based on the developed concepts, a Computer Aided Language Learning (CALL) application called VISL is built to teach FSCL. The application is integrated as an extension module in PAC, the computer system using FSCL for description and retrieval of information in qualitative analysis. In this application, users will learn FSCL through an interconnection of four modes: FSCL structures through the first two

modes and its specific query engine through the second two modes. After going through four modes, users will have developed full understanding for the language. This will help users to construct a consistent vocabulary database, produce descriptive sentences conducive to retrieval, and create appropriate query sentences for obtaining relevant search results.

# Acknowledgements

There are a number of people I would like to thank for their support throughout this study.

First and most importantly, I would like to express a deep gratitude to my supervisor, Dr. Eva Heinrich for her encouragement, enthusiasm and invaluable advice. Without her tremendous help, this study would not be possible.

I also would like to thank my family for their continuing encouragement, support and for always being there for me.

# TABLE OF CONTENTS

## TABLE OF FIGURES

## Chapter 1

# Introduction

The explosion of amount of information available in recent years has created an accompanying problem in that the large number of multimedia documents such as pictures, audios and videos has been a big challenge for information description and retrieval. This leads to requirements for creating new technologies to handle that huge volume of multimedia data. Christel et al. (1997, p23) claimed that "as digital video libraries emerge, users will need tools that handle effectively the dynamic nature of video; they will not likely invest hundreds of hours to find a single, relevant segment within a 1000-hour library". Voss (2001, p1) further confirmed that "A content provider like CNN has more than a hundred thousand hours of tape in its video archive—far too much for any human to view and annotate manually".

Natural language has widely been used for generating descriptive information as well as assisting information retrieval. However, natural language has significant drawbacks when confronted with large volumes of multimedia documents. The use of natural language in information description is significantly labour intensive (Sikora, 2001) and using of natural language for searching produces problems stemming from too many synonyms, too many meaning and spelling variations (Feldman, 1999). There are two kinds of context in which the use of natural language for descriptions may be not suitable:

- *Descriptions using in qualitative analysis in social psychology or education:* In this kind of analysis, descriptions are used to describe certain aspects of context from collected multimedia materials. The use of natural language causes difficulties for retrieving correct returns.
- *Multimedia descriptions for documents having structures:* In documents that describe multimedia contents, users need to retrieve not only entire the document content but also a single event or scene. In addition, users need to obtain very detailed information such as where, how, and when a single event should appear. When natural language is used for this kind of documents, users may get incorrect returns by using current searching techniques. In

addition, the use of natural language for this type of descriptions is time consuming.

With the requirements for new tools and new languages to handle, organise and retrieve multimedia data, many new languages are built and incorporated with computer applications to assist humans in describing and retrieving information of multimedia documents, as mentioned earlier. These languages are called *constructed languages* and can be presented in two different forms:

- *"Natural language like" languages:* These languages have many similar characteristics compared to natural language. Therefore, they are called "Natural language like" languages. These languages are used for qualitative analysis in social psychology or education and are commonly seen in studying behaviours. Examples of these languages are the Caber language (Patrick, 1985), Observer's language (Noldus Information Technology, 2003) or Flexible Structure Coding Language (FSCL - Heinrich, 1999).
- *Structured languages:* Descriptions of this kind of constructed languages are generated based on specific rules or schemas. The languages are used to describe documents that have structures such as multimedia documents like video clips or films. The most common language used for this purpose is the eXtensible Markup Language (XML - O'reilly Media, 2004).

The general purpose of these languages is to help users to attach notations or descriptions easily to multimedia documents and to retrieve correct information later on. When these descriptions are based on explicitly defined grammars, they can be easily analysed by computer systems. Therefore, the efficiency of searching information in a larger number of multimedia documents can be significantly increased.

To exploit possible advantages of constructed languages in describing and retrieving information, users have to understand the grammars of these constructed languages before using them. Currently, there is a shortage of studies and research that could provide an effective way to help users to understand those constructed languages. Therefore, the research presented in this thesis investigates issues concerning the

development of learning tools to help users understand constructed languages and to facilitate the use of those languages in information description and retrieval.

This introductory chapter starts with a short review on the grammars of natural languages and constructed languages. The chapter then outlines the motivation and objectives of the research for learning constructed language grammars. Next, the chapter presents the methodology used in this research. The structure of this thesis is presented at the end of the chapter.

## 1.1 What is language grammar?

Language grammars can be defined into natural language and constructed language grammars. Main features of these language grammars are presented in the following sections.

### 1.1.1 Natural language grammar

The grammar of a language deals with the issues of how sentences are constructed called syntax. Different languages have different syntaxes. The syntax of a language includes two factors:

*a. The orders of components such as subjects, verbs and objects in a sentence.*

Different natural languages may have different orders of subjects, verbs and other components. For example, the orders of the Toba Batak language (an Austronesian language of Indonesia – Valin, 2001) are syntactically different from English language:

| | | | | |
|---|---|---|---|---|
| The teacher is reading a book. | | | | *in English language* |
| Manjaha | buku | guru | i. | *in Toba Batak language* |
| Read | book | teacher | the | *the order of Toba Batak language compared to English language* |

*b. The combination of words into components.*

A word group may be formed by several words and each word may be in a different lexical category. In the example above, "the teacher" is a word group called the noun group that plays the role of a subject. This word group

includes an *article* category "the" and a *noun* category "book". Different languages have different word orders to form a word group. For example, the word order in a noun group in English and Vietnamese is different:

| | | | |
|---|---|---|---|
| The | red | book | *in English language* |
| Mot | quyen sach | do | *in Vietnamese language* |
| The | book | red | *the order of Vietnamese language compared to English language* |

With English language, there is one more important component called *morphology* that explains how a word is formed. For example, a word "student" has two forms: the singular form "student" and the plural form "students". Figure 1.1 presents relationships among different elements in an English sentence.

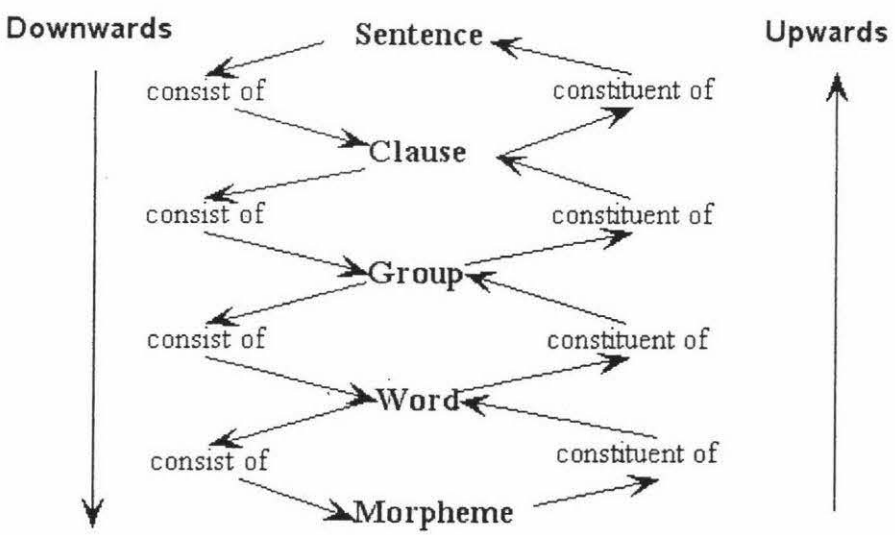


**Figure 1.1 Relationship among elements in an English sentence**

There are many natural languages as discussed above and each of them has specific structures. This research, from now on, will refer to the English language when using the term "natural language". In addition, the words "grammar", "structure", and "syntax" used in the research are equivalent.

### 1.1.2 Constructed language grammar

Constructed languages are created for incorporation with computer systems to help users to describe information easily and retrieve information correctly. In general, these languages' grammars are simplified from natural language or are structured so that the languages are logically precise and computationally tractable, but still human readable. Typical features of these languages can be summarised as the followings:

*"Natural language like" languages (NLL languages):* These languages have both similarities and differences compared to natural language. The differences lie on restricted vocabulary, formulated grammar and restricted structure:

- *Restricted vocabulary:* A restricted vocabulary is a list of words that is agreed on among members. Therefore, vocabulary size and meaning for a particular application domain are limited (Mitamura, 1999). In some languages, users can then organise words into hierarchies as in FSCL (Heinrich, 1999).
- *Formulated grammar:* In NLL languages, their grammars may be formulated explicitly such as LL based grammar in FSCL (Heinrich, 1999) or LR grammar in CABER language (Patrick, 1985). Documents or descriptions are written following rules set by the constructed language grammar. One important purpose of formulated grammars is to improve the performance of the retrieval process.
- *Restricted structure to reduce ambiguity*: In addition with a formulated grammar, NLL languages also employ structures that can reduce ambiguities compared with those of natural language. The reason for this is that "common" natural language structures can cause difficulties for computer application to "understand". For example, computer applications do not process the natural language accurately and efficiently when pronouns are used. Therefore, in NLL languages, pronouns are usually not encouraged (Mitamura, 1999; Pulman, 1996). A learner/user naturally understands what a pronoun stands for in a sentence. However computer applications may not know which noun this pronoun refers to.

*Structured languages:* Structured languages have different structures compared to natural language. These languages can be seen as "markup languages" as in XML.

Users normally have to study how to use these languages. For example, to use XML, users have to learn how to create correct XML documents following defined schemas. They also need to know how to search for information in a XML document. Simple learning tools can be developed that help users to understand basic XML structures or how to create XML schemas.

## 1.2 Motivation and objectives of the research

This section motivates the research by emphasising the importance of learning constructed languages, and by reviewing approaches that support learners in mastering language grammar. Following the motivation, the objectives of the research are outlined.

### 1.2.1 Learning constructed languages

Learning grammar is essential for learning a language. In a traditional sense, learning grammar is "the study of the principles which govern the formation and interpretation of words, phrases and sentences" (Radford, 1997, p1). To emphasise the importance of learning grammar, Lynn (2003, p1) wrote "think of English vocabulary as the bricks of the language and grammar as the mortar that hold those bricks together, without the mortar the bricks can come tumbling down! A lack of grammatical skills can cause embarrassing misunderstandings." However, many people do not want to learn grammar as it as boring and not meaningful. Experience shows that students have to do a lot of grammar exercises to learn a language but these exercises seem to have little or no meaning to them (Onestop Magazine, 2003).

Learning structures of NLL languages is important for users in view of describing and retrieving information. Even there are similarities of structures between natural language and these languages, improper understanding the specifics of these languages' structures can lead users to the problems like the following:

- *Generation of descriptive information that is difficult for querying:* Although there are advantages of NLL languages compared to natural language, they still have limitations. For example, users may create correct sentences following a language structure but it is hard to retrieve these sentences. Therefore, users should be aware of these limitations of NLL languages.

- *Misunderstanding of query engine leads to incorrect returns:* NLL languages usually employ specific query engines for searching for information. When creating improper queries, users may obtain redundant results or fail to detect descriptive sentences or documents that are appropriate to the users' search topic.

Learning structures is also essential for using structured languages, as their structures are totally different compared to natural language. Learning the structures of this kind of languages can be compared to learning the rules of a game. To be able to play a game, players must be familiar with the rules. If players do not know the rules, the game cannot be started or will soon collapse.

Therefore, to fully obtain advantages of constructed languages, users need to know the complete details of those structures. Even though different constructed languages may have different structures, there are some common principles that users are required to learn. The following sections outline important aspects that users should learn of constructed languages.

- *The structure of constructed languages*: Users must understand how elements of a constructed language can be combined into descriptions. For example, in FSCL, users must understand how words in different categories can be combined into word groups and how word groups can be integrated to create a sentence. With XML, users must understand how to create documents that are correct according the XML syntax, Data Type Definition (DTD), or XML schemas.
- *Query-matching mechanisms:* There are many different constructed languages used in many computer systems and many of these languages possess specific query engines. With NLL languages, the understanding of how the languages' query engines work will help users to set up suitable vocabularies and create appropriate descriptions to get best matches between queries and descriptions. With structured languages, users need to create correct queries to retrieve accurate returns.

There are several studies that focus on assisting users in generating correct descriptions using constructed languages. For example, there are some editor tools

that help users to check the compliance of user input with the structure rules and assist them to write accurate descriptions (Schwitter et. al., 2003). In these editors, users have to stick with the predefined vocabulary and grammar rules to generate informative and consistent documents. However, these editors cannot help users to understand the query mechanisms that are embedded in many constructed languages. As a result, users may not discover the full potential of constructed languages for information retrieval.

To address these issues, the research presented here focuses on developing concepts to be applied in computer applications for teaching users constructed languages. The research then implements those concepts in a teaching tool for teaching a specific constructed language, FSCL, as a case study.

### 1.2.2 Factors that enhance understanding in language learning

There are several factors that can enhance users' understanding in language learning. The factors discussed in this section are: learning by using diagrams and computers, learning by example, and active learning.

#### *Diagrams in language learning*

Many authors of books or research on language theories use diagrams to present sentences structures. The benefits of diagrams in presenting language structures are, according to Ye Hedge School (2003), "Diagrams are unquestionably the best way to map the word relationships within a sentence; only when we can understand how words are related in a sentence can we study complex thought". Advantages of diagrams in enhancing users' understanding of language structures will be presented in more details in the following chapter.

#### *Computer in language learning*

Books and classroom tuition are traditional ways for learners to learn grammar. However, the invention of computers has leaded to an innovative method to teach language structures. Computer applications, called Computer Assisted Language Learning (CALL), offer powerful self-assessment facilities to learners. CALL can help learners to study many areas of languages such as grammar, vocabulary, reading and writing. As diagrams can be easily drawn by computer applications, they are used

by many CALL applications as supporting tools to help users to learn natural languages. However, there are limitations of current research in exploring the advantages of diagrams in CALL applications for teaching *constructed language structures* and their specific query engines.

**Learning by examples**

Perry (1993) pointed out that command descriptions, syntax format and language references are not enough for new learners to learn a new language. He suggested that numerous examples could provide a better method to help users to understand the language. Addressing the same issue, Walsh (2001) emphasised that computer concepts are best taught with multiple examples and that using a variety of practical and real world examples can lead learners step-by-step on the road of learning. Pérez (2004) claimed that humans have great ability to learn concepts from a set of examples. Stauffer (2004) also concluded that learning by example allows learners to start with initial concepts (what they have known already) then understand new concepts (what they do not know yet) easily.

**Active learning**

Many researchers suggested that a passive learning environment in which students are involved only passively in learning, i.e., in listening to the instructor or reading the text book, generally leads to a limited retention of knowledge (McKeachie, 1998). In contrast, in active leaning, learners can obtain better knowledge as they not only merely listen or view the presentation of information but also work with the content and reflect on the process of learning (Oppermann, 2004). In an active learning environment, learners "are simply more likely to internalise, understand, and remember material learnt through active engagement in the learning process" (BYU Faculty Center, 2000, p.1). Active learning takes advantages of peoples' natural motivations, abilities and interests to solve problems, discover relationships and make comparisons. To learn language grammar, Eastwood (1999) claimed that the best way is to follow the active learning approach.

### 1.2.3 Research objectives

Following the motivations above, the main objectives of this research are:

- Investigate the advantages of using language structure diagrams for teaching *natural* languages.
- Identify the strategies of CALL applications to teach *natural* languages.
- Adjust the use of language structure diagrams and CALL strategies to the teaching of *constructed* languages and their queries engines.
- Implement these resulting concepts in a language-teaching tool for a specific constructed language.

### 1.3 Research methodology and steps

The research was carried out based on popular frameworks built by Galliers (1992) and Yin (1994). Using these frameworks, a research methodology that is suitable for fulfilling the stated research objectives has been developed:

*Literature review*: Examine diagramming techniques and CALL applications that assist language learning;

*Concept development*: Based on the literature review, develop concepts for teaching constructed languages;

*Prototype implementation for a case study*: Apply these concepts in an application for teaching a specific constructed language, FSCL;

*Experiments to verify the concepts* (due to time restrictions outside the scope of this thesis).

Applying this methodology leads to the following steps:

- Literature review, including:
  - Examining diagramming techniques using in natural language theories and constructed languages: The project will study how the diagramming techniques can support users in understanding language structures.
  - Investigating CALL applications for teaching natural language structures: The research will study features of CALL applications and their advantages for teaching constructed languages.
  - Inspecting in detail an existing constructed language, FSCL: The research examines the language structures, its structure diagrams and the query engine. This language will be used as a case study.

- Developing concepts to teach learners constructed languages using diagrams. These learning concepts are developed based on/adjusted from teaching concepts learnt from the literature review. The research employs these concepts for teaching FSCL and then outlines how those concepts can be adapted to teach other constructed languages.
- Implementing those concepts into a prototype application called Visualised Interactive Structure Learning (VISL) to teach FSCL. Specific technical issues relating to the development of this application will be presented in detail. The research will apply software engineering concepts using the Unified Modelling Language for analysis and design, and the Java programming language for implementation. VISL's interface will be developed based on principles of Human-Computer-Interaction (HCI).

### 1.4 Thesis structure

The thesis is divided into seven chapters. After the introduction in Chapter 1, Chapter 2 reviews the importance of diagrams in enhancing users' understanding of language structures. It also inspects typical CALL applications for teaching language structures. Chapter 3 examines a constructed language, FSCL, in detail. In Chapter 4, concepts for teaching constructed languages are developed. The chapter also illustrates how these concepts can be applied for teaching FSCL. Chapter 5 discusses the key issues relating to the development of a prototype CALL application called VISL for teaching FSCL. Chapter 6 presents the VISL development process through the three phases of analysis, design and implementation. Finally, Chapter 7 of the thesis summaries the contributions and highlights future topics arising from this research. The diagram in Figure 1.2 illustrates this structure of the thesis.

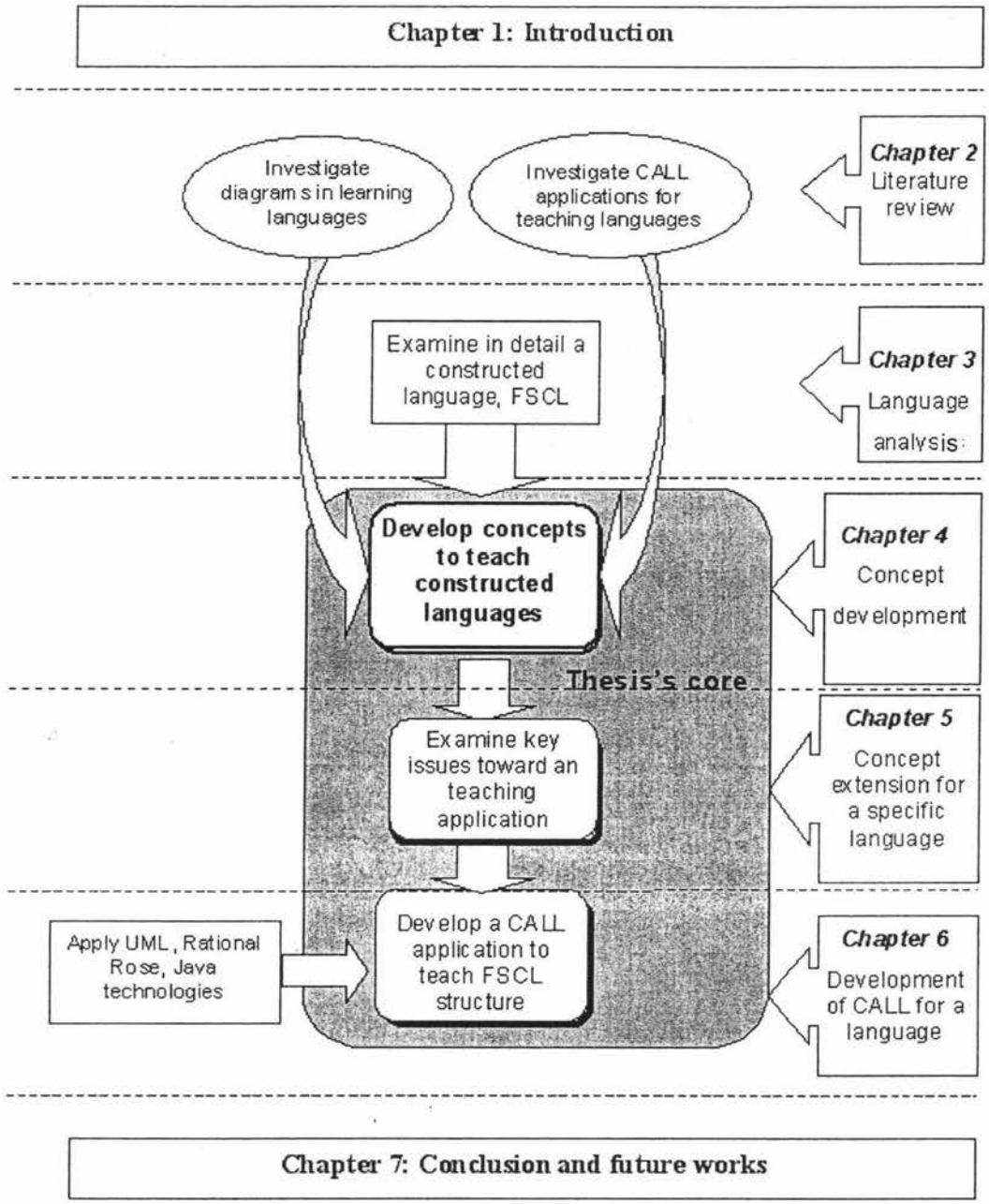


**Figure 1.2 Thesis's structure diagram**