University of Groningen

A document enrichment approach to facilitate reading comprehension

Olango, Proscovia

DOI:
[10.33612/diss.111697052](https://doi.org/10.33612/diss.111697052)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2020

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Olango, P. (2020). *A document enrichment approach to facilitate reading comprehension*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen, SOM research school.
<https://doi.org/10.33612/diss.111697052>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

A Document Enrichment Approach to Facilitate Reading Comprehension

Proscovia Olango

Publisher: University of Groningen, Groningen, The Netherlands

Printed by: Ipskamp Printing, Enschede, The Netherlands

ISBN: 978-94-034-2148-3 / 978-94-034-2147-6 (ebook)

© Proscovia Olango

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system of any nature, or transmitted in any form of by any means, electronic, mechanical, now known or hereafter invented, including photocopying or recording, without written permission of the publisher.

university of
 groningen

A Document Enrichment Approach to Facilitate Reading Comprehension

PhD Thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. C. Wijmenga
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Thursday 16 January 2020, at 12.45 hours

by

Proscovia Olango

born on 22 November 1973
in Gulu, Uganda

Supervisor(s)

Prof. H.G. Sol

Prof. J. Nerbonne

Assessment committee

Prof. J. Lubega

Prof. G.B. Huitema

Prof. F. Zwarts

To Abalo, Clement and Joel, with a hope that you will all awaken in the new earth.

Preface and Acknowledgements

University students have time-bound assignments which require a lot of reading, thus it is vital for them to quickly understand the content of documents provided for such assignments. The students' reading process can be hindered by factors related to inadequate reading comprehension skills. Such skills include insufficient vocabulary and background knowledge, as indicated by a stakeholders' survey carried out during this research. Such factors create gaps in the students' reading process. Causing them to look for information about difficult terms and concepts from reference works outside the content of their course documents. This disrupts the students' reading process and costs time. This research was therefore motivated by the need to keep students reading the content of their course documents by providing in-text definitions and background knowledge for difficult technical terms and concepts. The stakeholders' survey indicated that providing easy access to such knowledge would facilitate the reading comprehension process. Thus, realizing the idea of applying language technology in a design science research methodology for facilitating the reading comprehension process. TermPedia, a document enrichment approach for facilitating reading comprehension was proposed and its algorithms were designed and tested. The approach was then evaluated for its usefulness and usability by using students from the University of Groningen and Gulu University. The document enrichment approach makes contribution to both theory and practice of facilitating reading comprehension.

The intricate journey of carrying out this PhD research would not have been possible without the invaluable support of my supervisors, colleagues, family and friends. I am not able to mention everyone who contributed to the success of this research by name, but know that I am much obliged to you individually.

My sincere gratitude goes to my supervisors, prof.dr.ir. John Nerbonne and prof.dr. Henk G. Sol. Erudite prof. Nerbonne, thank you for the articulate guidance provided. I greatly benefited from the clear comments and suggestions on research ideas.

The discussions we had during scheduled meetings and lunch breaks were definitely encouraging eye openers. Thank you for opening your home to me while I was in the Netherlands during holiday seasons. These interactions gave me confidence to live and work in the Netherlands. I am most especially grateful that you did not give up on me.

Highly learned prof. Sol, thank you for picking me from the *lost and found* by taking interest in my research. It has been an exiting learning experience and your academic excellence made this PhD worthwhile and achievable. I particularly benefited from the guidance you gave as we explored the possibility of applying design science research to facilitating reading comprehension. The support and commitment you offered to this PhD research is invaluable. *Dank u zeer.*

Heartfelt thanks go to all members of staff of the Center for Language and Cognition Groningen (CLCG) for their cooperation. Thank you dr. Gertjan van Noord, dr. Elwin Koster, dr. Leonie Bosveld-de Smet, dr. Peter Kleiweg and dr. Henny E. Klein. Special gratitude goes to dr. Gosse Bouma who worked with me during the initial stages of my research. Sincere gratitude is extended to the people at the International Bureau who took care of all logistics during my stay in Groningen. Thank you Erik Haarbrink, Gonny Lakerveld and Marieke Farchi. Sincere thanks also go to the management of Gulu University. I am especially grateful to the former Vice Chancellor prof. dr. Jack. H. Nyeko Pen-Mogi and the current Vice Chancellor prof. dr. George L. Openjuru for their genuine support, patience and encouragement.

A pat on my back by dr. Mercy Amio with the words; “*You can do it. Be confident, it is your work*”. Such affection picked me from my lowest point and provided the needed courage not to give up. I would like to acknowledge Mercy together with dr. Hasifah K. Namatovu for coaching me on the application of design science research principles. Hasifah and Mercy, thank you for having confidence in me and for providing the academic and social support that I needed to keep working. I am also indebted to dr. Fridah Katushemererwe who diligently read my work with special attention to grammar and diction. Thank you too for the moral support extended to me in both Groningen and Kampala.

I would also like to acknowledge the support extended to me by the people who worked in CLCG at the time I was there. Thank you Geoffrey Andogah, Martijn Wieling, Peter Nabbende, Dörte and Daniël de Kok, to mention but a few. To my colleagues at the Department of Computer Science, Gulu University, thank you for shouldering my lecture loads during the time when my academics demanded high concentration. I am grateful for the patience and unity that exists in this department. It was also an honour building interactions with accomplished researchers such as dr. Deborah Mudali, dr. Drake P. Mirembe and dr. Robert Tweheyo.

Debbie, thank you too for your concern and the delicious meals you prepared during Ugandan dinners that made me feel at home in Groningen.

This thesis could not have been accomplished without the responses from university students who participated in providing data for the surveys and evaluations carried out during this research. Thank you for taking time to provide helpful information.

I acknowledge the support extended to me by my family members. Deep gratitude goes to my father and friend Mr. Clement Olango. *Baba*, thank you for being a close companion in my academic career. This PhD is yours as much as it is mine, because it was made achievable by your encouragement, trust and support. Your interest in my education was shown when you used to personally help me with my primary school homework. Thank you for giving me equal opportunity to education and for marking sure I was in the best schools as far as your resources could allow. This gave me the solid academic foundation needed for my PhD. I am proud to be your daughter. *Maa*, Mrs. Regina Olango Alaroker, you have been my excellent role model of hard work, care and generosity. By working hard to give me a healthy body you gave me a healthy mind for academics. *Apwoyo Mama*. Uncle David Okeny Ojok, I deeply appreciate the comfort, courage and advice you gave to me during the time I lost the laptop computer which contained all my research works. To my siblings: Pamela, Alphonse, Ivy, Celine and Lisa, thank you all for the unconditional love and moral support that you extended to me during this exciting academic journey. Your words of encouragement and humour lightened my spirit and gave me the strength to work diligently. I will forever hold dear those heartfelt smiles, hugs and kisses that made me feel worthwhile.

Glory be to Jehovah, the Almighty God who comforted me through the world wide brotherhood of Jehovah's Witnesses. Especially through the presence of Abalober, Alphonse, Anna, Faith, Like, Maureen, Paul, Sarah, Tini, and Vicky. Thank you all for the earnest concerns you showed during my deep sorrows. This made me rejoice. I count you friends for life eternal.

Proscovia Olango
Gulu

Contents

Preface and Acknowledgements	i
List of Tables	x
List of Figures	xii
1 Understanding Reading Comprehension	1
1.1 Practical Issues in Reading Comprehension	1
1.2 Models of Reading Comprehension	5
1.3 Motivation and Problem Statement	14
1.4 Text Wikification	16
1.5 Document Enrichment	19
1.6 Research Questions	21
1.7 Research Approach	22
1.8 Thesis Outline	27
2 Dealing with Technical Terms	29
2.1 Issues in Dealing with Technical Terms	29
2.2 Recognizing Technical Terms	30
2.3 Technical Terms' Sense Disambiguation	33
2.4 Defining Technical Terms	36
2.5 Background Information for Technical Terms	38
2.6 Visual Clues for Technical Terms	39
3 The Need for Document Enrichment	43
3.1 Exploratory Survey Objectives and Approach	43
3.2 Presentation and Discussion of Results	45
3.3 Key Processes in Reading Comprehension	47
3.4 Key Factors Hindering Reading at Universities	50

3.5	General Effects of Reading Difficulties	58
3.6	Reading Process Facilitation: Respondents' View	61
3.7	Document Enrichment Requirements	63
4	Designing TermPedia's Algorithms	69
4.1	Overview of TermPedia's Approach	69
4.2	Document Enrichment Algorithms Design	72
4.3	Data Used in Designing the Algorithms	75
4.4	Technical Terms List Generation Criteria	81
4.5	Technical Term Sense Disambiguation Criteria	85
4.6	TermPedia's Document Enrichment Algorithms	86
5	TermPedia: The DE Approach	91
5.1	Ways of Framework	91
5.2	Way of Thinking	91
5.3	Way of Governance	94
5.4	Way of Working	95
5.5	Way of Modelling	100
5.6	Way of Supporting	101
6	Testing TermPedia's Algorithms	103
6.1	Objectives of Testing TermPedia's Algorithms	103
6.2	Methods of Testing the Algorithms	104
6.3	Description of Test Data	109
6.4	Presentation of Test Results	113
7	Evaluation of TermPedia	133
7.1	Objectives for Evaluating TermPedia	133
7.2	TermPedia User Study at UG	134
7.3	TermPedia User Study at Gulu University	140
7.4	Comparative Discussion of Evaluation Results	158
8	Epilogue	161
8.1	Thesis Overview	161
8.2	Reflection on the Thesis Approach	166
8.3	Research Contribution	168
8.4	Generalizability of TermPedia	171
8.5	Direction for Future Research	173
	References	175

A TermPedia Evaluation Tools and Data	195
A.1 Traffic to TermPedia User Interface	195
A.2 TermPedia User Survey Questionnaire for RUG.....	196
A.3 r and α of Mean Scores for TermPedia User Survey.....	198
A.4 Reactions to TermPedia by RUG Students	200
A.5 Test Scores of Gulu University Students	201
A.6 TermPedia User Study Questionnaire for GU.....	203
A.7 Summary Scores for GU TermPedia User Study	205
A.8 Comments and Suggestions for TermPedia by GU Students	207
List of Abbreviations and Acronyms.....	209
Summary.....	211
Samenvatting	215
Curriculum Vitae	219

List of Tables

3.1	Biographies of the respondents	46
4.1	Important files generated by WikipediaMiner	75
4.2	Possible ambiguous concepts of the term <i>temporal</i>	76
4.3	Summary statistics for the English Wikipedia dump of 2010	78
4.4	Tuples and values from generated database for wikilink labels.....	79
4.5	Tuples and values from database of unambiguous wikilink labels.....	80
4.6	Examples of stop words that are used as wikilink labels	82
4.7	Sample stop word list for Wikipedia documents	84
6.1	Sample titles of disease related articles extracted from the Wikipedia dump	110
6.2	Term prediction by BTP, LTP, and FTP from example sentence 6.4.1	114
6.3	Over all testing of Wikipedia data technical term prediction	116
6.4	Statistics of terms predicted by LTP against gold set of EMA.....	122
6.5	Statistics of terms predicted by LTP against gold set of Merck.....	124
6.6	Example of target selection for a predicted ambiguous technical term...	126
6.7	Over all testing of Wikipedia data term target selection.....	128
7.1	Summary visits between Oct 19 and Nov 19, 2010 to TermPedia user interface from Google Analytics	136
7.2	Ranks of answers to the study questionnaire used at UG	137
7.3	Cronbach's alpha reliability coefficient for RUG questionnaire scores.....	138
7.4	Summary of students' closed-book test scores before they used TermPe- dia	142
7.5	Summary of students open-book test scores after using TermPedia.....	143

7.6	Summary of students open-book scores with outliers ignored	146
7.7	Table showing summary statistics of time spent by students on open-book test	147
7.8	Ranks for Gulu University questionnaire responses.....	148
7.9	Summary scores for all the 22 students as ranked for sections A, B, and C.....	148
7.10	Summary for user scores to questions 3, 5, 8, and 9 in section C	150
7.11	Summary for user scores to questions 6 and 7 in section C	151
7.12	Summary for user scores to questions 1, 2, and 10 in section C.....	151
7.13	Summary for user scores to questions in section B of Gulu user study questionnaire.....	153
7.14	Table showing summary statistics of scores given to questions GA1 and GA5.....	154
7.15	Table showing summary statistics of mean scores for questions 2, 3, and 4 in section A of Gulu user study questionnaire.....	154
A.1	Users comments and suggestions to improve TermPedia user interface and techniques	200
A.2	Students closed-book test scores before TermPedia.....	201
A.3	Students open-book test scores after using TermPedia	202
A.4	Sections A, B, and C summary scores for Gulu University user study questionnaire.....	205
A.5	Users comments and suggestions to improve TermPedia user interface and techniques.....	207

List of Figures

1.1	Illustration of the automatic text wikification process	16
1.2	Screen shot from https://en.wikipedia.org/wiki/Transistor , showing the richness of information from Wikipedia.....	21
1.3	The three-cycle view of Design Science Research [Hevner, 2007]	24
1.4	Research strategy, adapted from Gonzalez and Sol [2012]; Sol [1982]	25
3.1	A screen shot of the text wikification system architecture taken from from Mihalcea and Csomai [2007].....	67
4.1	TermPedia Document Enrichment System Architecture; Modified from Manning et al. [2008]’s Text Wikification System Architecture.	70
4.2	Zipf’s plot for ranked inverse word frequencies in Wikipedia.....	83
5.1	Sol [1988]’s Framework for Understanding Design Approaches	92
5.2	The TermPedia Document Enrichment Approach Use-Case Diagram ...	95
5.3	Steps in Applying TermPedia for Facilitating Reading Comprehension	97
5.4	Document enrichment with Wikipedia text.....	99
5.5	Document enrichment with Wikipedia text and images.....	99
5.6	TermPedia’s document enrichment approach sequence diagram.....	100
1.1	Graph showing the effect of reducing the percentage of n-grams permitted to be predicted as technical terms on precision, recall and f-score while using the LTP algorithm	117
2.2	Permitted percentage of terms to be predicted for each document	119
3.3	Upper limits of term likelihood probability for terms to be predicted ...	120
7.1	Screen Sorts of TermPedia User Interface	135

7.2 Snap shot of TermPedia user interface 143

7.3 Box plots for open-book test scores of students 144

7.4 Normal Q-Q plot of open-book test scores for students who were not by
helped TermPedia.....145

7.5 Bar plot for frequencies of ranked responses to questions 1, and 5 in section
A of the user study questionnaire 155

Chapter 1

Understanding Reading Comprehension

Neither comprehension nor
learning can take place in an
atmosphere of anxiety.

Rose Kennedy

This chapter opens by discussing practical issues in reading comprehension and shows that these issues affect some university students and other people in general. The chapter also considers models of reading comprehension in an effort to understand what goes on during the process of reading. The process of reading comprehension can be affected by insufficient vocabulary and background knowledge and poor clues to text. A possibility of enhancing reading comprehension points to a document enrichment approach. The chapter also stipulates the research motivation, problem, questions and approach. It closes by providing the thesis outline.

1.1 Practical Issues in Reading Comprehension

Each year hundreds and thousands of students in Uganda and many other countries around the world join universities. These students bring with them diverse language and background skills, but their common interest is to learn. Learning in an environment of people with diverse skills is an old phenomenon. It is a venture that can be challenging to students. One clear challenge is that learners have to read and comprehend large volumes of course content in a short time [Hoeft, 2012; Fairbairn and Fairbairn, 2001]. Document content comprehension is an important component of reading without which the goal for learning from documents would be futile. Therefore, students at university level need to have strong comprehension skills in order to quickly understand the content they read. In addition, comprehension skills help students to develop good qualities in decoding text, analyzing, explaining and expressing their own ideas about written materials [Gilakjani and Ahmadi, 2011]. Such qualities are important for furthering academic development.

Different reading comprehension skills such as fluency, lexical knowledge, and preexisting knowledge need to be applied concurrently and expeditiously in the process of reading so that there is comprehension [Abbas and Narjes, 2016]. However, there is evidence that a number of university students have low proficiency in reading comprehension because of poor reading skills [Unruh, 2015; Shafie and Nayan, 2011; Roehl and Shiue, 2003]. As a result, such students struggle to understand what they read, which makes the reading experience anything but easy or cost-effective. The struggle is escalated by unfamiliar discipline-specific texts [Hobson, 2004] and topics. Students, especially those in their first year of study lack skills of working with discipline-specific texts. This can be frustrating and often leads students to give up on their readings [Schwartz, 2012]. Students who find their course text difficult to read also find problems in preparing for exams [Pressley et al., 1997], and this may cause them to fail or dropout [Moodley and Singh, 2015; Fiester, 2010]. Those who struggle with reading comprehension throughout their university education may become graduates who have not deeply internalized their course material, making them neither confident nor productive graduates.

The lexical knowledge skill of reading, also known as vocabulary knowledge is listed in many works on reading comprehension as an important factor influencing document content comprehension [Abbas and Narjes, 2016; Palmer et al., 2008; Graves and Graves, 2003; Nagy, 1988]. Students may have to consult external reference resources like dictionaries and encyclopedias in order to comprehend vocabulary they cannot recognize. This costs time and effort and may dampen the reading mood. Just like vocabulary knowledge, the other reading comprehension skills may also be underdeveloped in some university students especially for those who use English as a second language [Shehu, 2015; Hendricks, 2013; Fender, 2008], in cases where English is the official language at their universities. Our research does not directly consider problems of reading related to English as a second language but we use this reality to support the notion that some students at universities struggle with reading comprehension. Nevertheless, the language of interest to our research is English because it is the official language in Uganda. Another reason why we are interested in working with the English language is because most of the university textbooks and course materials are written in English.

Compared with other languages, English has an extremely large vocabulary [Jennings et al., 2014]. Moreover, the English language just like any other language constantly expands its vocabulary by borrowing, coining, and combining words to represent new ideas, (technology) and development [Engineer, 2005]. For example, the noun *fitspiration* was recently added to the English language and it refers to a person or thing that serves as motivation for someone to sustain or improve health and fitness [Oxford Living Dictionaries, 2017]. Example 1.1.1 shows another case in which

language change introduced three new domain-specific vocabulary. These vocabularies were borrowed or coined in the recent years to express ideas related to Short Message Services (SMS) and Internet technology. For a person who is not familiar with these technologies the semantics of the terms *twitter* and *tweet* may be conspicuously revealed by considering them in context. On the contrary, the meaning of *microblogging* may not be that obvious from this sentence context.

Example 1.1.1: Language change

In 2009 the noun *twitter* was borrowed as trademark of a social network that provides *microblogging* services, enabling its users to send and receive messages called *tweets*.

The large variety of vocabulary and the fact that this vocabulary grows every year makes mastering English vocabulary a lifelong task. This means that the problem of reading comprehension caused by insufficient vocabulary knowledge is general and does not only affects university students but everyone. When people are unable to recognize and understand words, their reading efficiency and text comprehension are affected. That in turn impedes effective learning among post-secondary learners and affects their opportunities for competitive employment [Barth et al., 2016].

Apart from ill-developed reading skills, sometimes document contents are difficult to understand. Some factors that influence content difficulty include sentence structure, length and author elaboration [Graves and Graves, 2003]. Reading is slowed to a near halt and comprehension is seriously compromised when proficient adult readers struggle with “difficult” contents [McNamara, 2007]. This makes reading a dreadful and time-consuming experience and yet we may have to read such documents for one reason or another.

Example 1.1.2: Difficult Sentence structure¹

“For purposes of paragraph (3), an organization described in paragraph (2) shall be deemed to include an organization described in section 501(c)(4), (5), or (6) which would be described in paragraph (2) if it were an organization described in section 501(c)(3).”

For instance, Americans have to read through obtuse legal documents that have difficult sentence structures as shown in example 1.1.2, during their annual tax-filing

¹Sentence was taken from <https://www.economist.com/blogs/johnson/2011/04/legal-language> on 23-November-2017

sessions. This sentence is a tax code from section 509(a) of America's book of tax law. Much as the sentence is confusing because of the cross references, it would be absurdly long [Johnson, 2011] if the various referred sections were actually written out. Such writings need to be simplified so that people can easily understand them. Elaborate text facilitates reading comprehension as illustrated in example 1.1.3.

In this example the author clearly defines what text *tokenization* means and conveniently includes a graphic support for his definition. The graphical support helps the reader to visualize the *tokenization* process and create a mental impression of it. Both the text and graphic support assist the reader in understanding, remembering and applying the concept of *tokenization*. Such elaborate text provides sufficient clues to text that facilitate reading comprehension and knowledge retention and application.

Example 1.1.3: Elaborate text²

Tokenization

Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called *tokens*, perhaps at the same time throwing away certain characters, such as punctuation. Here is an example of tokenization:

Input: Friends, Romans, Countrymen, lend me your ears;

Output:

Friends

Romans

Countrymen

lend

me

your

ears

Example 1.1.4: Text which may not be elaborate, but may need further support (clues)

The term *unstructured data* refers to data which does not have clear, *semantically* overt, easy-for-a-computer structure. It is the opposite of *structured data*, the canonical example of which is a *relational database*, of the sort companies usually use to maintain *product inventories* and *personnel records*. [Manning et al., 2009]

If we compare example 1.1.3 and example 1.1.4, we see that the latter may not be elaborate enough for understanding the meaning of *unstructured data*. Much as the author defines *unstructured data*, he refers to other concepts in his definition like *text semantics*, which are easily read by computers, *structured data* and *relational databases*. The reader has to have background knowledge about these referred concepts in order to understand what *unstructured data* means.

²Screen shot taken from Manning et al. [2009]

Example 1.1.4 illustrates text that may not provide sufficient clues for facilitating reading comprehension because the author uses domain-specific concepts in his definition without graphical support.

In this digital world university students normally read electronic documents and, when they interact with text that they find difficult to understand for one reason or the other, they preferably consult other electronic resources on the World Wide Web to help them understand. These resources are preferred because they provide background information in various knowledge domains, vocabulary definitions and graphical help for understanding difficult text. Other advantages of the electronic resources are access opportunities, which are absent in the print resources, which include remote access, 24-hour access, and multiple users for single sources [Liu, 2006]. The essential advantage of electronic resources for our research is that they can be used to enhance the comprehension of difficult electronic resources by use of computer software. Our interest is to work with electronic resources for university education. We are interested in investigating ways of applying computer software to enhance comprehension of electronic course content by university students. The reason for this investigation is because electronic course content is widely used by students and they contain technical terms that hinder the students' comprehension. Reading comprehension is a concept that goes beyond vocabulary knowledge and accessible documents. The skills of reading comprehension indicate that there are cognitive aspects to it, including a systematic way of acquiring comprehension skills. Various reading models have been developed in an effort to understand what goes on during the process of reading comprehension. Understanding the reading comprehension process throws light on how to improve it, even by innovative development through application of computer software. In the following section, we focus on four models of reading comprehension namely: *bottom-up*, *top-down*, *interactive* and *situation* models, because they are proposed to account for the comprehension process [An, 2013]. These models can be applied to people reading at all levels since reading comprehension is a lifelong activity [IRA, 1990; Horton et al., 2015; K12Reader, 2016].

1.2 Models of Reading Comprehension

Before we discuss the models of reading comprehension, we will first explain what reading comprehension means. Woolley [2011] defines reading comprehension as the process of making meaning from written text. This definition carries the idea that reading comprehension requires a coordination of text with context in a way that goes far beyond simply chaining together the meanings of a string of decoded words [Spiro et al., 1980]. However, to make meaning from the entire text, one must visually

process the individual words, identify and access their phonological, orthographic, and semantic representations, and connect these representations to form an understanding of the underlying meaning [Hoover and Gough, 1990; Kendeou et al., 2014].

A similar definition of reading comprehension is given by Snow [2002], as the process of simultaneously extracting and constructing meaning through interaction and involvement with written language. This definition highlights three important key features of comprehension: (i) the accurate decoding of written language, (ii) a process of meaning construction through, which inferences and information not available from the text are incorporated into the meaning representation, and (iii) active motivated engagement from the reader [Snow, 2010]. This suggests that reading comprehension stems from a reader's ability to efficiently integrate previously acquired knowledge with information provided in written language [Duffy et al., 1984] and the writer's intent.

To illustrate the importance of prior knowledge, consider Example (1.2.1). The sentence in this example requires prior information on concepts of *Elagolix*, *GnRH* and *estradiol* for its comprehension. It is common to find this kind of sentences in medical journals, for which insufficient prior knowledge in the specific medical domain will lead to failure in comprehension. In such cases, reading comprehension is failed because its features are affected.

Example 1.2.1: Importance of Prior Knowledge

Elagolix is a novel, oral *GnRH* antagonist that dose-dependently suppresses *estradiol* levels [Ács et al., 2015].

For instance, the written language will not be decoded accurately because of lack of understanding of the medical terms. More to that, meaning cannot be constructed because there is no prior knowledge from which to infer it. Consequently, active involvement of the reader would be affected because of the “empty” concepts encountered. As such, reading comprehension adequately often is a major basis for success or failure from kindergarten to college and throughout professional life [Rayner and Reichle, 2010; Onwuegbuzie and Collins, 2002] depending on the relevance of a reader's prior knowledge to specific text.

These definitions render reading comprehension as a complex interaction among automatic and strategic cognitive skills and processes, which enable the reader to create a mental representation of the text [Moore, 2016]. Many difficult occasions arise that lead to reading comprehension failure [Woolley, 2011] because it is a complex process. The complexity of reading comprehension is captured in the following models which, describe the cognitive and linguistic processes it involves.

Bottom-Up Model

The bottom-up model considers reading as a text-driven decoding process where the reader reconstructs meaning from the smallest units of text [Hameed, 2008]. Text decoding generates a reader's cognitive activity [Cruz and Escudero, 2012] and it is presented as a sequential process. Steps in the sequence include visual focus on the identification of letters, noticing combination of the letters, recognition of words from letter combinations, establishing sentences via their syntactic structures and finally integrating sentences into coherent discourse [Shahnazari and Dabaghi, 2014; Davoudi and Moghadam, 2015]. Until the meaning of text being read is eventually determined. The model therefore renders word recognition as an important role in reading comprehension [Goodman, 1967]. This is why inability to recognize words is said to hinder reading comprehension [Perfetti, 1985; Stanovich, 1988; William, 1988; Rauha, 2011].

However, the sequential bottom-up approach does not bear itself out in actual practice [Hameed, 2008] since the model assumes that there is strictly a single meaning of text coded by the writer and the reader needs to decode this meaning without going beyond it. Therefore, it is not possible to make use of higher-order reading skills such as making inferences. Consequently, a reader's background knowledge plays virtually no role in deriving and interpreting the meaning of text, in this model [Shahnazari and Dabaghi, 2014].

Top-Down Model

The top-down model of reading comprehension was presented by Goodman[1967]. He suggests that the reader's preconceptions and background knowledge largely impact the lower-level processes of reading such as orthographic and phonological processing, as well as the word recognition skill. Goodman's top-down model views reading as a psycholinguistic guessing game, entailing a cycled sequence of cognitive processes used by the human brain for text processing. These processes were identified as predicting the meaning of text content, confirming the true predictions, correcting the false predictions and finally terminating these processes [Davoudi and Moghadam, 2015] when comprehension is achieved. The model focuses on what readers bring to written text, which helps them to understand the text meaning [Shahnazari and Dabaghi, 2014]. It presumes that the reading process is not guided by decoding of letters and parsing of syntax and semantics, but by a reader's background knowledge and expectations [Cruz and Escudero, 2012].

In the top-down model, reading comprehension is not just a process of extracting meaning from text as presented by the bottom-up model. On the contrary, it is a process of connecting information in the text being read to the background knowledge

of a reader [Grabe, 1988]. Consequently, the top-down model renders background knowledge essential for successful reading comprehension [Chervenick, 1992; Wanja, 2012]. The model suggests that the reader needs to possess adequate and relevant prior knowledge on specific subjects in order for them to understand the content of documents which discuss such subjects.

Interactive Model

The comprehension process is sequential in both the bottom-up and top-down models of reading. For this reason, the bottom-up model does not allow for higher-level processing strategies to influence lower-level processing, and the top-down model does not account for the situation in which a reader has little knowledge of a text topic. Therefore, such a reader cannot form predictions [Liu, 2010]. To resolve these discrepancies Rumelhart [1977] proposed an interactive model of reading comprehension. This model is based on the concept that meaning does not reside in the text alone, but is a co-construction of the writers' text and the readers' interpretation [Davoudi and Moghadam, 2015]. The model recognizes the interaction between the bottom-up and top-down processes simultaneously throughout the process of reading comprehension [Andres, 2014]. Reading is viewed as an active process that depends on the text, the reader characteristics and the reading situation where information from several knowledge sources are considered simultaneously. Examples of the knowledge sources may be given as letter-sound relationship, word meanings and event sequence. Every reader brings a multitude of skills and knowledge to the task: decoding skills, word-recognition skills, vocabulary knowledge, knowledge of grammatical structures, and conceptual abilities [Hameed, 2008].

The interactive model fundamentally promotes the development of theories in reading, especially the schema theory [An, 2013]. Schema theory views reading as an interactive process. The interaction occurs at three stages: between the bottom-up and the top-down processing, between the high level and lower level reading skills and between the readers background knowledge and the background knowledge presented in the text. Reutzel and Cooter Jr. [2015] present the schema theory as a hypothesis that explains how the knowledge, we have stored in our minds help us to understand and gain new knowledge. They define the term schema as a kind of organized storage cabinet in our brains with file folders containing different information about concepts, events, emotions, and roles drawn from our life experiences. Therefore, a schema is organized knowledge that a reader already has [Gunning, 1996] stored in his mind. Each schema connects to other related schemata forming a large network of interconnected knowledge and experiences [Talanquer, 2013]. The schema theory suggests that prior knowledge and experiences are essential in helping a reader to

comprehend text information as well as to make important inferences about characters in the text [Woolley, 2011; Zwaan and Radvansky, 1998].

The process of successful reading comprehension is guided by the principle that information put into the mind of a reader is mapped against some existing schema and that all aspects of the schema must be compatible with that information [Shen, 2008; Zhang, 2010]. This principle highlights the importance of prior knowledge, revealing that new information, concepts and ideas only have meaning when they are related to something a reader already knows. However, in a situation where new information or experience cannot fully be accommodated in an existing schema, the schema must expand to become consistent with the new concept [Adekoya, 2013]. If the new concept cannot be accommodated by expanding an existing schema then a new schema is created for it.

The schema theory shows that reading comprehension is a complex process of simultaneously applying multiple cognitive skills like fluency, term and semantic processing, working memory and reasoning and inference [Katz et al., 2012; Maulizan, 2015; Moore, 2016]. The cognitive strategy of successfully applying these skills lies in the ability of the reader to activate prior knowledge for effective comprehension [Moore, 2016]. Prior knowledge is especially needed for the cognitive skill of reasoning and inference.

Situation Models

Beside prior knowledge, Zhang [2010] reports that reading comprehension may be affected if the writer does not provide sufficient clues in the text for the readers to effectively use their information processing skills to activate their prior knowledge. Situation models, also called mental models, are complex mental representations that can simulate a situation described in text [Radvansky and Copeland, 2001]. Simply stated, situation models are integrated mental representations of a described state of affairs [Zwaan and Radvansky, 1998]. They focus on the mental representation of events described in text and they include event-indexing, structure building, and construction-integration models of reading comprehension.

Events are thought to be related to one another on five main situation elements: time, space, entity, causation, and motivation [Zwaan, 2015]. The event-indexing model attempts to specify relations among these elements, in an effort to generally represent the events, which are described in text. The structure-building model focuses on identifying and describing the processes that operate during the comprehension of various media such as texts and pictures [Mohammad and Hamidreza, 2015]. As implied by its name the construction-integration model is a two-phased process of construction and integration. During the construction stage:

A crude mental representation of text, in the form of an associative network, is constructed from both linguistic input and the comprehender's own knowledge based in a simple bottom-up, data driven manner [Wharton and Kintsch, 1991].

At the integration stage, the associated network of crude mental representations of text are fine-tuned in the reader's mind and this helps the reader in interpreting the meaning of text. In support of the construction phase of the construction-integration situation model of reading comprehension, Miller and Johnson-Laird [1976] argued that humans create mental representations of what they read. In reading comprehension, the mental model theory states that a reader creates a mental representation of the text being read which simulates the circumstances in the "world" being described as the reader understands it [Casper et al., 1998]. The theory assumes that a reader focuses on a main character and creates a mental model of the circumstances surrounding the character in the text. The mental model is updated to integrate new circumstances of the character as the situations in the text unfold [Gunning, 1996]. In constructing a mental model, a reader is believed to link smaller units of text like words, phrases and sentences to build mental pictures [Woolley, 2011]. These pictures relate the reader's background knowledge to the text and makes higher-level inferences about it. The pictures are representations of people, objects, locations, events and actions and not of individual words, phrases, clauses, sentences and paragraphs of the text [Pardo, 2004; Palmer, 2016].

A fundamental hypothesis of this model is that sometimes written text is presented in a form analogous to that of perceived or imagined events [Johnson-Laird, 1981]. Skilled writers have the ability to evoke such presentations so that their readers seem to "experience the events" rather than merely read them. This suggests that writers must provide sufficient clues for readers to construct appropriate mental representations of text. Bower and Morrow [1990] report that readers tend to remember the mental model they construct from text rather than the text itself.

Part of the cognitive process involved in the mental model theory was explained by Wang and Gafurov [2003] as a procedure by which the brain searches relations between a given object or attribute and other objects, attributes, and relations in the long-term memory, and establishes a representation model for the object or attribute by connecting it to appropriate clusters of memory. The results of this process are fluid mental images that change as the reader absorbs new text [Woolley, 2010]. The ability to construct an active mental image of text is a cognitive skill known as visualization. Meaning that readers process both visual representations of verbal information and of objects to create meaning [Moore, 2016]. These explanations assume that both prior knowledge and the construction of a mental model are essential for successful reading comprehension. They support the idea that writers should provide sufficient clues in text to facilitate the

brain's process of identifying objects and attributes and for relating these to a readers' prior knowledge in the long-term memory throughout the text discourse.

However, the mental images generated are abstract. It takes at least three seconds for a reader to create a mental image. The threshold of three seconds is too short to create a detailed image. Normal word reading times are at least 10 times faster, it seems unlikely that readers typically generate detailed visual images during comprehension.

In general, the reading comprehension models show that textual features like content, style, linguistic, and cognitive features play an important role in determining the comprehensible level of text difficulty [Woolley, 2011].

Problems Addressed by the Reading Comprehension Models

There are various problems addressed by the reading comprehension models but the one which stands out from our discussion is insufficient prior knowledge which is mentioned in most of the models considered. Many researchers also agree that prior knowledge is one of the strongest predictors of reading comprehension [MancillaMartinez and Lesaux, 2010; Fisher and Frey, 2013; Elwér, 2014]. Three of the reading comprehension models discussed above support the argument that a reader needs to have specific, sufficient and suitable prior knowledge in order to comprehend what they read. However, there are a vast variety of dynamic knowledge domains that no single human can have adequate prior knowledge in all of them, no matter how widely they read.

The situation models in particular support the idea that authors need to provide sufficient clues to text so that readers can create accurate mental representations of the written text. These representations are useful for knowledge retention since readers tend to remember them and not the text. Clues in the text can be provided in form of illustrations and through elaborate writing by authors. Unfortunately, not all sorts of writing allow an elaborate writing style and multiple illustrations. For example, when authors write academic papers for journals or conferences their article may be restricted to fit on a certain number of pages. This will limit the number of illustrations that can be included in such articles, because illustrations can easily increase page numbers, since they require larger space compared to text.

Research has shown that successful reading comprehension and knowledge retention depends on the ability of a reader to reliably access and integrate prior knowledge, and the ability to generate, maintain and update iterative forms of meaning constructions [Griffiths et al., 2011; Kharismawati, 2015]. The following two sections discuss the problems of insufficient prior knowledge and insufficient clues to text.

Insufficient Prior Knowledge

Prior knowledge can help or hinder reading comprehension. If it is activated, sufficient, appropriate and accurate it helps, otherwise it hinders reading comprehension [Svinicki, 1994]. It affects how readers perceive, organize and connect new information. Vocabulary proficiency is one of the indicators of prior knowledge [Fisher, 2009]. Perfetti and Stafura [2014] introduced a framework with vocabulary knowledge at the centre of reading comprehension. They claim that there is a link between identifying the meaning of a word and reading comprehension. This claim is supported by Cromley and Azevedo [2007] who report that vocabulary and background knowledge make the largest contributions to comprehension. They suggest that vocabulary and background knowledge interventions might be the best way to begin improving academic reading comprehension.

Research about reading comprehension is mainly applied to children [Cain et al., 2000; Graham and Bellert, 2005; Tan, 2015; Herlina, 2016], but the concern about vocabulary causing failure in it can be applied to adults too [Murnane et al., 2012]. Actually, as readers grow older the demand placed on vocabulary and prior knowledge accelerates because they are expected to activate and apply these assets to subjects that are conceptually more difficult [Fisher, 2009]. These subjects require domainspecific literacy by the readers because they present special vocabularies specific to their knowledge areas. Adult readers cannot have sufficient prior knowledge in all domains, since they tend to concentrate on text that is of interest to their profession or research. For example, it may be difficult for a person from an information technology background to understand writings from law authors because legal language is peppered with French vocabulary, which makes it difficult for readers who are unfamiliar with law to understand [Duckworth and Spyrou, 1995]. It can be said that the problem of reading comprehension in consideration of vocabulary knowledge affects both children and adults.

Education plays an important role in building vocabulary knowledge for individual learners and their reading comprehension skills in understanding semantically related texts. For this reason, it is normally assumed that older people with high education have sufficient vocabulary and reading skills to read fluently. However, Cain et al. [2000] report that a difference between less skilled comprehenders and a comprehension-age match does not prove causality. They indicate that strengths in the comprehension skills measured are more plausibly the cause of reading comprehension success. This implies that reading comprehension does not depend on how young or old a reader is, but rather on the strength of their cognitive processes and skills in a knowledge domain of interest.

Insufficient Clues to Text

Since readers tend to remember the mental representation of text rather than the text itself, the ability of the reader to construct a mental model of written text is vital for knowledge retention [Shihab, 2011]. As mentioned earlier, the process of creating a mental model of text may be facilitated by visual clues, which can be provided through illustrations.

Example 1.2.2: Hint for Knowledge Retention

I hear I forget, I see I remember, I do I understand.

(Chinese Proverb)

The proverb in example 1.2.2 implies that a reader can easily remember text if it is illustrated with visible images i.e. “I see I remember”. As seen earlier, in cognitive science comprehension is characterized as the construction of a mental model that represents the objects and semantic relations described in a text [Thuring et al., 1995]. This implies that if a reader spends little effort in the process of constructing mental images, then text may be easily comprehensible. Technology can also be used to integrate text illustration to facilitate the process of constructing mental models by a reader, which improves learning by facilitating reading comprehension and knowledge retention. A related idea for improving learning is the *multimedia principle* by Mayer [2005]. This principle states that “students learn more deeply from multimedia presentations involving words and pictures than from words alone.” Mayer [2005] says that this multimedia principle is consistent with the findings of Rieber [1990] that students learn better from computer-based science lessons when animated graphics are included. Despite the proven benefit of pictures or graphical illustrations in learning, many electronic tools for accessing knowledge like web search engines and question and answering systems still present their output as text snippets generally ignoring any pictures that occur in the knowledge sources [Bosma et al., 2008].

Visual coding, visual argument and conjoint retention are educational perspective theories which suggest that the use of realistic graphics in teaching materials increases the probability of improving knowledge retention [Vekiri, 2002; Nsamenang and Tchombé, 2012]. Wills [2004] acknowledges that graphics are used to overcome deficits in patient education material because the comprehension of such material is often impeded by text written at levels too high for the patient population. It can be concluded that graphics should be used alongside text clues like definitions. If high level text is well explained, comprehension is facilitated and knowledge retention can be enhanced by graphical representations of difficult text.

1.3 Motivation and Problem Statement

This thesis is motivated by the specific reading comprehension needs of university students. For example, they need to prepare for and excel in numerous time-bound tasks like course assignments, assessments, tests and examinations. University students also need to retain and apply the knowledge they acquire to help them in achieving these time bound tasks and in reviewing and mastering new course materials [Lindsey et al., 2014]. These students also have the need to experience the joy in reading and learning. To meet these needs, they wish to use minimum effort and time for reading comprehension in order to avoid frustration or anxiety caused by failure in it. Background knowledge, vocabulary knowledge and sufficient clues to text can help in reducing the effort and time needed by a student to comprehend what they read. On many occasions this is not the case, because students are challenged by insufficient background knowledge, bewildering vocabulary [Jabbar, 2015] and poor clues to text, which make their reading comprehension difficult.

When bewildering vocabulary is used at the rapid pace of an expert reader, the words translate into jargon and their overuse creates gaps in the students' ability to process new information [Svinicki, 1994]. Example 1.2.1 on page 6 portrays a good scenario of overusing bewildering vocabulary. From a cognitive science perspective, it is thought that humans have limited processing capacity [Ur, 2004] for attention and absorption of information. A person with good processing capacity is generally a proficient reader, who depends on the ability to recognize words quickly and effortlessly. If word recognition is difficult, people use too much of their processing capacity to read individual words, this slows down their ability to comprehend what is read. It is true that reading comprehension goes beyond vocabulary knowledge, but research has shown that this is an important indicator for text comprehension [Nagy, 1988; Fisher, 2009; Perfetti and Stafura, 2014]. Sedita [2005] says that vocabulary knowledge is important because it encompasses all the words we must know to access our background knowledge, express our ideas and communicate effectively, and learn about new concepts.

Insufficient vocabulary knowledge is compounded by the reality that languages change overtime [Murray, 1996]. When this happens, new vocabularies and concepts are formed. For students who are not familiar with certain knowledge areas, these new vocabularies present unfamiliar cases that slow down reading comprehension, since meaning cannot be seamlessly recognized from such vocabulary. Moreover, most vocabularies are ambiguous in relation to the context in which they are used. For instance, the word *tweet* has various meanings depending on the context, as shown in example 1.3.1.

Example 1.3.1: The word tweet could mean³:

- Tweet (bird call), a type of bird vocalization
- Tweet, a message sent using Twitter
- Cessna T-37 Tweet, a twin-engine United States trainer-attack type aircraft
- Tweet (singer) (born 1971), American R&B and soul singer-songwriter

While insufficient vocabulary knowledge slows down word recognition and thus the process of reading, insufficient background knowledge in specific domains slows down comprehension of textual meaning by offering no foundation for making inferences. There is such a variety of knowledge domains that no single student can be conversant with all of them. In reality most students read text related to their subject area and it is less likely that they will acquire background knowledge in other domains or build vocabulary in knowledge areas that are not related to their courses of interest. Knowledge acquired by the students may be easily lost if it does not make a clear and simple impression in their minds. This problem is a result of insufficient clues to text in the students' reading material that affect the capacity of their brains in generating mental pictures. A struggle in generating mental pictures of read content affects knowledge retention.

We have identified four major problems that hinder students' reading comprehension, namely:

- (i) Insufficient background knowledge
- (ii) Insufficient vocabulary knowledge
- (iii) Ambiguous vocabulary
- (iv) Missing graphical support

Most university textbooks and other course documents are written in technical language, therefore they contain domain-specific vocabulary that may be unfamiliar to students. The authors of such documents sometimes presume that their readers have some background knowledge in the addressed subject area. This presumption causes the authors to provide minimal clues in text as they write none elaborate content. Consequently, the recent years have seen an enthusiastic growth of research in text wikification [Cai et al., 2013; Ferragina and Scaiella, 2010; Milne and Witten, 2008; Mihalcea and Csomai, 2007] as a way to enhance the comprehension of electronic resources that contain difficult vocabulary and poor clues to text.

³Retrieved from <https://en.wikipedia.org/wiki/Tweet> on 4-6-2018

1.4 Text Wikification

Text wikification is the process of automatically linking keywords and key phrases found in plain electronic resources to Wikipedia articles that have the correct meaning of the linked terms [Cai et al., 2013]. This process is generally known as automatic hypertext generation, which refers to tagging text with anchors that lead to external or internal information resources of a document [Domingue et al., 2001]. The difference is that general hypertexts may lead to any web document, whereas wikification hypertexts strictly lead to Wikipedia articles. Keywords and key phrases are a good presentation of domain-specific vocabulary also known as technical terms. For ease of writing and references, we shall refer to domain-specific vocabulary as technical terms because they are widely known as such in the interdisciplinary field of human language technology, a field that we explore for enhancing reading comprehension for students at universities by dealing with technical terms in context.

Two major tasks involved in text wikification are automatic keyword extraction and word sense disambiguation (WSD). Keyword extraction is the task of automatically recognizing terms that indicate the subject of a document. For example, if the keywords in a document are *pregnancy*, *C-section* and *traditional birth attendant* then its subject is probably *childbirth*. In natural language processing WSD is the task of determining which meaning of a word is active in a particular document context. WSD is necessary because the meaning of a word is ambiguous in different document contexts as inherent from natural language. The WSD task in text wikification can be illustrated by figure 1.1.

Figure 1.1: Illustration of the automatic text wikification process

Given the sentence:

Please *run* the *word processor* in a new *window*.

a wikification application should automatically extract the three keywords *run*, *word processor* and *window*, and link them to Wikipedia pages which have the correct definition of these keywords in relation to the context where they occur. For example, in this case the keyword *run* will be linked to a Wikipedia page with the title *Execution (computing)* because that is what it means in this context. The dotted arrows in figure 1.1 show other Wikipedia pages to which the keyword *run* could be linked. Word sense disambiguation is responsible for linking the extracted keywords to correct Wikipedia pages by using various computer algorithms. There are of course more alternative Wikipedia pages to link the keyword *run* depending on the context where it occurs. These four example pages are selected to show the difficulty in the task of word sense disambiguation during the process of automatic text wikification. The beauty with automatic text wikification is that it can extract keywords made of more than one word, for example *wordprocessor*. This is especially advantageous when dealing with technical terms because the majority of them are compound words or word phrases [Hiroshi and Tatsunori, 2000].

Wikipedia is a free-content on-line encyclopedia, which is a product of the continuous collaborative effort of many volunteer contributors. Although a few critics have questioned the credibility and coverage of Wikipedia, in the year 2005 a special report on science articles indicated that Wikipedia is similar to Encyclopedia Britannica in both coverage and accuracy [Giles, 2005]. A similar more recent study reports that the quality of Wikipedia articles is on par and that Wikipedia contains over 2.5 billion words, over 60 times more than encyclopedia Britannica [Schaefer, 2014]. Questions about the authenticity of Wikipedia arise because of the collaborative nature through which the encyclopedia grows, because many of the contributors are not accredited authors. The co-founders of Wikipedia apparently see this as an advantage because they anticipate that any error noticed in the content pages shall be boldly corrected by the people who notice them. An advantage that has been wisely utilized by colleagues from the fields of science, technology, engineering and mathematics who are transforming Wikipedia to benefit higher education globally, and to the benefit of the global public [Haslam, 2017]. Haslam's research showed that Wikipedia can be a vital part of the undergraduate and graduate curricula especially in natural sciences.

Wikification of Education Resources

A number of text wikification researches have been done for electronic education resources. This is a logical development that arises from the need to enhance students'

comprehension of such resources that may be littered with unfamiliar vocabulary. For example, Csomai and Mihalcea [2007] tailored a mini-version of the Wikify! [Mihalcea and Csomai, 2007] system for linking educational materials to Wikipedia Knowledge. Wikify! introduced the use of Wikipedia as a resource for automatic keyword extraction and word sense disambiguation. If a document is given as input to Wikify!, the system has the ability to identify the key phrases in the text and link them to corresponding Wikipedia pages. Eventually, these links provide users with quick access to additional information. Csomai and Mihalcea [2007] demonstrated that providing such information improves both the quality of the knowledge acquired and the time needed to obtain such knowledge. Results from their experiment indicated that:

“providing students with information relevant to the topic of study, and bringing such information within easy reach through hyperlinking, are both successful strategies for increased effectiveness in pedagogical tasks.” [Csomai and Mihalcea, 2007]

Another example of education text wikification can be given as that done by Lui et al. [2011] who present a study that considers the conversion of plain text documents into hypertext for web-based learning. They generate two kinds of hyper-links, one is the self-referencing links for elaboration and references and the other is links to Wikipedia articles for exploratory learning. Lui et al. noticed that only a few self-referencing links could be generated using a test set of course text. This indicates that the elaboration and referencing links need to be supported by the exploratory learning experience through linking course text to Wikipedia in order to effectively enhance the comprehension of web-based course text and learning.

Automatic text wikification offers a number of advantages such as identifying keyword phrases, providing nonlinear navigation for seeking relevant information, aiding the generation of new knowledge and creating opportunities for exploring new knowledge. Besides their advantages, text wikification has certain drawbacks. For example, it creates hyper-links that direct the reader to a Wikipedia article. This could easily take away attention of the reader from read text, cost time and translate into “noise” [Zhang et al., 2014]. Wikification neither provides precise definition of the domain-specific vocabulary nor does it provide illustrations for such concepts where appropriate. It is good for providing additional clues to text that may enhance reading comprehension of electronic resources but this advantage is seriously compromised by the limitations it presents. In addition to text wikification, we would like to investigate the possibility of integrating Wikipedia knowledge into education material as a way of providing vocabulary and background knowledge “on the fly”. In this way, students can have quick access to the meaning of vocabulary they do not understand and background knowledge on unfamiliar concepts that they may be considering.

Allowing students to easily build their vocabulary levels and background knowledge as they read without linking away from the read text. We refer to the approach of integrating relevant information into text as document enrichment.

1.5 Document Enrichment

Document enrichment is a relatively new research area in information science, which emphasizes adding information and knowledge to existing text [Zhang et al., 2014; Motta et al., 2000; Mulholland et al., 2000] by creating hypertexts that link to relevant web resources. Document enrichment is an essential task for reading comprehension because as we have seen, authors may omit basic information to make a document concise [Zhang et al., 2014] and they may use technical terms. Concise writing and the use of technical terms generally gluts text with knowledge gaps as authors omit information and assume a certain amount of background and vocabulary knowledge for readers. Comprehending documents with knowledge gaps in their content becomes difficult [Sousa, 2011] if a reader does not have sufficient background and vocabulary knowledge in the topic being considered, especially for domain-specific text. Document enrichment seeks to fill such knowledge gaps.

Some researchers have used the concept of document enrichment to refer to the process of text wikification [Zhang et al., 2014; Bron et al., 2011]. That is, creating hypertexts in plain text documents that link to Wikipedia pages with relevant information. Keyword and key phrase extraction, word sense disambiguation and hypertext generation are the fundamental techniques of this type of document enrichment. We conceive a different document enrichment concept as follows: in addition to the fundamental techniques of contemporary document enrichment which stop at generating hypertexts which link to Wikipedia knowledge, our priority is to seamlessly integrate knowledge into text. The idea is to maintain a student in the document being read as long as possible. Without linking away to search for information to fill the gaps created by their insufficient background or vocabulary knowledge. To do this, we explored the possibility of modifying the text wikification concept to fill knowledge gaps in-text as a way of enhancing reading comprehension for students at universities. This would save effort, time and encourage concentration on reading while reducing interruptions that may be caused by linking directly to a Wikipedia article. Otherwise, a student still has to find the relevant information if they link directly to Wikipedia, let alone be drawn away and lost into the vast sea of information on the Internet.

So far, we have identified four problems that affect the reading comprehension process and these include insufficient background knowledge, insufficient vocabulary knowledge, ambiguous vocabulary and insufficient graphical support in educational

resources. Example 1.5.1 provides educational text that was difficult for university students to understand because of all the aforementioned four problems that affect reading comprehension. The example depicts text from a reading assignment that was given to first year students of Bachelor of Computer Science and Bachelor of Information Technology at Gulu University. The reading assignment was for an Embedded Systems course unit. This assignment presented problems in reading comprehension especially for the Information Technology (IT) students who did not have a science background and hence insufficient knowledge of concepts like *transistors*, *semi-conductor* and *microchip*. This gave no opportunity for the students to make inference from their prior knowledge schemata and bring meaning to what a *microprocessor* really means.

Example 1.5.1: Educational text that may be difficult for some readers

A *microprocessor* is an electronic computer component created from miniature sized *transistors* and some other *circuitry elements* on a solitary *semi-conductor* IC (*integrated circuit*) or *microchip*... There are mainly five sorts of *microprocessors*; these *minute* units provide computers with a “brain”.

The scientific concepts also presented unfamiliar technical terms, which did not allow the IT students with no science background to understand the meaning of *microprocessor* because they did not know what these terms mean. A more interesting phenomenon was the misinterpretation of the ambiguous word *minute*. Most of the students had knowledge of this word as *the sixtieth part (1/60) of an hour, as in sixty seconds* and not something *extremely small, as in size*. This caused confusion and the students were not able to relate *five sorts of microprocessors* to *sixty seconds*. The definition in example 1.5.1 was given without graphical support for the technical terms like *transistors*, which made it difficult for the IT students to make mental impressions of such terms.

Since we are interested in modifying the text wikification concept for facilitating reading comprehension by students at universities, we investigated a data-driven or encyclopedia-driven document enrichment approach. Based on the successful application of text wikification to educational documents, we realized that it was worth investigating a document enrichment approach that extends the text wikification features to facilitate reading comprehension of students at universities. In the approach to be investigated, we intend to resolve ambiguity in the meaning of technical terms in context using word sense disambiguation algorithms. This will help in providing contextually relevant information for terms such as *minute*. The problem of insufficient background knowledge and vocabulary gaps can be filled using articles from the Wikipedia encyclopedia, and missing graphical support can hopefully be found

there too. In figure 1.2, we see that Wikipedia provides background knowledge for and defines the technical term *transistor* and provides a graphic illustration for it. Such information from Wikipedia can be used for facilitating university students' reading comprehension through document enrichment. A detailed discussion on this document enrichment approach can be found in chapter 5 of this thesis.

Figure 1.2: Screen shot from <https://en.wikipedia.org/wiki/Transistor>, showing the richness of information from Wikipedia.

1.6 Research Questions

Research in facilitating reading comprehension of education documents content in the area of automatic text wikification is a process that concentrates on generating hypertexts. The hypertexts are generated with keywords found in the documents as anchor text which are linked directly to Wikipedia pages. This leads the students away from the document being read. The text wikification process is actually a document enrichment approach, if we consider the contemporary meaning of the document enrichment concept. Therefore, the main research question guiding this research is:

How can reading comprehension be facilitated among university students without leading away from the content of course documents?

The following are specific research questions that helped to investigate the problem of reading comprehension among university students and assist to devise a design artifact that facilitates the grounded problem:

- (i) What are the key processes involved in reading comprehension?
- (ii) What are the key factors that hinder the reading comprehension processes?
- (iii) What are the key tasks in a document enrichment approach for facilitating reading comprehension?
- (iv) What is the performance of the document enrichment algorithms?
- (v) What is the perceived usefulness and usability of the document enrichment approach?

1.7 Research Approach

Research approaches are the plans and procedures for inquiry that span the steps from broad assumptions to detailed methods of data collection, analysis, and interpretation [Creswell, 2013]. This statement implies that a research approach outlines the research philosophy, strategy, methodology and techniques that are used throughout the research process [Tweheyo, 2018; Mirembe, 2015; Sol, 1982]. Selecting an appropriate approach for a research may be done by considering the research object, which is indicated by the nature of the problem and the anticipated solution. Similarly, after examining information systems research, Orlikowski and Baroudi [1991] concluded that such research does not exhibit a single set of philosophical assumptions regarding the nature of the phenomena studied and what constitutes valid knowledge about those phenomena. Plans and procedures of inquiry, which adopt design science principles within an Engaged Scholarship approach and following a strategy of Singerian inquiry in a pragmatic framework of abduction, were selected as the philosophy and strategy of research.

Research Philosophy

Research philosophies are the perceptions, beliefs, assumptions, the nature of reality and truth (knowledge of that reality), that influence the way in which a research is undertaken, from design to conclusion [Flowers, 2009]. According to Guba and Lincoln [1994], each research philosophy has a unique position of explaining *ontology*,

epistemology and *axiology* (values). *Epistemology* deals with our view on how knowledge is acquired and *ontology* deals with the nature of reality. There are three epistemological choices for social science research which include positivism, interpretivism, and pragmatism [Gonzalez and Sol, 2012], with their corresponding ontological positions. These epistemology choices which are explained below can be borrowed and applied to this research [Ricciardi, 2009].

“Positivism is based on deductive or theory testing by measuring or observing social realities. Interpretivism also referred to as constructivism is based on the belief that individuals and groups make sense of the social world basing on their experiences, memories and expectations. Pragmatism is based on the notion that the most important determinant of research philosophy is the research question. It is not committed to any one system of philosophy or reality” [Tweheyo, 2018].

In developing philosophical positions logical assumptions are made concerning the nature of society and science, this leads to a discussion that justifies the philosophy stand of an inquiry [Mirembe, 2015]. The research philosophy selected for this study is Engaged Scholarship with design science principles based on an interpretive and pragmatic epistemology stance with a critical realist ontology position. The selection of Engaged Scholarship as our research philosophy was influenced by the need for action-oriented solutions to the problem of reading comprehension in universities. Engaged Scholarship is a form of inquiry where researchers involve others and leverage their different perspectives to learn about a problem domain [Van de Ven, 2009]. This research engages university students at various education levels as domain stakeholders, in identifying problems and solutions in reading comprehension. It is argued that the principles of interpretivism and critical realism are suitable for Engaged Scholarship research because of their focus on different stakeholders' perspectives in overcoming the difficulties associated with positivism [Tweheyo, 2018].

Design science principles are applied to the Engaged Scholarship philosophy because they are fundamental for solving questions relevant to human problem through the creation of innovative artifacts [Hevner et al., 2004]. These innovations seek to define the ideas, practices, technical capabilities, and products through which the analysis, design, implementation, management, and use of information systems can be effectively and efficiently accomplished [Hevner et al., 2004]. Creating successful innovative artifacts requires engaging people from diverse backgrounds and perspectives [Hevner, 2007] which in turn requires a method of triangulation. This is the fundamental reason why design science principles are applied to the Engaged Scholarship research philosophy. Hevner [2007] analyzed design science research as an embodied of three closely related cycles of activities

Figure 1.3: The three-cycle view of Design Science Research [Hevner, 2007]

as shown in figure 1.3. The *Relevance Cycle* inputs requirements from the contextual environment into the research and introduces the research artifacts into environmental field testing. The *Rigor Cycle* provides grounding theories and methods along with domain experience and expertise from the foundations’ knowledge base into the research and adds the new knowledge generated by the research to the growing knowledge base. The central *Design Cycle* supports a tighter loop of research activity for the construction and evaluation of design artifacts and processes.

Critical realism maintains a strong emphasis on the nature of reality, this is consequential to its dogma that the world exists independently of what we think about it [Zachariadis et al., 2010]. Critical realists argue for a shift from prediction to explanation, the use of abstraction, and reliance on interpretive forms of investigation [Wikgren, 2005]. This argument implies that critical realism involves stakeholders in the process of identifying problems and solutions, making it a good ontology position for the Engaged Scholarship research philosophy.

Research Strategy

A research strategy provides the overall direction of the research, including the process by which the research is conducted and it underlines Singerian inquiry as one of its probing systems [Tumwebaze, 2016]. Singerian inquiry believes that reality is not proven, but arrives when enough decision makers are convinced as to what is real [Boggs and Hammer, 2010]. Our research adapted a research strategy based on Singerian inquiry in a pragmatist framework of abductive reasoning [Gonzalez and

Sol, 2012; Sol, 1982]. Such a strategy is good for dealing with ill-defined problems in complex and volatile environments as is the case for problems of reading comprehension in universities, which concerns our research. Singerian pragmatism is an epistemology and ethical theory stating that the value of knowledge should be expressed in terms of how the knowledge improves the human condition [Wijnhoven, 2013]. Our research seeks to improve the experience of reading comprehension by students at universities. Therefore, the Singerian inquiry in a pragmatist framework is a suitable strategy for such a research.

Abductive reasoning is a kind of interpretive inference [Ross, 2010], which begins with an incomplete set of observations and proceeds to the likeliest explanation for the set [Patajoki, 2013]. This kind of reasoning will help us to infer the best explanation for reading comprehension problems in universities, based on the experience of students. Students were engaged in an open-ended on-line questionnaire allowing us to make the best conclusion about their reading comprehension problems from their reading experiences. Abductive reasoning is able to introduce new ideas, solve problems, and lead to new explanations of life reality. It thus has the capacity to generate knowledge, which is important for the rigor cycle in design science, since this cycle deals with grounding the problem and adding to the knowledge base of reading comprehension in our case.

Figure 1.4: Research strategy, adapted from Gonzalez and Sol [2012]; Sol [1982]

The adapted research strategy allowed five consequential phases, which supported the application of our design philosophy. These phases included *initiation*, *abstraction*, *theory formulation*, *instantiation*, *testing* and *evaluation* (see figure 1.4).

- (i) *Initiation*: The initiation phase focused on scoping, formulating and grounding the problem in reading comprehension for students at universities. Problem grounding seeks to identify issues that are relevant to an environment. This is done so that artifacts designed to solve such issues are intrinsic [Ziemke, 1999], meaningful, useful and usable in the selected environment. The research relevance and rigor as well as contributions to the reading comprehension knowledge base [Hevner et al., 2004; Van de Ven, 2007] were also identified in the initiation phase. During this phase, an extensive survey of literature was done on reading comprehension theories and it was found out that technical terms are one of the major hindrances to reading comprehension. Literature review was also done on contemporary techniques that are used in dealing with technical terms in an effort to facilitate reading comprehension. An empirical description of the research problem was resultant to the initiation phase.
- (ii) *Abstraction*: According to Kamarudin et al. [2016] abstraction is a process of facilitating an issue to help designers solve problems efficiently. They continue to say that abstraction reduces complexity and increases creativity by guiding designers to focus on the key factors of a problem towards producing a broader solution perspective. In our research, identifying the key factors that hinder reading comprehension in universities was achieved through a stakeholders' survey. This survey was carried out by using an on-line open-ended interview questionnaire. The on-line interview was also interested in suggestions for technology-based solutions for resolving issues in reading comprehension. The abstraction phase therefore carried out an exploration to get a generic understanding of the problem domain. The outcome of this stage provided a set of functional services that were needed for an approach to facilitate reading comprehension on electronic media.
- (iii) *Theory formulation*: After conceptualizing and ascertaining relevance of the reading comprehension problem for students at universities through initiation and abstraction, theory formulation was carried out to devise an appropriate solution for the problem [Tumwebaze, 2016]. Devising solutions for relevant problems is a major principle of design science and it seeks to extend boundaries of human and organization capabilities by creating new and innovative artifacts [Hevner et al., 2004]. Such artifacts include algorithms, which were developed in our case for a document enrichment approach to facilitate reading comprehension. The theory formulation phase was thus a problem-solving phase which provided appropriate ways of addressing the reading comprehension problems that were described and understood in the initiation and abstraction phases respectively. Through abductive reasoning on the problem, we were able to find a satisfying

explanation for hindrances to reading comprehension of students at universities and to implement suitable solutions as suggested by the stakeholders at the abstraction stage. The theory formulation stage resulted in a prescriptive conceptual model for the document enrichment approach to facilitate reading comprehension.

- (iv) *Instantiation and testing*: Instantiation is referred to as a material artifact while the other types of artifacts are referred to as abstract artifacts [Peffers et al., 2007; desrict.org, 2015]. These material artifacts allow for the implementation of abstract artifacts. In our case the instantiation was realized by a document enrichment approach to facilitate reading comprehension by designing document enrichment algorithms. Testing was carried out on the performance of the performance of the document enrichment algorithms. The algorithms created for the document enrichment approach were tested for their performance using precision and recall. This helped to see if the algorithms performed as specified in the abstraction phase. A prescriptive empirical model was the outcome of the instantiation and testing phase. The prescriptive empirical model made it possible to evaluate the document enrichment approach.
- (v) *Evaluation*: The document enrichment approach to facilitate reading comprehension was evaluated for its usability and usefulness by means of a user survey and a user study. The evaluation was done in two universities after enriching students' course contents with information that facilitates reading comprehension. Feedback from the evaluation phase was gathered using qualitative questionnaires.

1.8 Thesis Outline

This thesis contains 8 chapters. Chapter 1 reviews practical issues in the concept of reading comprehension and presents this concept using various models. The chapter also explains the motivation for our research and stipulates its problem and questions. By considering how contemporary researchers have dealt with problems encountered in the process of reading comprehension for students at universities, the chapter suggests a document enrichment approach as a possible way to facilitate reading comprehension. Our research approach including its philosophy and strategy are also discussed in this chapter.

Chapter 2 reviews literature concerning contemporary issues in dealing with the problem of technical terms in reading comprehension. The literature review was specific to dealing with technical terms because the literature survey in chapter 1 showed that such terms hinder reading comprehension. Moreover, computers have

challenges in dealing with them because computers do not process world knowledge. The chapter lays a foundation of state-of-the-art technologies that have been applied in the process of facilitating the reading comprehension process by dealing with technical terms. These technologies support a document enrichment approach to facilitating reading comprehension.

Chapter 3 reports a stakeholders' survey that was carried out to ground the problem caused by technical terms in reading comprehension of university students. The report highlights key factors hindering the reading comprehension process of students at universities and the general effect of the factors on the process. It also discusses the stakeholders' views towards facilitating the reading comprehension process, given the challenges they experience in it. The chapter concludes that a document enrichment approach named TermPedia would be appropriate for facilitating reading comprehension according to reviewed literature and the stakeholders' views.

Chapter 4 explains the functional design of a set of services reliant on results from the stakeholders' survey discussed in chapter 3. It describes the processes which lead to designing the document enrichment algorithms for TermPedia. The chapter also discusses the materials, and procedures that have been used in the process of designing the algorithms. The criteria used by each algorithm for document enrichment were also described in this chapter.

Chapter 5 presents TermPedia the document enrichment approach for facilitating reading comprehension using Sol [1988]'s framework. This framework described TermPedia's way of thinking, governance, working, modelling and supporting.

The process of testing the performance of TermPedia's document enrichment algorithms was described in chapter 6. The chapter explains the methods and data used in testing the performance of the algorithms and concludes by presenting and discussing results of the tests.

In chapter 7, detailed descriptions of a user survey and a user study were given. The user survey and study were carried out to evaluate TermPedia. The main objective of the evaluation was to investigate the usefulness of TermPedia. The chapter reports the results of the user survey and study. It concludes by comparing the opinions of participants from the user survey to those of participants from the user study on the idea of using document enrichment in facilitating reading comprehension.

In chapter 8 a conclusive remark is given about the entire research by reflecting on how the suggested research approach was applied to facilitate the problem caused by technical terms in reading comprehension. The chapter clearly identifies the contribution made to the knowledge of facilitating reading comprehension and how the research results may be applied. It concludes by giving recommendations and laying a ground for future work.

Chapter 2

Dealing with Technical Terms

The ambiguities of language, both in terms of vocabulary and syntax are fascinating.

Marilyn Hacker

This chapter presents a literature survey of contemporary information technology concepts, approaches and methods, which have been applied to automatically deal with the issue of technical terms in electronic documents. The objective of the literature survey was to guide our research in identifying the problems presented while using information technology in dealing with technical terms. The problems identified include automatic term recognition and disambiguating the meaning of technical terms in document context and automatic provision of definitions and background information for the recognized terms. The literature survey shows that information technology can be successfully applied in dealing with technical terms despite the problems presented in such attempts. The survey also helped to generate an original view of the important aspects in dealing with technical terms for facilitating reading comprehension.

2.1 Issues in Dealing with Technical Terms

Practical issues in reading comprehension and reading comprehension models showed that unfamiliar vocabularies are a hindrance to understanding document content (See chapter 1 of this thesis). When vocabulary gaps hinder reading comprehension, we keep in mind that the real issue that makes them bewildering or unfamiliar to the reader is lack of relevant prior knowledge. Vocabulary in text can be as common as *runs* but if the reader does not have the prior knowledge schema on which to hang its interpretation within the context, then there will be a challenge in comprehension. For example, people who are not familiar with baseball struggle to understand the concepts of *runs*, *innings*, *outs*, and *points* because they lack the specific prior knowledge schema on which to hang their interpretations [Murnane et al., 2012] within this knowledge domain. This example shows that sometimes the interpretation of vocabulary is

specific to knowledge domains and requires relevant prior knowledge for comprehension. Such vocabularies are referred to as domain-specific terms or technical terms. As mentioned in subsection 1.3 on page 14 of this thesis, we shall refer to them as technical terms in our writing.

By technical terms we refer to anything that is a word, a group of words, an acronym, or an abbreviation, which designates a special meaning in the context of a specific knowledge area. Although the meanings of terms in general are often clear from text context [Cain, 2007], it is also true that technical terms are often used without proper definitions [Busch, 2004; López, 2013]. This is because when authors address issues of interest in a specific knowledge domain, they assume that the readers are familiar with that domain's specific terminology and see no need to define the terms. On the contrary however, text written for a specific knowledge domain may catch the interest of a person, who is not familiar with that domain. In which case the undefined technical terms will surely hinder reading comprehension [Binotti et al., 2001; Moore, 2013]. Similarly, a recent pre-print article shows that the readability of science is steadily decreasing, which trend is indicative of a growing usage of general scientific jargon [Pontus et al., 2017].

In order to automatically address the problem of technical terms in text, they need to be recognized and disambiguated with reference to the text context in which they occur. After that they can be defined and provided with relevant background information and clues to text meaning. Providing relevant background information involves linking the identified term to contextually relevant information. The reason for disambiguating technical terms is that they can be ambiguous with reference to specific knowledge domains, see example 2.3.1. Background information is provided because sometimes the definition of a technical term includes other unfamiliar vocabularies, see example 2.4.2.

In the following sections we provide a literature survey of contemporary information technology concepts, approaches and methods which have been applied to automatically deal with the issue of technical terms in text. The objective of this literature survey was to guide our research in identifying the problems presented by using information technology to deal with technical terms in text and generate an original view to the important aspects of automatically dealing with technical terms.

2.2 Recognizing Technical Terms

Automatic term recognition also referred to as automatic term extraction, has been a sub-field in information retrieval and natural language processing for a long time [Ananiadou, 1994]. It is the process by which a computer automatically extracts lexical units that designates concepts in thematically restrictive domains.

This means that the extracted lexical units refer to common, rare, or new concepts with special meanings in the context of specific knowledge domains, in which case they qualify as technical terms. For example, Yu et al. [2012] published an article with the new term *nanoresonators* which refers to tiny mechanical devices that could help cell phone users avoid the nuisance of dropped calls and slow downloads. Besides new terms, there are terms that have been in use for many years and refer to specific concepts just as illustrated by the term *runs* in section 1.2 of this chapter. It is easy for humans to recognize such words as technical terms but the same cannot be said for computers. The vital problem in automatic term recognition involves their detection. Therefore, term extractors provide term candidates and not terms. For a lexicon unit to qualify as a term candidate it should have the following properties; *unithood*, *termhood*, and *specialised usage* [Vivaldi and Rodríguez, 2010]. The first property refers to the internal coherence of a unit; the second to the degree a given candidate is related to a domain-specific concept and the latter to the specialized usage of the unit. It is obvious that measuring such properties is not an easy task, thereby providing a big challenge for the process of automatic term recognition. This is why there has been extensive research in the area of automatic term recognition since the late 1980s [Cabr et al., 2001], with significant improvements over the years [Milios et al., 2003; Korkontzelos et al., 2008; Barrn-Cedeo et al., 2009].

An approach to automatic term recognition according to Knoth et al. [2009] consists of a combination of linguistic and statistical phases. The linguistic phase employs a linguistic filter, based on part-of-speech (POS) tags and syntax, to extract a set of candidate terms, while the statistical phase uses statistical methods to assign weights to each term candidate. A detailed study of the statistical methods used for automatic term recognition can be found in the comparative study by Knoth et al. [2009], some of which include term frequency, a combination of term frequency and inverse document frequency (TFIDF), weirdness, and likelihood ratio. Knoth et al. [2009] reported that the best approach to automatic term recognition was the combination of weirdness and TFIDF. Weirdness coefficient is a measure for term candidates that contrast the frequency distribution of a word in a special (domain-specific) corpus to the frequency distribution of the same word in a general corpus. Words with high weirdness coefficients and relative frequencies are considered as good term candidates [Sebastiani, 2003]. The weirdness coefficient is expressed by the following formula:

$$\text{Weirdness coefficient} = \frac{f_s/N_s}{f_g/N_g} \quad (2.1)$$

Where f_s and f_g are the frequencies of a word in the special and general corpora respectively and N_s and N_g are the total number of words in the special and general

corpora respectively.

Another approach to automatic term recognition uses Wikipedia for concept identification. For example, Vivaldi and Rodríguez [2010] present an approach for obtaining the terminology of a given domain using the category and page structure of Wikipedia in a domain and language independent way. Their idea was to take advantage of the category graph of Wikipedia starting with a set of categories that they associate with a selected domain. After obtaining the full set of categories belonging to the selected domain, the collection of corresponding pages is extracted using some constraints. The set of titles of recovered Wikipedia pages and categories is selected as initial domain terms. Their system has been tuned, applied, and evaluated on the production and validation of a set of term candidates. The applied production and validation of term candidates has been applied to several domains and languages showing acceptable performance with the state-of-the-art in term extraction.

In another approach Akalya and Sherine [2010] develop a general framework for term recognition and extraction using Wikipedia data but they concentrate on extracting terms for ontology construction, which can also be used for enhancing reading comprehension. Akalya and Sherine [2010] use Wikipedia links to describe context and provide background knowledge for investigated topics and they use this information to create domain taxonomies. This means that Wikipedia links can also be used to create a dictionary of technical terms since the investigated topics are a good example of such term.

Similarly, Grineva et al. [2009] also discuss an approach to automatic term recognition that uses Wikipedia page titles as technical terms. In their approach, Grineva et al. [2009] model a document as a graph of semantic relations between terms and that document. Each text document is parsed and all possible n-grams are extracted and the variations for each n-gram are constructed using different morphological forms of its words. All n-gram variations are then sought for among Wikipedia articles. Candidate terms are weighed and the semantic relations among them are determined by exploiting information extracted from Wikipedia. A semantic graph of terms extracted from a document is analyzed using an algorithm for detecting the community structure of a network. Using terms in detecting semantically related communities in networks can be reduced to detecting semantic relatedness in a single document depending purely on context information. The approach of communities in networks provide several Wikipedia articles for each recognized term whereas only one article should be provided for each recognized term after term sense disambiguation if the interest is in clear contextual explanations of the recognized term.

Jones [2007] made an evaluation of the various techniques used in term extraction for searching ontologies and saw that the techniques, which use Wikipedia were significantly more effective than the other techniques. This evaluation result was a

motivation for us to use Wikipedia article titles and links embedded within the articles as a baseline for our technical term extraction technique. In the same year of this evaluation report, Fahmi et al. [2007] experimented with the method of automatic term recognition that combines linguistic and statistical phases.

Fahmi et al. [2007] focused on the use of existing terms from glossaries, thesauri, or ontologies to extract new terms from a domain-specific text. Their baseline system combines a linguistic pattern matcher for extracting candidate noun phrases with a statistical method for ranking candidate phrases according to their association strength in a domain-specific corpus. They developed a method for ranking candidate terms, extracted from Dutch medical corpora, with the help of the Unified Medical Language Systems (UMLS) as an external knowledge source. Fahmi et al. [2007] considered the extraction of phrasal terms for a question answering (QA) task. Although limited by the task at hand and knowledge sources, Fahmi et al. [2007] did a commendable job in identifying technical terms in the medical domain as shown by their results. Evidently not every knowledge domain has available dictionaries, thesauri, and ontologies, which can be used in the process of automatic term recognition. In such cases it would be helpful to use Wikipedia as the external knowledge source because it has a good coverage of nearly all existing knowledge domains.

2.3 Technical Terms' Sense Disambiguation

Technical terms can be ambiguous because they carry different meanings in specific knowledge domains. A case of an ambiguous term is given in example 2.3.1 where the word *ontology* refers to different concepts when considered within *computer science* or *philosophy* knowledge domains. For humans, it is normally possible to distinguish the interpretation of an ambiguous term intuitively, when it occurs in a certain context, if they have sufficient prior knowledge of it. Otherwise, the ambiguous term will hinder text comprehension due to the lack of prior knowledge or misinterpretation of its meaning in text context.

Example 2.3.1: Ambiguous term

WordNet¹ defines the word *ontology* as “a rigorous and exhaustive organization of some knowledge domain that is usually hierarchical and contains all the relevant entities and their relations”, with reference to the field of *computer science* or “the metaphysical study of the nature of being and existence”, with reference to the field of *philosophy*.

¹WordNet is a semantic lexicon for the English language

As humans we can easily resolve word senses because our brains process “world knowledge”. Since computers do not possess the “world knowledge” used by humans for word sense disambiguation, they need to use other resources for fulfilling this task [Kasture and Agrawal, 2012]. A rich variety of techniques have been applied in the process of word sense disambiguation, including knowledge-based, supervised machine learning, semi-supervised, and unsupervised methods [Edmonds and Agirre, 2008].

Knowledge-based approaches to word sense disambiguation depend on different knowledge sources such as machine-readable dictionaries or sense inventories and thesauri [Pal and Saha, 2015]. In the knowledge-based approach, a word sense is guessed by counting overlaps between dictionary definitions of various senses of the word and the context where the word appears in text [See Lesk, 1986]. Lesk [1986] is well known for introducing the Lesk algorithm, which is a classical method for word sense disambiguation. He reports that his system disambiguates word sense with an accuracy range between 50% to 70%, on text from *Pride and Prejudice* and selected papers of the *Associated Press*. Only single words are disambiguated in Lesk’s paper while the majority of technical terms are compound nouns or word phrases. Hiroshi and Tatsunori [2000], mention that “85% of domain-specific terms are said to be compound nouns”. Therefore, the techniques of word sense disambiguation can be applied to disambiguating the sense of technical terms if the technique can be modified to account for the fact that most terms are compound nouns or a combination of word phrases. Knowledge-based approaches to word sense disambiguation often give high precision scores, but their performance is affected by sparsity in word overlaps and their reliance on dictionary definitions [Pal and Saha, 2015].

In the supervised machine learning methods of word sense disambiguation, a classifier is trained for each distinct word on a corpus of manually sense-annotated examples. For example, Kasture and Agrawal [2012] present a supervised method for the word sense disambiguation task using a tree-matching approach. Their method required a context knowledge base containing a corpus of sentences. The corpus was created by collecting sentences from many electronic text sources, which included thousands of books mostly written by professionals. Kasture and Agrawal [2012] used an algorithm which matches the tree of input sentence with the trees of all the glosses of the target word available in the database of their system one by one. The evaluation results of this algorithm reflected that the system works well in disambiguating word senses if a comparatively smaller knowledge base is considered. Despite its high performance, like all other supervised methods of word sense disambiguation, this method has an important drawback, in that its applicability is limited to those few words for which sense-tagged data is available, and its accuracy depends on the amount of labelled data available at hand [Turdakov and Velikhov, 2008; Mihalcea,

2007]. Such systems also suffer from a data scarcity problem due to the intense manual labour required in tagging words with senses. Moreover, it is equally hard to acquire sufficient contextual information about senses of a large number of words in natural languages. A specific disadvantage to the supervised method used by Kasture and Agrawal [2012] is that the trees of different glosses of the ambiguous word should be available within the system in order to perform tree-matching. Semi-supervised and unsupervised methods tend to provide solutions for the drawbacks in supervised methods by not depending exclusively on annotated words, but these incur other special difficulties.

Semi-supervised methods make use of small annotated corpus as seed data in the bootstrapping process as proposed by Hearst [1991] where she describes an algorithm that checks the context surrounding a target noun against that of previously observed instances and chooses the sense for which the most evidence is found. The evidence consisted of a set of orthographic, syntactic and lexical features. However, the method does not require the most complex mechanisms in natural language processing techniques such as knowledge bases, semantic feature representations, or semantic inference because the methods are coarse. The algorithm described by Hearst [1991] was called Catchword, and its objective was as follows: given an English sentence or sentence fragment containing a target noun, determine which of a set of pre-determined senses should be assigned to that noun. Hearst [1991] showed that noun homograph disambiguation can be accomplished by making use of simple syntactical and lexicon-semantic information culled from a large corpus and her results were comparable to earlier efforts using machine readable dictionaries and large corpora. More importantly, this algorithm provided a means for the integration of information from machine readable dictionaries with that of large corpora towards the goal of more robust natural language processing. This algorithm is not directly relevant for our research, but it is presented here to provide an example of a method of word sense disambiguation that uses a semi-supervised method.

Unsupervised methods of word sense disambiguation cluster occurrences of words, thereby inducing word senses. The underlying assumption is that similar senses occur in similar contexts, and thus senses can be induced from text by clustering word occurrences using some measure of similarity of context [Edmonds and Agirre, 2008]. Then, new occurrences of the word can be classified into the closest induced cluster or senses. For example, a “Google and WordNet based word sense disambiguation” was presented by Klapaftis and Manandhar [2005] as an unsupervised method for automatic disambiguation of noun senses found in domain-specific unrestricted corpora. The method used Google to find contextually relevant terms, which in turn helped in assigning the correct WordNet sense to each term under disambiguation. Klapaftis and Manandhar [2005] start with a list of terms extracted from a collection of domain-specific documents.

Each sentence containing the domain-specific term is then sent to Google and the top four ranked documents are used in combination with WordNet synset to disambiguate each term. The system evaluation revealed a disambiguation accuracy of 58.90%. It is important to note that this system disambiguates both single nouns and compound nouns by limiting the disambiguation process to the list of words found in WordNet, which is quite restrictive. In many cases, the term to be disambiguated may not be listed in WordNet. This problem could be overcome by using Wikipedia, which is expected to have a larger coverage of terms compared to WordNet [Gelbukh, 2014]. The main disadvantage of fully unsupervised systems is that, because they do not exploit any dictionary, they cannot rely on a shared reference inventory of senses [Sreedhar et al., 2012].

Through word sense disambiguation experiments performed on a Wikipedia-based sense-tagged corpus generated for a subset of the SENSEVAL ambiguous words, Mihalcea [2007] show that the Wikipedia annotations are reliable, and that the quality of a sense-tagging classifier built on this data set exceeds by a large margin the accuracy of an informed baseline that selects the most frequent word sense by default. They use a word sense disambiguation system that integrates local and topical features within a machine-learning framework. Their disambiguation algorithm starts with a process where text is tokenized and annotated with part-of-speech tags. Then local and topical features are extracted from the context of the ambiguous word, like the part-of-speech of the current word. Finally, the collected features are integrated in a Naive Bayes classifier. After evaluating their supervised approach to word sense disambiguation, Mihalcea [2007] reported that Wikipedia sense annotations can be used to build a word sense disambiguation system, leading to a relative error rate reduction of 30–44%, as compared to most frequent word sense and Lesk-corpus approaches.

2.4 Defining Technical Terms

Automatic term definition is the process by which a computer provides the correct definition of a domain-specific vocabulary. The approach to automatic term definition may be supervised or unsupervised. In the supervised approach, a lexicon knowledge database with all possible definitions of terms is generated and then the computer picks the most relevant definition of the term with respect to its context. Automatic term definition therefore depends on the techniques of term sense disambiguation in order to provide relevant definitions in context.

For the case of unsupervised term sense disambiguation, the computer tries to find the correct definition of a term from the document where it occurs without the help of an external lexicon knowledge database. The unsupervised approach therefore

uses linguistic features like the part-of-speech and words surrounding the technical terms in order to provide a definition for it. For example, if the technical terms occur in a sentence with any such phrases as *is a*, *is the*, *is defined as*, *refers to*, or *is explained as*, then the sentence is a possible definition of the term in question as shown in example 2.4.1.

Example 2.4.1: Sentence defining the term chambers

A judge's chambers, often just called his or her chambers, is the office of a judge².

Torralbo et al. [2005] present a piece of work on automatic term definition from the perspective of document summarization. They exclusively use knowledge bases to find summaries for already existing technical terms. The purpose of their proposed system was as follows: given a list of domain-specific terms, generate a definition of each of the terms. It can be thought of as a complement of term identification procedures, so they are able to provide a small definition of each of the new terms identified. In this approach, they use the World Wide Web, as the source of documents, so the module can be applied to every domain for which there is information on the Web. The scope of the World Wide Web is too huge and it is likely that Torralbo et al. [2005] noticed this as well when they say that, "this source is not fully reliable, as it may contain inexact or erroneous information." Working with authoritative documents to define technical terms would be a way to improve the accuracy of automatic cross-references. As mentioned, work done by Torralbo et al. [2005] crucially depends on an existing list of technical terms but there are methods devised to define technical terms without a predefined list of terms [Androutsopoulos and Galanis, 2005]. These methods have the advantage of defining new terms, which may not be included in the list of available terms. This would also be an attractive approach for acquiring relevant definitions for each term as used in a particular document.

Sometimes even when a technical term is defined within the text, the definition contains other vocabulary items, which may not be known to a reader who is not conversant with the knowledge domain of the text. The definition in such a case will not provide sufficient clues to reading comprehension but rather compound the problem because of the vocabularies used. As seen in example 2.4.2 additional contextual information would be necessary for the reader to understand such a definition. The example presents the following domain-specific (medical) terms:

- i. *epidural hematoma*,
- ii. *hematoma*,

²Sentence retrieved from [http://en.wikipedia.org/wiki/Chambers\(law\)](http://en.wikipedia.org/wiki/Chambers(law)), on 19 December, 2012

iii. *dura mater*, and

iv. *cranium*

The example also presents the abbreviation “*EDH*” and it is clear that this abbreviation refers to *epidural hematoma*, but what exactly does this mean? Although the domain-specific term *epidural hematoma* is defined in the sentence example, the definition still makes no sense if the medical terms *epidural*, *hematoma*, *dura mater* and *cranium* do not “ring a bell” in the reader’s mind. We see that sometimes even when technical terms are defined within text, additional background information would be necessary for reading comprehension.

Example 2.4.2: Futile term definition

Epidural hematoma (EDH) is a rapidly accumulating hematoma between the dura mater and the cranium.³

2.5 Background Information for Technical Terms

In case a reader lacks prior knowledge of a particular technical term, this can be inferred from external sources like the Internet, glossaries, dictionaries and encyclopedias. In practice these inferences can be made accessible by automatic link generation or automatic hypertext generation where the technical term is used as the link text which refers to additional information. This helps to provide easy access to background information for technical terms with which readers are unfamiliar and saves time in the process of filling the knowledge gap. It is also less cumbersome because the reader gets a direct link to background information and does not necessarily have to make extra effort to look in other knowledge sources if the technical term is incomprehensible.

Automatic link generation is the process by which a computer seamlessly creates hypertext, which refers to contextually relevant information for the created hypertext. The objective of automatic link generation is to create meaningful links for some n-grams of text. N-grams used in automatic link generation may be single words or phrases, which represent major concepts within the text. It is therefore logical to point out that term recognition techniques can be applied for predicting which sections of a text document can be used for automatic link generation. An example of research in which links are generated automatically is Hyperwords. Hyperwords

³Sentence examples was taken from, http://en.wikipedia.org/wiki/Head_injury. on Aug., 2008

is a project that extensively enriches the World Wide Web with information from on-line knowledge bases.

“Hyperwords are interactive words; words you can issue commands on. (This concept is very broad and particularly designed to help knowledge workers move more effectively through their information environment, gather useful information and have access to better tools)”⁴

It is definitely necessary to have relevant and sufficient information but making every word interactive would provide too much information and most likely obstruct the reader’s ability to follow the documents by drawing attention away from important concepts. Providing relevant information for only important concepts in relation to text context would be more helpful.

It was motivating to consider the work done by Rotard et al. [2007] under the title, “*Semantic lenses: Seamless augmentation of web pages with context information from implicit queries*”, who say:

“In analytical processes huge amounts of complex data have to be collected, filtered, and provided to the users in an appropriate way”.

Rotard et al. [2007], emphasize the representation of appropriate data. Their work indicates that available contextual information can be seamlessly integrated into text and visualized for easy comprehension.

Another work in relation to automatic link generation is one done by Goffinet and Noirhomme-Fraiture [1995]. This work deals with the problem of automatically generating cross-reference links when converting text to hypertext. They use a statistical approach, based on techniques commonly used in Information Retrieval. They also use complementary probabilistic methods in order to create more relevant links than the ones generated without any knowledge.

2.6 Visual Clues for Technical Terms

Picture illustrations provide visual clues to text meaning and consistently improve reading comprehension and learning performance [Chongtay et al., 2013]. Much as visual clues are important, Chongtay et al. [2013] argue that such clues need to be complemented with text information to make them concrete and more meaningful. Considering the importance of using visual clues to text, Aramini et al. [2015] presented an unsupervised method for illustrating short text inputs. They applied the idea of automatically providing visual clues to text by using keywords

⁴See, <http://www.hyperwords.net/definition.html>. Referenced in Aug., 2008

(technical terms) from the text to search for web pictures that provide illustrations for the text content. Their concern was to exploit photos and their captions from Google image search engines in order to avoid the use of a predefined annotated image dataset and the tedious task of manually selecting relevant images from a list of retrieved images. During the task of extracting keywords in short text, the system proposed by Aramini et al. [2015] also returns a model of the keywords semantic meaning (Text Semantic Space). Other information retrieved along with images from the web by this system is processed to create a model of the concepts related to such images (Image Semantic Space). The Text Semantic Space and the Image Semantic Space were then intersected to find a Common Subspace, for ranking images based on words from the subspace. The system's test results show good results for automatic text illustration. Although Aramini et al. [2015] are not directly concerned with enhancing reading comprehension, it is enlightening to see that they successfully used keywords as a lead to relevant pictures for illustrating text. This shows that technical terms embedded in documents are an asset in automatically providing visual clues which are relevant to text context, which can enhance reading comprehension.

Joshi et al. [2006] presented another unsupervised approach to automatic text illustration with an interest in story picturing. They applied the idea of extracting semantic keywords to a story text and used the extracted keywords to search for relevant pictures from an annotated image database for illustrating the story. The Story Picturing Engine presented by Joshi et al. [2006] consisted of three components (i) the story processing and image selection, (ii) the estimation of similarity between pairs of images based on their visual and lexical features, and (iii) the mutual reinforcement-based rank estimation process. Mutual reinforcement is one of many graph-based ranking algorithms that are used in the context of document summarization [Gao et al., 2012] where each entity contributes towards the rank of others based on some similarity among them. Joshi et al. [2006] used this idea to assign a measure of importance to individual images based on their similarity to other images then they select the images with the highest ranks for story-picturing. They carried out a story-picturing user survey where humans manually selected images for picturing the same stories as were pictured by the Story-Picturing Engine. The humans also used pictures from the same annotated image databased as the Story-Picturing Engine during the survey. The survey results showed that the engine performed equally well, on average compared with a human being in story-picturing. This indicates that is possible to automate the process of providing visual clues to text and still get averagely good results as would be done manually.

Concerned by the difficulty in understanding medical documents, Bosma et al. [2008] presented a method that extends the answers returned by a QA-system with appropriate illustrations. Their method was applied in Interactive Multimodal

Information eXtraction (IMIX) and the picture illustrations were generated from *Merck Manual Medisch Handboek* [Berkow et al., 2005] and *Winkler Prins Medische Encyclopedia* [Barbier et al., 1972]. In their method, for each textual answer two illustrations were provided, one for which the textual answer and the pictures caption was used as associated text, and the other for which the picture was associated with the smallest unit of surrounding text from its original document. The purpose of the picture illustrations in [Bosma et al., 2008] was limited to the medical knowledge domain. After evaluating their methods for picture selection, they realized that their picture corpus was limited and did not contain illustrations for many questions' answers. Therefore, it is necessary to use a big picture corpus when considering the problem of automatically providing visual clues to text.

Chapter 3

The Need for Document Enrichment

By consultation plans will succeed.

Proverbs 20:18 NWT

This chapter reports a stakeholders' survey that was carried out to ground the problem caused by technical terms in the reading process of university students. The report highlights key factors hindering the reading comprehension process of students at Ugandan universities and the general effect of the factors on the process. It also discusses the stakeholders' views towards facilitating the reading comprehension process, given the challenges they experience in it. The chapter concludes by ascertaining that a document enrichment approach would be appropriate for facilitating reading comprehension according to the reviewed literature and the stakeholders' views.

3.1 Exploratory Survey Objectives and Approach

Reading comprehension is a complicated process of constructing meaning from text and is successful when a reader can decode the author's words and use his or her prior knowledge to construct the relevant meaning of the message. For this to happen, the reader needs to have developed sufficient reading comprehension skills. It is normally taken for granted that students at university level have sufficient reading skills, which can help them to learn from textbooks without any problems [Brophy, 2009]. Unfortunately, this is not the case. We saw in chapter 1 of this thesis that it is possible for students at universities to be affected by the factors that hinder reading comprehension. The objective of the exploratory survey was to understand the key process that students at Ugandan universities go through in reading their course content, the key factors that hinder their reading process and the requirements necessary for facilitating the reading comprehension process.

The exploratory survey employed a qualitative research approach, because it is concerned with the development of concepts, which help us to understand a social phenomenon in its natural rather than experimental setting [Pope and Mays, 1995].

This gives due emphasis to the meaning, experiences and views of the participants. These sentiments were gathered by e-mailing open-ended interview questionnaires to the students in various Ugandan universities. The interviews helped us to generate in-depth qualitative data from a small number of students compared to their population at both public and private universities in Uganda. The number of students involved in the stakeholders' survey was small because qualitative methods collect a lot of unstructured data. This data can be challenging to analyze if too many responses are collected. Like analyses of all other qualitative data, the analysis of this data was drawn from concepts expressed by the respondents and insights from the researcher. Analysis results were illustrative explanations and individual responses. The population sampling was theoretical because the original decisions are not based on a preconceived theoretical framework. E-mail interviews were used because we wanted to reach many students in various Ugandan universities without transport costs. Besides, email reduces interview bias since the respondent is not facing with the researcher, and the respondent has time to think about their response before writing it down. The data collected would be electronic and will be easier to analyze. In addition, the problem of interviewer's accent does not arise so interview questions would be correctly interpreted.

The qualitative data collected was analyzed using thematic content analysis. This included data familiarization, pattern construction, theme searching, themes review and their documentation [Braun and Clarke, 2006]. Patterns were constructed in order to build a list of relevant topics that arose from the interviews. The constructed patterns were considered important to reading comprehension of university students. The list of relevant topics was then iteratively grouped into themes. The on-line interview targeted undergraduate and postgraduate students from both public and private universities in Uganda. Questions to the undergraduate students addressed their opinions directly and their experience in reading course content while questions to graduate students were directed at their opinions and their experience in reading research material.

The reading comprehension e-mail interview questionnaires were divided into seven sections. Questions in section 1 were designed to collect information about the respondents' biographic data. Section 2 was designed to explore reading motivation, so that responses to the questions in this section would indicate what environments, situations and course content motivates students at Ugandan universities to read. Section 3 was dedicated to finding out the form in which course content is used at Ugandan universities and what form is preferred including the reason for preference. In section 4 the researcher intended to find out if the respondents used electronic devices for reading, and if so, which kinds and what else the respondents do while using these electronic devices. Section 5 was specifically dedicated to investigating

the respondents' experiences of reading difficult course content. This section helped in understanding the effect of respondents' background knowledge on reading difficult course content, their behavior while reading difficult course content including their fluency and efficiency at decoding and recognizing words. Interview questions in section 6 gave the respondents an opportunity to participate in the design activity of a reading device that would meet their needs especially while reading difficult course content. Section 7 was a wrap-up section that allowed the respondents to express what they felt the interview questionnaire did not address regarding difficult course content.

3.2 Presentation and Discussion of Results

The presentation and discussion of results helps us to move from the data we collected into explaining, understanding and interpreting the people and situations we are investigating. This may involve not only the primary message but also the evaluative attitude of the respondent towards the message including whether the content of the message is meant to represent individual ideas, or the degree to which the respondent is presenting actual versus hypothetical experiences.

We received 60 responses from students at Ugandan universities and used 40 of them. The criteria for selecting the responses to be used were based on their completeness and relevance to the reading comprehension process. The 40 selected responses were from students at 3 public universities and 2 private universities. The public universities included Gulu University (GU), Makerere University (MUK) and Muni University (MU) while the private universities included Uganda Christian University (UCU) and Uganda Technical and Management University (UTAMU). The majority of the selected responses were from public universities, representing 55.0% of the total percentage of selected responses as shown in table 3.1. We consider a rather large number of respondents compared to other qualitative research because it provides the resource for carefully examining the characteristics which address our research questions. The large number of responses also helped to clearly distinguish conceptual categories of interest that emerged from the data collected and increased the prospect of resolving relationships among such categories [Dworkin, 2012].

Table 3.1 shows that there were 14 female and 26 male students among the selected respondents. Imbalance in the gender numbers of the selected respondents does not cause concern since analysis is based on individual opinions of the respondents [Sullivan and Sargeant, 2011]. The majority of these respondents were in the age group between 20 and 29 years, indicating that they are at the right age for pursuing undergraduate studies. It is evident that most of the selected responses were taken from undergraduate students since they account for 60% of such responses.

Table 3.1: Biographies of the respondents

Sl. No.	Bio-characters	Frequency	Percentage
<i>Gender</i>			
1.	Female	14	35
2.	Male	26	65
<i>Age Groups</i>			
1.	19 and below	1	2.5
2.	20 to 29	24	60.0
3.	30 to 39	8	20.0
4.	40 to 49	7	17.5
<i>Institutions</i>			
1.	Gulu University	8	20.0
2.	Makerere University	3	7.5
3.	Muni University	11	27.5
4.	Uganda Christian University	1	2.5
5.	UTAMU University	17	42.5
	Total	40	100.0
<i>Level of Study</i>			
1.	Master Students	8	20.0
2.	PhD Candidates	3	7.5
3.	PhD Students	4	10.0
4.	Postdoctoral Candidates	1	2.5
5.	Undergraduate Students (Year One = 11), (Year Two = 9), (Year Three = 4)	24	60.0
	Total	40	100.0
<i>Knowledge Domains of Interest</i>			
1.	Bachelor of Information Systems	12	30.0
2.	Bachelor of Information Systems and Technology	8	20.0
3.	Bachelor of Science in Computer Science	4	10.0
4.	Adolescent Health	1	2.5
5.	Computing	2	5.0
6.	eHealth	1	2.5
7.	E-learning and MOOCs	2	5.0
8.	Fraud in eBanking Institutions	1	2.5
9.	GIS and Remote Sensing and Spatial Analysis	2	5.0
10.	Information Systems	1	2.5
11.	Information Technology	3	7.5
12.	Knowledge Management	1	2.5
13.	Networked Learning and Problem Based Learning	1	2.5
14.	Project Management	1	2.5
15.	Spatial Analysis	1	2.5
	Total	40	100.0

Besides undergraduate students, other respondents included, masters and PhD students and PhD and postdoctoral candidates. The age groups of selected respondents indicated that they are at the right age for studying at Ugandan

universities. Their opinions and experiences should reflect the actual situation in reading comprehension at universities in Uganda for all sorts of students.

Most of the selected responses were from undergraduate students pursuing courses related to Information Technology. For example, 30% of the selected responses were from students pursuing a Bachelor of Information Technology. The selection presented a good group to work with because they were familiar with computer technology and electronic devices. They had the required knowledge for answering the on-line questionnaire and to give relevant experiences and opinions in connection with reading on electronic devices. As shown in table 3.1, most of the undergraduate students who responded to the on-line interview were in their first year of study. This is an advantage because one of our research concerns is that such students are confronted with new subjects for which they do not have sufficient background and vocabulary knowledge for successful reading comprehension [See section 1.1]. Their responses helped to confirm this concern.

The selected postgraduate respondents were interested in 12 knowledge domains as listed in table 3.1. Majority of whom were interested in Information Technology 7.5%. Other subjects of interest included e-Health, e-Banking, Project Management and Spatial Analysis. Since the respondents are interested in different knowledge domains, it is interesting to see if they have similar opinions about reading comprehension independent of their subject interests.

3.3 Key Processes in Reading Comprehension

The literature review carried out in chapter 1 reveals that the key processes in reading comprehension include the ability of a person to process text, understand and relate it to what he/she already knows. Analysis of the qualitative data ascertains these key processes as seen in the following response:

I read and conceptualize the subject matter for the purposes of understanding and applicability. (Master student)

The above response also implies that the ability to retain and apply knowledge acquired from read texts are important processes in reading comprehension for students at Ugandan universities. However, the ability of a student to process text is vital because the other key processes in reading comprehension depend on it. When students are able to process text, they can then understand and integrate it with their prior knowledge, including retaining and applying acquired knowledge.

The complex nature of reading comprehension demands prior-preparation and application of certain skills for its success. The following emergent themes from the

analyzed qualitative data indicate the preparation and skills applied by students at Ugandan university to support the reading comprehension process.

Setting Up Goals and Objectives for Reading

As university students prepare to read, they setup goals and objectives. That is seen in the following response of an undergraduate student.

I start asking myself questions and then I set up my goals of reading that course content.

Goals and objectives are set by university students in order to help them take control of their own learning and get a sense of accomplishment when the goal is achieved. A reading plan is important because university students normally have a lot to read in a short time. This is seen in the response by a PhD student that “*there is too much to read within the same time*”.

Good Vocabulary Skills

A person with strong vocabulary skills is able to easily recognize most of the words in read text and understand their meaning. Strong vocabulary skills are essential for understanding as students at Ugandan universities read. Its relevance was confirmed by responses from undergraduate students such as:

I understand every word in a sentence before, to understand it (my course content) better;

I try to separate sentences into portions and find meaning of words, to get the right interpretation of what am reading.

These responses indicate the importance of the bottom-up reading comprehension model as discussed in subsection 1.2 of this thesis by showing that recognizing the smallest units of text is necessary for constructing meaning from text.

Making Written Summaries

People reading at Ugandan universities make written summaries of what they read in order to help them understand what they read and retain the knowledge they acquire [Bazerman, 2015]. This theme was supported by the following responses.

I make summaries directly on the hard copy of my course content, increasing memory rate. (Undergraduate student)

I read a few concepts and note down some ideas from the reading to easily remember what I really read and what I got from the reading. (Master Student)

I try to create summaries of key points I can understand. (PhD student)

Written summaries also help people reading at universities to put new ideas in their own words and give them the opportunity to make notes for easy overview.

Active Reading

When people read with an awareness of the purpose for their reading and make intentional efforts to understand what they read, it is said that they are active readers. Active reading also involves deliberate evaluation of the relevance of the document being read to the reader's needs. The following students' responses verify that people at Ugandan universities apply active reading skills during their reading process:

While reading difficult research material, I skim and scan through it first;

I synthesize the main points in the text/content am reading. This will help me to pull together the pieces of what am learning at (different) points during the journey of reading my course content;

I start asking myself questions then I set up my goals of reading my course content.

These responses also show that people at universities are critically involved in what they read. Skimming and scanning through documents helps them to get an overall impression of a document content and to find the information, which is important for their reading purpose. Active reading can help people at universities to achieve their reading purposes through asking questions as they read. Reading a document in a focused and speedy way is another active reading technique, which is used by people at universities to extract meaning from document content. People reading at universities also test their memory as part of active reading, that is why we saw earlier that they make written notes as a review of their reading. Writing notes can help them to find out if they remember what they have read. The skill of active reading is generally applied by people at Ugandan universities to save time [Wilfrid Laurier University, 2008] during their reading process.

Searching for Clues to Text

Another theme that emerged from the qualitative data analysis to show what happens during the reading process of people at Ugandan universities was that they search

for clues to understanding. This happens when they find documents, which are difficult to understand. The following responses from undergraduate students show that people at universities look for word meanings and additional information on certain topics.

I use electronic devices for reading my course content because they provide a wide range of reading resources like dictionaries to help me understand content;

While reading from an electronic document, I can be searching for more information on a particular topic from the Internet.

These responses show that people at Ugandan universities use electronic devices while reading. More important than that, the responses show that some concepts in the documents read by people at Ugandan universities are difficult for them to understand and that they have insufficient vocabulary for recognizing some of the words in such documents. A detailed presentation of key factors that hinder reading comprehension and lead the stakeholders to a search for clues to text is made in the next section.

3.4 Key Factors Hindering Reading at Universities

The themes that emerged from the qualitative data analysis to indicated key factors hindering the process of reading comprehension at Ugandan universities were as follows:

Insufficient Background Knowledge

Students at Ugandan universities ascertained that a key factor that hinders their reading comprehension process is insufficient background knowledge in certain knowledge domains as seen in the following responses:

Research material is made difficult to understand if you don't know anything about the field. (Master student)

Course content is made difficult to understand because they contain new and strange things that you have never come across which can sometimes be complex. (Undergraduate student)

Besides these responses indicating that insufficient background knowledge in specific knowledge domains hinder reading comprehension of students at Ugandan universities, they also show the importance of the top-down model of reading comprehension. This model argues that for reading comprehension to be successful,

the reader needs to bring what they know to the text. [See section 1.2] Insufficient background knowledge in a specific field triggers the search for information outside the course or research material read by university students. That is why an undergraduate student said he uses electronic devices for reading because:

At least from electronic devices when there is Internet, I can go on-line and research more on what I find difficult to understand.

Undergraduate students emphasized that insufficient background knowledge is a factor that hinders reading comprehension, as was anticipated in chapter 1 of this thesis. Actually, an undergraduate student said that:

Over time I have learned smart reading because some of the course contents are really unique compared to high school.

Insufficient Vocabulary Knowledge

Stakeholders at Ugandan universities indicated that they have insufficient vocabulary knowledge because they come across unfamiliar vocabulary in their course or research material and that such vocabulary hinders their reading comprehension process. For example, they said the following about the things they can do while reading from electronic devices that they cannot do with printed text:

When I find a word I do not understand, it is easy for me to check for the meaning on-line or with mobile dictionary. (Undergraduate student)

Searching for alternative meanings to some terms. (Master student)

This shows that students at Ugandan universities need to consult reference resources outside their course content for definitions and explanations of unfamiliar vocabulary because they “*need to learn and fully understand concepts.*” (Undergraduate student) The stakeholders ascertained that unfamiliar vocabulary hinders their reading comprehension process because such vocabulary makes course and research material difficult to understand. This analysis was indicated by the following responses:

Research material is made difficult to understand when there is use of too many complex terms and methodologies. (Master student).

Course content is made difficult to understand because of hard language with strong words that might be hard to interpret. (Undergraduate student)

Insufficient vocabulary knowledge hinders the reading comprehension process of students at Ugandan universities because of the following reasons; note that these reasons are discussed in no specific order of importance.

(i) Reading takes too much time and effort

Having to stop and find the meaning of unfamiliar vocabulary slows down the reading process. Contemplating the meaning of unfamiliar vocabulary slows down the rate at which students understand their read course content, leading them to divert a lot of their mental processing capacity. These aforementioned analyses are indicative in the following responses from undergraduate students.

Vocabulary that I am not familiar with affects the way I understand my course content because it makes me a slow learner;

Unfamiliar vocabulary affects the way I understand my course content because it slows down my rate of understanding while reading and makes me think a lot.

In essence the reading comprehension process is slowed down if students cannot recognize certain words at a glance so they try to sound them out as indicated in this response:

I overcome the difficulties in reading my research material by reading aloud so that I keep myself reminded. (Master student)

They hope that sounding out the words will help them remember but on the contrary it makes it harder for them to remember all the words in a sentence and understand how they fit together.

As I read difficult course content, I push my mind to the limit and try to absorb it. (Undergraduate student)

(ii) Loss of interest in reading

When the meaning of unfamiliar vocabulary cannot be resolved, it leads to a loss of interest in reading as ascertained in the following responses from undergraduate students.

At times some words are hard to understand which leads me to first check in a dictionary or at times I don't which makes me lose interest.

Unfamiliar vocabulary lowers my morale to continue reading.

(iii) Misinterpretation of course content

Unfamiliar vocabulary hinders the reading comprehension process by causing university students to misinterpret the course content of their reading materials. This analysis is confirmed by the following undergraduate responses as they explain how unfamiliar vocabulary affects their understanding of course content.

Unfamiliar vocabulary can cause misinterpretation but with electronic devices you can get other explanations for that vocabulary;

I may derive a different meaning out of information due to the vocabulary used;

It gives me hard time that what I perceive becomes contradictory to what is true.

In addition to the above undergraduate responses, a PhD candidate also said that:

Unfamiliar vocabulary affects my understanding of research material because I don't get a proper understanding of the material.

(iv) *Causing confusion*

Unfamiliar vocabulary also hinders the reading comprehension process of people at Ugandan universities because they make course and research material confusing as seen in the following response.

Unfamiliar vocabulary affects my understanding of course content because it confuses me and makes the whole document boring. (Master student)

(v) *Inability to construct phrases and sentences*

Another way through which unfamiliar vocabulary hinders the reading comprehension process of people reading at Ugandan universities is that such vocabulary makes it difficult for them to combine words into phrases. Unfamiliar vocabulary also makes it difficult for these people to construct sentences. This analysis was prompted by the following response:

Vocabulary that I am not familiar with affects my understanding of research material in such a way that it is a stumbling block to the construction of my sentences and phrases as well. (PhD student)

This response supports the argument of the bottom-up model of reading comprehension. The model argues that it is important to recognize individual words in sentences in order to join these into phrases and sentences and eventually make sense of the concepts expressed by an author in a piece of writing. [See Chapter 1]

(vi) *Gaps in the reading process*

Insufficient vocabulary knowledge hinders the reading process of students at Ugandan universities by creating gaps during the process. These gaps are

created because the readers do not understand the meaning of certain words in their reading material. As they try to find the meaning of such words, they lose focus of their reading and get disconnected from what they are reading. This analysis was indicated in responses such as:

Unfamiliar vocabulary distorts my reading and creates gaps in the process. I need to consult the dictionary so often that I lose track of the main argument; (PhD student)

Unfamiliar vocabulary may disconnect me from the flow of my reading requiring me to first understand it and then continue. (Master student)

(vii) Building vocabulary levels and motivation to learn

On the positive side, we saw that it is important to have unfamiliar vocabulary in the stakeholders' course content because it helps them to develop their vocabulary levels and motivates them to learn as seen in the following responses to how unfamiliar vocabulary affects the reading process of the stakeholders.

I can read those new words and expand my own vocabulary. (Postdoctoral candidate)

Unfamiliar vocabulary motivates me to get their meaning. (Undergraduate student)

An undergraduate student explains that:

"Unfamiliar vocabulary does not affect my reading process because electronic devices always have dictionaries, if not then I go on-line."

This shows that electronic devices used in reading can help students at Ugandan universities to easily find the meaning of unfamiliar vocabulary. In turn this helps them to understand their course or research material. This interpretation is supported by the following response:

My experience of reading difficult course content is that sometimes I find other things very difficult but when I go on-line and consult, they become better. (Undergraduate student)

It is indicative that if the meaning of unfamiliar vocabulary can be easily found, the reading comprehension process is less affected and stakeholders are helped to expand their vocabulary knowledge. However, analysis of the qualitative data indicates that insufficient vocabulary knowledge has a largely daunting effect on the reading process of university students.

Insufficient Clues to Text

Analysis of the qualitative data showed that another factor that hinders the reading process of university students in Uganda is the insufficiency of the clues to text as shown in the following responses:

My course content is made difficult to understand because it is not straightforward and some have fewer examples. Therefore, it's hard to relate to; (Undergraduate student)

My research material is made difficult to understand if it is abstract, not clearly presented. (PhD candidate)

Course or research material with insufficient clues to text hinder the process of reading comprehension because they do not provide enough information to understand the written concepts. This is a fundamental hypothesis of the situation model of reading comprehension [Johnson-Laird, 1981]. It was indicated by the following response:

My course content is made difficult to understand because they do not provide information to make me understand the concept. (Undergraduate student)

In order to counteract the effect of insufficient clues to text, the stakeholders look for visual and audio clues and create mental pictures as indicated in the following responses:

To help me understand difficult course content, I normally use visual and audio aids; (Undergraduate student)

As I read difficult research material, I create an imagination (mental picture), which makes me tired soonest. (PhD student)

It is indicative that visual clues can help the students in understanding research material that has insufficient clues to text. As argued in section 1.2, clues to text, which include visual images help the comprehension process by facilitating the creation of mental pictures during reading. The PhD's response ascertains that mental pictures are created in the mind of the students at Ugandan universities as they read. This response also indicates that it is exhausting to create mental images from content with insufficient clues to text.

Difficult Writing Styles

The stakeholders ascertained that sometimes authors have difficult writing styles, which cause problems in understanding their course or research material. Such writing styles include the use of long sentences and complex grammar as

indicated in the following experiences of reading difficult course content by students at Ugandan universities.

My experience of reading difficult course content is that at times the style of writing of some content makes reading hard; (Undergraduate student)

The grammar and writing skills used make research material difficult to understand; (Master student)

My research material is made difficult to understand because of long sentences and the ambiguity of the words; (PhD candidate)

My research material may be made difficult to understand because of the text structure. (PhD student)

Grammar complexity is measured using sentence length, word frequency and/or word length [Murphy, 2013], among other things. Long sentences are in general regarded as confusing and difficult to understand [Sandra et al., 2008; Graham and Bellert, 2005; Graves and Graves, 2003]. Such sentences are wordy and they lack focus and proper structure compelling a reader to search for the writer's intended meaning in long phrases. The PhD candidate's response indicates that difficult writing styles come along with ambiguous words. Ambiguous words are those that can have various meaning depending on the context where they are used. Such words can hinder the reading process if a reader does not recognize their meaning in the context of the read course content. Besides, while searching for the meaning of an ambiguous word, it is possible to find the meaning that does not relate to its use in the read course content. Interestingly, the qualitative data analysis indicates that ambiguous terms hinder the reading process of students at Ugandan universities, because it is one of the problems that motivated our research. [See section 1.3]

Too Many Time Bound Reading Tasks

Too many time bound reading tasks is another factor that was found to hinder the reading comprehension process of students at Ugandan universities. Undergraduate students gave the following explanation regarding the reasons why their reading habits have changed over time.

My reading habit has changed over time because there is too much expected from me, which involves course works, so I end up concentrating on some course units and ignoring others;

My reading habit has changed over time because of the pressure to have knowledge and also excel in my exams.

Therefore, undergraduate students changed their reading habit to acquire a lot of knowledge in a short time in order to complete time-bound tasks like course assignment, test and examinations. This analysis agrees with the argument that university students need to read and comprehend large volumes of their course content in a short time. [See section 1.1] This analysis was directly indicated by the following undergraduate response:

My reading habit has changed over time to make me acquire more knowledge at a high-speed rate.

Too many time-bound reading tasks hinder the reading process because they are overwhelming and do not support knowledge retention as indicated in the following undergraduate students' responses.

My course content is made difficult because it's so hard to find what is necessary to read since there is so much content, some of which may be unnecessary and

Too much to read interferes with my memory capacity.

Unfamiliar Phrases, Acronyms and Abbreviations

Students at Ugandan universities explained that unfamiliar phrases, acronyms and abbreviations are other factors that hinder their reading process. The problem is caused by the lack of prior knowledge of such text elements as in the case of unfamiliar vocabulary. Therefore, they do not have the possibility to draw inferences from their background knowledge to bring meaning to the content where such text elements occur, as discussed in chapter 1 of this thesis. Analysis of the qualitative data shows that unfamiliar phrases, acronyms and abbreviations disrupt the reading process of university students. The problems are similar to those caused by unfamiliar vocabulary. These include causing confusion, preventing understanding, affecting content interpretation, slowing down reading and causing a loss of interest in reading. These interpretations were indicated in the following responses:

Unfamiliar phrases delay the interpretation of course content where you have to take time to understand the phrase before understanding the statements in which it is used;
(Undergraduate student) and

Sometimes they bring about confusion. (Master student)

A unique response about the effect of unfamiliar phrases, acronyms and abbreviations on the reading process of students at Ugandan universities was that such text elements does not support knowledge retention as shown by the following responses.

Some lengthy ones that I can't memorize make me keep referring to them and hence I lose interest in the long run. (Master student)

Unfamiliar phrases, acronyms and abbreviations make me forget each and everything. (Undergraduate student)

3.5 General Effects of Reading Difficulties

We have discussed specific effects of the factors hindering the reading process of students at Ugandan universities. For example, the behavior of students at Ugandan universities while reading difficult course content reveals that insufficient vocabulary and background knowledge affects their reading fluency and efficiency. Such reading difficulties make reading a tedious experience because the stakeholders have to critically read through the same text multiple times while looking for the meaning of new concepts and unfamiliar vocabulary. A close look at the qualitative data reveals that there are common effects to the reading process caused by the factors hindering it. These effects are presented in this section in no specific order of importance.

Anxiety and Stress

Factors hindering the reading process of students at Ugandan universities make them anxious and stressed as ascertained in the following response.

What happens in my mind as I read difficult course content is that, I generally feel that I will not understand it anymore because some of the things are really so hard that I try more and more reading until I understand them. It gives me the impression, if it's like this at this level, how about the next level (year 2)? (Undergraduate student)

Another undergraduate student explains that the following questions run in his mind as he reads difficult course literature:

*Will I make it?
Will things enter my brain? Will
I remember?
Will the things be examined?*

University students are anxious to understand and remember what they read and be able to apply the knowledge they acquire in other areas of their study. The reason these anxieties arise is because they find some of their content too difficult to understand, making them doubt whether they will meet their reading expectations.

When university students are overwhelmed with anxiety as a result of reading difficulties, they experience stress. For example, an undergraduate student explains that:

The way I read has changed over time such that I am overstressed.

Other undergraduate students explain that:

I feel heated up and overworked in my mind as I read difficult course content and

My experience is that it is stressing and annoying to read difficult course content, it wastes time.

The above explanations ascertain that difficult course content stresses students at Ugandan universities because they cost a lot of time and energy to read and understand.

Frustration

Difficult course content frustrates people reading at Ugandan universities as indicated by the following experiences:

The difference between using my electronic device for reading to reading printed copies of my course content is that it keeps my brain more active. Sometimes when I meet some hard content, I even almost give up but then when I go on-line, I definitely get the explanations; (Undergraduate student) and

Difficult course content is hard to understand because I cannot relate it to the research topic. This makes me feel frustrated. (Master student)

The above experiences agree with the argument that lack of prior knowledge in specific knowledge areas hinder the process of reading comprehension and can be frustrating, often leading students to give up on their reading. The experience of this undergraduate student in particular shows that the provision of background information for difficult concepts can facilitate reading comprehension, since they are able to get explanations for difficult course content on-line.

Deviated from Read Text

Majority of the stakeholders ascertained that they overcome reading difficulties concerning unfamiliar text elements by looking for the meaning of such elements in various reference works as indicated in the following experiences:

When I find a word I do not understand, it is easy for me to check for its meaning on-line or with mobile dictionary while reading on an electronic device. Something that I cannot do with printed copies of my course content; (Undergraduate student) and

When I come across unfamiliar phrases, acronyms and abbreviations, I use the Webster or Oxford advanced dictionaries for reference. (PhD student)

These experiences indicate that the stakeholders search outside the text they are reading for the meaning of unfamiliar text elements like new vocabulary, text phrases, acronyms, and abbreviations. In agreement with this analysis the following undergraduate students explain that they search for the meaning of unfamiliar vocabulary while reading on electronic devices by:

Swapping though application programs; and

Copying the word I do not understand and look for its meaning.

Students at Ugandan universities are distracted from the text they are reading as they try to find the meaning and explanation of unfamiliar text elements that exist in their course content. The need to find the meaning of unfamiliar text elements also wastes productive time and costs effort because the meanings of unfamiliar text elements are looked for on the on-line, in printed dictionaries, textbooks, and in conversation or communication among peers. Looking for such information on-line can make university students lost in the vast sea of information on the Internet. A concern that was mentioned in section 1.5 as a weakness of contemporary document enrichment techniques. This weakness is evident in the following undergraduate students' opinions:

What I dislike about the electronic device I use for reading is that I can easily get distracted by the Internet;

I dislike the electronic device I use for reading because most times there is lack of in depth reading and concentration due to multi-tasking.

A lecture directly states that:

I dislike the electronic device I use for reading because it takes attention away from the environment.

It is clear from the above opinions that the Internet deviates focus from the reading material and that multi-tasking on electronic devices used for reading affects the level of concentration during the reading process. Loss of focus and deviated attention affects course content comprehension, which is the ultimate goal of the reading process.

3.6 Reading Process Facilitation: Respondents' View

Considering that the contemporary culture of people around the world tends to reading on electronic media [Liu, 2006], we were concerned with facilitating the reading process on such media. The following undergraduate student's response suggests that even students at Ugandan universities mostly read electronic copies of their course material.

The way I read has changed overtime in that I read more over the Internet for research purposes than looking for printed course materials. I have been introduced to many websites that enhance course learning on-line and memorization of what is being taught in class. My reading habit has changed mostly because of the pressure to have knowledge and also excel in my exams.

In order to find out the characteristics of an ideal electronic reading device that would facilitate the reading process, we asked the participants to provide their opinions about the functions of such a device. The elementary characteristics of a good electronic reading device were summarized in the response of the following master student:

I would like the electronic device for reading difficult course content to highlight the words and provide definitions or links to web pages with more information on the terms in order to aid my understanding of the content when I find vocabulary that I do not understand.

The master student's response suggests that the reading device that facilitates the reading process should be able in principle to do the following things: predict and define, disambiguate and explain unfamiliar text elements. As discussed, earlier unfamiliar text elements may include new vocabulary, abbreviations, acronyms, and text phrases.

Predict and define unfamiliar text elements

The stakeholders indicated that they would like an electronic device used for reading to have an application that automatically predicts and defines perceived unfamiliar text elements. This analysis was drawn from the following responses about what the stakeholders would like an electronic device for reading difficult course content to do:

Identify perceived terminologies and get the meaning of these terms; (Postdoctoral candidate) and

Have an application that automatically defines terms or gives me alternative words (synonyms). (PhD student)

Students at Ugandan universities would like the definitions of difficult text elements to be integrated into the text being read so that they do not have to leave what they are reading. This analysis was indicated in the following undergraduate student's response:

When I find vocabulary that I do not understand, I would like the electronic device that I use for reading to get its meaning without leaving the course material am reading.

The following university lecture's response indicated that providing the definition of unfamiliar text elements in pop-up windows would be a way to keep the stakeholders in the text they are reading.

When I find a text phrase that I do not understand, I would like the electronic device I use for reading to give a short explanation for the phrase by popping up an alternative phrase because this would make reading and meaning flow while I read.

Other responses that support the lecture's idea for facilitating reading comprehension without leaving the read text were found in what the following master students said about what an electronic device used for reading should do when they find acronyms and abbreviations that they do not understand.

Display information when a cursor is held on an acronym or abbreviation that I do not understand, so that it displays its full meaning and

Provide full meaning when mouse is moved over acronym or abbreviation that I do not understand.

Disambiguate the meaning of unfamiliar text elements

The stakeholders notice that some unfamiliar text elements like new vocabulary can be ambiguous depending on the context of text where they are used. This is why they suggested that the correct meaning of the unfamiliar vocabulary needs to be given as the reading device defines the vocabulary. This suggestion was indicated in the following undergraduate student's opinion:

When I find vocabulary that I do not understand, I would like the electronic device for reading to connect to the Internet and retrieve for me the meaning of the word in relation to what am reading because this will make my reading easier.

Provide background information for unfamiliar text elements

The qualitative data analysis indicated that in some cases the stakeholders would like more explanation on unfamiliar text elements besides their definitions. They suggested that linking unfamiliar text elements to content-related background information on the Internet could provide the necessary explanation as indicated in the following opinion.

I would like the electronic device for reading difficult course content to highlight the words and provide definitions or links to web pages with more information on the terms in order to aid my understanding of the content when I find vocabulary that I do not understand. (Master student)

A postdoctoral candidate directly suggested that:

When I find vocabulary that I do not understand, I would like the electronic device for reading to mark those terms and link to their explanations.

The above responses indicate that sometime the definition of unfamiliar text elements does not provide enough information for reading comprehension as explained in section 2.4.

Illustrate unfamiliar text elements

In addition to the above-mentioned characteristics of an electronic device used for reading, a master student noted that providing visual clues for unfamiliar text elements could facilitate the process of reading comprehension. He says that:

Use of pictorial images while reading difficult course content could actually help ease the reader's understanding.

This opinion agrees with the argument in section 1.2 that graphical clues often facilitate reading comprehension and increases the probability of improving knowledge retention. A close look at the suggested functions of an electronic device that can be used for facilitating the reading process points to a document enrichment approach. Therefore, these suggestions can be used for determining requirements of a document enrichment approach that facilitates the reading process.

3.7 Document Enrichment Requirements

The literature review carried out in chapter 1 shows that university students generally have problems in understanding course materials that are domain-specific. In support

of this notion, the analysis of the stakeholders' survey indicates that students at Ugandan universities find some of their reading materials difficult to understand. For these students, difficulty in understanding results from their insufficient background and vocabulary knowledge in specific subject domains. Authors contribute to reading difficulty by providing insufficient clues to text.

Section 3.6 pointed to a document enrichment approach as an appropriate means for facilitating the reading comprehension process. All the suggestions given in that section imply that the use of information technology may have the effect of facilitating reading comprehension. For example, they suggest that the definitions and explanations of difficult technical terms that may be encountered in reading materials should be integrated into text for easy access to meaning. The students also suggest that providing pictorial clues to text would be an appropriate means to facilitate reading comprehension. Critical analysis of the students' opinions shows that the requirements for a document enrichment approach that facilitates the reading process are both general and specific.

The general requirements were identified as follows:

Integrate context relevant definitions of technical terms into text Providing context relevant definitions for the predicted term is an important step in facilitating the process of reading comprehension because it will help students to understand the meaning of such terms in context. The students will not have to search for the meaning of technical terms, which they do not understand from external sources like the Internet and dictionaries. That will help them not to deviate from the text they are reading, thereby not interrupting concentration and the reading process. Less time and effort should be needed during the reading process since the students would not have to look for the meaning of technical terms that they find difficult to understand. Another advantage is that students would build their vocabulary levels. However, the technical term definitions provided should be relevant to the text context where they occur because words are ambiguous [Krovetz and Croft, 1992], carrying different meaning in specific contexts.

Provide quick access to relevant background information

It is important to provide background information for technical terms that students find difficult to understand because sometimes the definitions of such terms are insufficient for comprehension as mentioned earlier. A similar sentiment is expressed by results from the stakeholders' survey, which shows that not all the information retrieved by the students in search for understanding is useful. For example, an undergraduate student stated that:

Difficult course content at times calls for alternatives for better understanding, which at times prove to be difficult, sometimes not.

Quick access to content-relevant and comprehensible background information

should reduce the time and effort needed for reading by providing information for supporting required inferences in case the students have insufficient background knowledge in the subject area addressed by that content.

Provide pictorial clues to text where appropriate

Pictorial clues to text are important for both facilitating the reading process and enhancing knowledge retention. It is important that students retain the knowledge they acquire because they need it for knowledge application and for integrating what they already know with new ideas. When students are able to do these things, their reading comprehension is successful.

For the above general requirements of a document enrichment approach that facilitates the reading comprehension process to be realized, the following specific requirements for the approach should be considered.

Raw Electronic Texts

Raw electronic texts are the course or research material read by students on electronic devices. These texts contain unidentified technical terms that may lead to difficulties in the reading comprehension process if students do not understand their meaning. Raw electronic text is a requirement for the document enrichment approach because it can be used in predicting technical terms, integrating term definitions into text and linking the predicted terms to external reference sources for additional background information.

Automatically Predict Technical Terms

Automatic prediction of technical terms from the raw electronic text is an important step towards integrating term definitions into text and providing background information and pictorial clues to text. Once technical terms are automatically predicted then their context relevant definitions can be searched for and integrated into text. Automatic term prediction can be done through various human language technologies, which require tailored algorithms that depend on text statistics or syntax.

Automatically Disambiguate Technical Terms Senses

In order to resolve the contextually relevant definition of a technical term, its sense needs to be disambiguated within the context where it occurs. When the relevant contextual meaning of a technical term is provided, it can be easily understood. The reading process can also be facilitated in this way because the meaning given will help the reader to seamlessly understand the text content.

Technical Terms Dictionary

The document enrichment approach requires a built-in dictionary to provide a list of technical term and their various meanings in relation to the different knowledge domains. In essence the dictionary plays the role of a term sense inventory. After a term sense or meaning has been automatically disambiguated in context, its correct definition can then be picked from the dictionary and integrated into text. The dictionary will also be used during automatic term prediction from raw electronic text to ascertain the existence of such predicted terms. For the definition of technical terms in the context of our research please see section 1.3 of this thesis.

On-line Reference Resource

An on-line reference resource is required for providing background information for technical terms that have been predicted from raw electronic text. We keep in mind that the information required should be relevant to content of the text. Such information can be provided through the technical terms because their meanings will have been resolved in context. The background information to be provided through technical terms will therefore be relevant to text content. We propose the Wikipedia encyclopedia as the on-line reference resource because it comprises almost all known knowledge domains and its structure allows easy definition of technical terms. The encyclopedia is frequently updated by a number of volunteer authors, allowing it to expand as the numbers of contemporary technical terms grow. There are also illustrations in Wikipedia that can be used for providing clues to text for enhancing reading comprehension and knowledge retention.

To address the general and specific requirements we suggest TermPedia, a data or encyclopedia-driven document enrichment approach for facilitating the reading comprehension process. TermPedia is proposed to integrate contextually relevant definitions of technical terms into text and provide background knowledge and pictorial illustrations for technical terms, through the help of Human Language Technologies (HLTs) and Wikipedia's special features. The system architecture for TermPedia will include predicting, disambiguating, defining and illustrating technical terms including providing relevant background information for them from Wikipedia. It was therefore appropriate to modify Mihalcea and Csomai [2007]'s system architecture for automatic text wikification shown in figure 3.1 to accommodate the requirements for TermPedia. Note that text wikification is a notion of document enrichment as discussed in section 1.5 of this thesis.

We choose to work with technical terms because they are a major factor in hindering the reading comprehension process and are present in most university textbooks and course materials. Another reason why we focus on technical terms is

Figure 3.1: A screen shot of the text wikification system architecture taken from Mihalcea and Csomai [2007]

because they are not part of the vocabulary all language learners acquire as they complete secondary school. We anticipate that using the TermPedia document enrichment approach to provide vocabulary and background knowledge for technical terms will facilitate reading comprehension at universities. The same approach provides illustrations for technical terms which, are expected to help students in acquiring and retaining knowledge. In the next subsection we explain how the acronym TermPedia was perceived.

Meaning and Perception of the TermPedia Acronym

The acronym TermPedia comes from two words: *term* and *pedia*. A *term* is a word or an expression that has a precise meaning in some uses or is peculiar to a science, art, profession, or subject.¹ This definition was retrieved from Merriam-Webster's Online Dictionary which is America's foremost publisher of language-related reference works. The word *pedia* is a plural form of the word *pedion*, which is a noun that refers to a crystal form having only a single face, without a symmetrical equivalent.² TermPedia therefore refers to the precise meaning of a word or an expression as used within a given text context without a symmetric equivalence of the meaning of that term in another context. Ideally the document enrichment approach, TermPedia will

¹Retrieved on Feb. 10, 2009, from, <http://www.merriam-webster.com/dictionary/term>

²Retrieved on Feb. 10, 2009, from, <http://dictionary.reference.com/browse/pedia>

be designed to interpret as accurately as possible the meaning of technical terms in the context where they occur. This will enable the tool to provide meaningful vocabulary and background knowledge for such terms. In the following chapter we will discuss TermPedia, the document enrichment approach for facilitating the reading process.

Chapter 4

Designing TermPedia's Algorithms

A man's wisdom lights up his face
and softens his stern appearance.

Ecclesiastes 8:1_b NWT

This chapter describes the steps which lead to designing the document enrichment algorithms for TermPedia. It discusses the functional design of a set of services reliant on the results of the stakeholders' survey as discussed in section 3.7 of this thesis. The chapter provides an overall description of the materials, and procedures that have been used while designing an encyclopedia-driven document enrichment approach for facilitating reading comprehension. We specifically talk about the collection, categorization, and usage of data from Wikipedia for designing the document enrichment algorithms. Detailed discussions of the techniques used in designing the document enrichment algorithms are also provided.

4.1 Overview of TermPedia's Approach

Based on our understanding of the reading comprehension process and the challenges involved in it, we found that it was appropriate to facilitate that process using an encyclopedia-driven document enrichment approach. This approach was named *TermPedia*. The approach is referred to as encyclopedia-driven because the data used to provide contextually relevant information for document enrichment derives from Wikipedia, an internet-based encyclopedia. See section 1.3 for a brief description of Wikipedia and why we use it for facilitating reading comprehension through document enrichment.

Encyclopedias are large sources of information, which can be used to enhance the learning process [Parker and Chao, 2007] but, to make good use of this information a user has to switch between interfaces, whether on the world wide web (WWW) or while using printed copies of these reference materials. Finding information from encyclopedias like Wikipedia can be time consuming and expensive [Smith, 1996]

Figure 4.1: TermPedia Document Enrichment System Architecture; Modified from Manning et al. [2008]’s Text Wikification System Architecture.

with regard to the effort that must be spent in searching for the information. We hope to minimize the time, effort, and distance between encyclopedic knowledge and documents by integrating information from Wikipedia into electronic documents.

Our approach to document enrichment builds on work done by Mihalcea and Csomai [2007], who linked education materials to Wikipedia knowledge. The “build” is that, in addition to linking technical terms directly to Wikipedia, our approach of document enrichment integrates their definitions and illustrations into the documents in which they occur. To achieve our idea of document enrichment we modified the text wikification system architecture by Manning et al. [2008] to incorporate the integration of text into electronic course materials read by university students. Compare figures 3.1 and 4.1. The importance of this modification was to keep university students in the documents they read without having to search for the meaning of difficult terms from external sources. It was expected that this would reduce the time and effort required for reading and knowledge acquisition. Knowledge retention would also be facilitated since the approach provides graphical illustrations of the technical terms where possible.

The encyclopedia-driven approach to document enrichment largely depends on Internet technology, with an implementation of information extraction, technical term prediction, term sense disambiguation and hypertext generation as the core techniques. Information extraction is used to create a predefined dictionary of technical terms and other pieces of information relevant for document enrichment. Term prediction highlights terms from the predefined dictionary in course materials used at universities. Term sense disambiguation resolves the meaning of the predicted terms in context so that relevant definitions, graphical illustrations and background information are provided for them. Hypertext generation is used for integrating contextually relevant technical terms definitions, illustrations and background information for the predicted technical terms into the course materials used by university students. Hypertext generation is a basic technology for creating informative and interactive text through the use of markup languages. We use hypertext mark-up language (HTML) for generating hypertext because it is the main markup language for displaying information on the WWW. Briefly, hypertext markup involves annotating pieces of text in a given document so that the text is syntactically distinguishable from other text in that document. With HTML, it is possible to link the annotated piece of text to information within or outside a given document [Domingue et al., 2001]. We extended the traditional behaviour of hypertext for document enrichment by enabling annotated pieces of text to display information in tool tips. This goes beyond traditional hypertext that leads to a new web page or a new section of a web page. We expect that this will provide a more user-friendly experience than the traditional hypertext links. See figure 5.4 on page 99 for an extended presentation of the behaviour of traditional hypertext.

Sarawagi [2008] refers to “traditional” information extraction as the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources. However, in this case we need to extract information from Wikipedia, a semi-structured Internet-based data source. For this reason, Internet information extraction was identified as the appropriate technology to carry out the task.

In contrast to “traditional” information extraction which roots in natural language processing (NLP) techniques such as linguistic analysis, Internet information extraction (relies) on syntactic structure identification marked by HTML tags.
[Chang et al., 2001]

We borrowed the idea of syntactic structures marked by HTML tags and applied it to syntactic structures identified by wiki markup language. A wiki is a website that allows people to create and collaboratively edit web pages via a web browser. Wiki markup is the syntax and keywords used by the MediaWiki software, a free web-

based wiki application used to run all the Wikimedia Foundation projects, which include but are not limited to Wikipedia, Wiktionary, and Wikinews. An example of wiki markup syntax can be given as `==Anatomy==`, which splits wiki articles into sections using various level headings. The two-level heading depicted here by the double equal signs to the left and right of the keyword *Anatomy* is the highest level of heading used by wiki article authors and editors. This example indicates that there is a section named *Anatomy* on a certain wiki page. Like this heading example, there are syntactic structure wiki markups for links and URLs, mathematical formulae, Greek letters and many others. Since the wiki data structure does not conform to the formal structure of data models associated with relational databases or other forms of data tables, we refer to it as semi-structured. Nonetheless the wiki data structure contains markups to separate semantic elements and enforce hierarchies of records and fields within the data. The Internet information extraction technique tailored to use wiki markup was therefore employed for extracting and analyzing data for the task of document enrichment, from wiki pages that form the English Wikipedia.

In order to adequately enrich documents, we use both the text and images from Wikipedia. Wikipedia text provides definitions of technical terms and other relevant information on the technical terms for extended reading, while the images provide illustrations for the technical terms. We believe that if a reader understands the meaning of a technical term and sees its visual illustration, then he or she is able to comprehend the text and also remember it for a long time.

4.2 Document Enrichment Algorithms Design

As already mentioned, document enrichment involves an integration of various human language techniques such as technical term spotting, resolving the meaning of technical terms in context, technical term illustration, and detecting definition passages. The semantic web technology of automatic link generation is also involved in document enrichment for highlighting spotted technical terms and linking them to their background knowledge. The following tasks needed to be carried out during the process of designing TermPedia's document enrichment algorithms.

- (i) Creating a dictionary of technical terms from different disciplines.
- (ii) Identifying which phrases in electronic documents represent technical terms.
- (iii) Figuring out which Wikipedia pages provide relevant information for the identified technical terms.
- (iv) Enriching documents with information relevant to the identified technical terms.

Algorithms based on human language technologies had to be designed for carrying out tasks (i) through (iv) above. (See section 4.6) The algorithms designed for tasks (i), and (ii), used wikilinks for generating a dictionary of technical terms from a data corpus containing a dump of the English Wikipedia. Wikilinks are hypertexts generated by wiki markup language; they contain information about which Wikipedia pages provide information for the technical terms (concept strings).

Wikilinks therefore provide access to two pieces of important information; (a) technical terms referred to by the wikilink labels, and (b) which Wikipedia article provides information for the technical term referred to by the wikilink targets. For example, in a wikilink where the label and target are separated by a pipe (|) symbol such as shown in example 4.2.1, *temporal* is the wikilink label and text which is displayed on a Wikipedia page that links to another Wikipedia page with the title *Temple (anatomy)*. In this case *temporal* is the technical term and *Temple (anatomy)* is the target, or Wikipedia page that provides information for the term. We use wikilink labels to represent technical terms because they occur within the text body of Wikipedia pages and are therefore more likely to provide genuine technical terms, which may occur in other documents compared to the wikilink targets which provide subject concepts. Looking at example 4.2.1 it is evident that *temporal* is a better technical term compared to *Temple (anatomy)*, which is clearly a subject concept. Wikilink labels also solve the problem of spelling variations, word stemming, and terms made up of multiple nouns, noun phrases and/or word strings to a large extent.

Example 4.2.1: Wikilink

[[Temple (anatomy)|temporal]] ““Temple”” indicates the side of the [[head (anatomy)|head]] behind the [[human eye|eye]]s. The bone beneath is the [[temporal bone]] as well as part of the [[sphenoid bone]]...

Example 4.2.2 (a) and (b) show United States (U.S.) spelling variations of the term *esophagus* with one form having the letter (o) at its beginning. These spelling variations are used interchangeably in written English. Furthermore the adjectival spellings for the same term varies with the letters (*ea*) in place of (*us*) at the end of the normal variations. Notice that all variant spellings of the term *esophagus* exist in Wikipedia. Thus, for a document that contains any of the spelling variants, we are able to provide a link to the relevant Wikipedia page.

Example 4.2.2: Spelling variation

- (a) esophagus
- (b) oesophagus
- (c) esophageal
- (d) oesophageal

Often technical terms that appear in documents have many morphological variants as shown in example 4.2.3. In most cases the morphological variants of technical terms have the same semantic interpretations that can be important notions for understanding documents. Word stemming is normally necessary to reduce a technical term to its root or base form so as to allow linking to all the morphological variations of a word. Fortunately, the lists of terms generated from wikilink labels contain many morphological variations of a technical term, as we see in this example, so that the need for stemming does not arise. By using many morphological variations of technical terms, we create a large-sized list, which is conveniently stored in a database to save processing time.

Example 4.2.3: Avoiding word stemming

coupling; coupled; coupler; couples; couplet; couple

In some cases, technical terms are expressed by phrases instead of single words, for example compound nouns. Wikilink labels are a good representation of such cases, as shown in example 4.2.4 (a) where both *medicine* and *cabinet* are two different nouns that are combined to form a term. In example 4.2.4 (b) we see that wikilink labels may also consist of noun phrases. In this particular example, the word *arteries* is the noun represented by the noun phrase *the great arteries*. Moreover, if a technical term consists of a long string of words like a journal name as shown in example 4.2.4 (c), wikilink labels may nonetheless represent them accurately.

Example 4.2.4: Multiple nouns, noun phrases, and word strings

- (a) Medicine cabinet
- (b) Transposition of the great arteries
- (c) Journal of the history of medicine and allied science

Generally, it is quite hard to determine which word sequences in an electronic document denote a single concept [Miller and Schwartz, 1994] or term. For example, it is hard to determine which word sequences represent a multi-word unit while using frequencies or other statistical models to try and predict them from text. The advantage of using wikilink labels is that they contain multi-word units that represent various

terms as shown in example 4.2.4. In the next section, we describe the data from the English dump of Wikipedia that was used in designing our approach of document enrichment for facilitating reading comprehension.

4.3 Data Used in Designing the Algorithms

An XML dump of raw data containing Wikimedia markup syntax, images, tables, references, and lists was downloaded from the Wikimedia website¹ on the Internet. The dump was then processed using WikipediaMiner², a toolkit for extracting information from the markup semantics of Wikimedia [Milne and Witten, 2009]. Like many other tools that exist today for processing Wikipedia data based on its rich structure, WikipediaMiner provided information about terms and concepts in Wikipedia and how they are related to each other. It also provides a simple object-oriented access to Wikipedia’s structure and content. After processing the XML dump of Wikipedia using WikipediaMiner, the important Comma Separated Values (CSV) files shown in table 4.1 were generated as were other files containing Wikipedia terms, concepts, and concept relations.

Line No.	File name	Content Sample
1.	anchor.csv	“autism”,25,928
2.	anchor.summary.csv	“autism”,“25:928;29113700:1;1871748:1;929970:1”
3.	page.csv	25,“Autism”,1
4.	pagelink.in.csv	25,“3433:4805:5237:5664:...”
5.	infobox.csv	25, Autism, Disease, Name, Autism

Table 4.1: Important files generated by WikipediaMiner

The *anchor.csv* file contained information about wikilink labels, i.e. what strings of text are used as wikilink labels, and how many times the wikilink label has been used to refer to a certain Wikipedia page. Considering the example in table 4.1, we see that the wikilink label *autism* has been used nine hundred and twenty-eight times (928) to create a link to the Wikipedia page with *id* 25. Summaries about wikilink labels were readily available in the *anchor.summary.csv* file that provides information about all the Wikipedia pages to which a particular wikilink label makes a link, including the frequencies of how often these links have been made. For example, we see in the second line of table 4.1 that the wikilink label *autism* was used nine hundred and twenty eight times (928) to create a link to the Wikipedia page with *id* 25, and

¹See <http://dumps.wikimedia.org/>. Cited on 25 September, 2012.

²Available at <http://wikipedia-miner.cms.waikato.ac.nz/>. Cited on 11 October, 2012.

one (1) time to create a link to the Wikipedia pages with ids *29113700*, *1871748* and *929970*.

The frequencies of how often a wikilink label is used to refer to a Wikipedia page are important since they are used for disambiguating the target of ambiguous technical terms. Ambiguous wikilink labels (technical terms) are used to refer to more than one Wikipedia page while unambiguous wikilink labels are used to refer to a single Wikipedia page. The ambiguity phenomenon is inherent in natural language since terms have various meanings depending on the context in which they occur within a document. For example, the term *temporal* discussed in the example 4.2.1 above may refer to the concepts of time or the human anatomy as shown in table 4.2. Depending on the context in which the term *temporal* occurs, we need to find the Wikipedia page that would give the correct definition for the term in that context. This is why technical term sense disambiguation is needed i.e., to resolve the meaning of the technical term in context so that we can provide relevant information for it during document enrichment.

The wikilink label (term) <i>temporal</i> can refer to the following concepts:	
Concepts	Title of Wikipedia Page (Target)
of or relating to time:	
Temporality in philosophy	Temporality
The Lords Temporal, secular members of the House of Lords	Lords_Temporal
Temporal (video game), a freeware platform and puzzle game	Temporal_(video_game)
of or relating to the temple (anatomy):	
Temporal bone, a part of the skull	Temporal_bone
Temporal lobe, a part of the brain	Temporal_lobe
Temporal scales in reptiles	Temporal_scales
in logic, it can refer to:	
Temporal logic	Temporal_logic

Table 4.2: Possible ambiguous concepts of the term *temporal*

From the *page.csv* file we can obtain information about the Wikipedia page ids and titles. We can see in table 4.1 that the Wikipedia page with id *25* is titled *Autism*. This information was used to resolve the page titles of the ids from the *anchor.csv* file, giving a name to the Wikipedia page which the wikilink labels were used to make a link to. Therefore, since the wikilink label *autism* makes a link to a Wikipedia page with id *25*, we can say that the wikilink label links to the Wikipedia page titled Autism.

The file *pagelink.in.csv* provides information about which Wikipedia pages link to a particular Wikipedia page. Looking again in table 4.1, this time at line 4, we see that the Wikipedia page with id 25 receives links (in-links) from more than four other Wikipedia pages. The in-link information was important for technical term target disambiguation just like the frequency information mentioned above.

We extended the WikipediaMiner software to extract information from Wikipedia infoboxes and store them in five-tuples in the *infobox.csv* file. The five-tuples include page-id, page-title, template-name, attribute-name, and attribute-value. Line 5 in table 4.1 shows a sample record of this file where 25 is the page-id, *Autism* is the page-title, *disease* is the template-name, *Name* is the attribute-name, and *Autism* is the attribute value. The template-name provides classification information for subjects in Wikipedia that are represented by the page titles. This made it possible to classify Wikipedia subjects according to the template-name of Wikipedia infoboxes.

Example 4.3.1: Content of plain text files

```
<doc id="515048003" url="http://en.wikipedia.org/wiki/Temple (anatomy)">
Temple (anatomy).
Temple indicates the side of the head behind the eyes. The bone
beneath is the temporal bone as well as part of the sphenoid
bone...
</doc>
```

In addition to the CSV files, plain text was extracted from the Wikipedia XML dump using *WikiExtractor.py*. This python script extracts and cleans text from a Wikipedia database dump [Medialab, 2012], discarding any other information or annotations present in Wikipedia pages, such as images, tables, references and lists. Only the text body of the Wikilink article was reserved, and this was the text between the `<text></text>` Wikimedia markup tags. The generated text files were stored in 65 bz2-compressed files of similar sizes. Each compressed file contained various Wikipedia articles with document content format shown in example 4.3.1. From these plain text files, it was possible to extract data for one of the disambiguation algorithms that uses a Lesk-like definition overlap for sense disambiguation. See section 2.3 for a brief explanation of the Lesk algorithm and the section below titled *Disambiguation by Lesk-like definition overlap* for the application of a Lesk-like algorithm.

Much as it is easy to access information from Wikipedia because of its rich data structure of markup semantics, the data retrieved is noisy and may not always represent the desired information. However, because of its large coverage it is often possible to locate desired pieces of information. The summary files generated by

	Types	Tokens	Averages
All wikilink labels	10,672,018	80,938,424	7.58
Unambiguous wikilink labels	10,093,442	38,637,384	3.83
Ambiguous wikilink labels	578,576	42,301,040	73.11
Page titles	8,605,586	8,605,586	1.00
Incoming links	8,039,487	70,060,185	8.71
Infobox names	18,040	1,242,427	68.87

Table 4.3: Summary statistics for the English Wikipedia dump of 2010

WikipediaMiner provided an accessible data source for wikilink labels and their targets among other necessary information for our document enrichment experiments. [See table 4.3] Among the summary information were 80,938,424 total occurrences of wikilink labels with 10,672,018 unique occurrences of those labels. This shows that on average there were approximately 7 occurrences of each wikilink label in the entire Wikipedia dump. Of the 10,672,018 unique wikilink labels 578,576 are ambiguous with a frequency of 42,301,040 (see Tokens column in Table 4.3). This indicates that more than half (52.26%) of the total occurrences of these wikilink labels are ambiguous and they occur millions of times. Therefore, there is a need to disambiguate their meaning in the text context where they may occur. The dump also contained 8,605,586 page titles. These page titles provided targets for the wikilink labels as described earlier. In total there were 70,060,185 occurrences of in-coming links to 8,605,586 pages, showing that each page had an average of 8 other pages in the Wikipedia dump linking to it. The high frequency of in-links occurrences made it possible for them to be used for the task of word sense disambiguation.

Collecting and Categorizing Data for Designing TermPedia’s Document Enrichment Algorithms

To address the reading comprehension problems discussed in sections 1.3 and 3.4 using a document enrichment approach, we collected data from the English Wikipedia dump described above. The data was collected for predicting and disambiguating the meaning of technical terms with respect to the context in which the terms occur. Data for predicting technical terms included a list of wikilink labels (terms) and targets. The targets are an indication of which Wikipedia pages provide relevant information for the terms. Data for disambiguating the sense of technical terms in context formed a dictionary because it contained a list of technical terms and various definitions for those particular terms. Information about the in-coming links

from other Wikipedia pages to a particular Wikipedia page was also collected for the disambiguation of technical terms senses in context. All these data were collected from the CSV files generated by WikipediaMiner toolkit as discussed in section 4.3. The collected data were stored in databases because this would facilitate design of the document enrichment algorithms by reducing data access costs and processing time. The following subsections explain how the collected data were stored and used for document enrichment.

Generating the list of technical terms

Three databases were created to contain lists of technical terms represented by wikilink labels. The databases contained four tuples consisting of:

- (i) the wikilink label,
- (ii) the ids of the Wikipedia pages which the wikilink label was used to refer to, (page-ids),
- (iii) the titles of the Wikipedia pages which the wikilink label was used to refer to, (page-titles),
- (iv) and the frequencies with which these wikilink labels were used to refer to the Wikipedia pages, (frequency).

This information was stored into three databases because wikilink labels are either unambiguous or ambiguous. One of the three databases that were created stored all the wikilink labels regardless of whether they are ambiguous or not. This information was necessary for designing the baseline document enrichment algorithm discussed in the section titled *Blindfold Term Prediction*. The other two databases contained the ambiguous and unambiguous wikilink labels respectively. Consider an example of the wikilink label *esophageal arteries*, tuples in the technical terms database would have the values shown in table 4.4 for this particular wikilink label.

Tuples	Values
wikilink-label	esophageal arteries
page-ids	8694323; 4484965
page-titles	Esophageal arteries; Esophageal branches of thoracic part of aorta
frequency	2; 1

Table 4.4: Tuples and values from generated database for wikilink labels

Tuples	Values
wikilink-label	cold agglutinin disease
page-ids	4517757
page-titles	Cold agglutinin disease
frequency	7

Table 4.5: Tuples and values from database of unambiguous wikilink labels

Table 4.4 is a good indication of the records stored in the ambiguous wikilink label databases and table 4.5 shows an example of a record contained in the database of unambiguous wikilink labels. Remember that we use wikilink labels to represent technical terms, so that these databases constituted the lists of terms and the information about which Wikipedia pages provide relevant information for them. Table 4.4 shows that the term *esophageal arteries* is linked to two Wikipedia pages one of which has the title *Esophageal arteries*. These databases were generated from the *anchor.csv* and *page.csv* files. [See table 4.1] However there were two criteria for selecting terms for the list of technical terms.

Collecting Data for Term Sense Disambiguation

For term sense disambiguation, we needed to create two databases to contain a list of technical terms and their definitions (a dictionary), and a list of all Wikipedia pages including the incoming links (in-links) to these pages, respectively.

Creating the technical terms dictionary

Definitions of terms that occur in the list of all technical terms from the Wikipedia dump are extracted from the first paragraphs of Wikipedia pages in the plain text documents. A close look at the plain text data reveals that the first paragraphs of Wikipedia pages almost always contain definition statements about the page title. To illustrate, the first paragraph of the Wikipedia page titled *Temple (anatomy)* provides a definition statement of the term *temple*, which also happens to be the definition of the term *temporal* in the context of anatomy. See example 4.3.2, which also shows that data in the technical terms dictionary database contains two tuples namely a wikilink label (technical term) and the first paragraph of the Wikipedia page(s) to which the wikilink label refers (technical term definition). In this case *temporal* is the technical term and *Temple indicates the side of the head behind the eyes...* is the definition of the technical term.

Example 4.3.2: First paragraph of a plain text Wikipedia page titled Temple

```
<doc id="515048003" url="http://en.wikipedia.org/wiki/Temple (anatomy)">
Temple (anatomy).
Temple indicates the side of the head behind the eyes. The bone
beneath is the temporal bone as well as part of the sphenoid bone.
...
</doc>
```

Wikipedia page in-links collection

From the *pageinlink.csv* and *page.csv* files in the English Wikipedia dump of 2010, it was possible to create a database containing information about in-links to particular Wikipedia pages. Information about the ids of all Wikipedia pages that link to a particular Wikipedia page is provided by the *pageinlink.csv* file. In example 4.3.3 we see that pages with the ids 3433, 4805, 5237, 5664, ... link to the Wikipedia page with id 25. The *page.csv* file provides these page ids with titles where we find that Wikipedia pages with titles *Blindness*, *Behavior*, *Catatonia*, *Consciousness*, ... link to the Wikipedia page with title *Autism*. However, in the page in-links database we only stored the page ids and resolved them with page titles after deciding which Wikipedia page provides the best information for a technical term in question. [See section 4.5 (iii)]

Example 4.3.3: Wikipedia page in-links

```
25,"3433:4805:5237:5664:..."
```

4.4 Technical Terms List Generation Criteria

The criteria for selecting terms for the list of technical terms depended on a stop word list and the frequency with which a wikilink label is used to refer to Wikipedia pages. Stop words are defined as words that occur frequently in documents but carry negligible amount of information [Lo et al., 2005]. These words therefore have very low information value when it comes to document enrichment and are better ignored while searching for difficult sections of a document that may need further explanations.

A stop word list is necessary because words that form this list are sometimes used as wikilink labels and it would be irrelevant to include them in the list of terms generated from Wikipedia. For example, in table 4.6 we see that the word *in* is used as a wikilink label that refers to a Wikipedia page with id 14775 and title *Inch*.

Stop word	Wikipedia page referred to	
	id	title
in	14775	Inch
is	3409	Being
of	658770	Of

Table 4.6: Examples of stop words that are used as wikilink labels

Generating the stop words list

The stop word list was generated for the collection of clean text documents from the Wikipedia dump of October 11, 2010. The list was generated from one million five hundred thousand (1,500,000) clean text articles, where each article represented a document in the Wikipedia clean text corpus. The benefit of using a corpus specific stop word list is that it excludes less informative terms for a given data collection, thus avoiding the problem of providing irrelevant information for documents in that data collection.

$$idf_{xNorm} = \log \left(\frac{(NDoc - D_x) + K}{D_x + K} \right) \quad (4.1)$$

where $NDoc$ is the total number of documents in the collection, D_x is the number of documents containing term x , and K is a constant of value 0.5 added to moderate extreme values.

While comparing approaches to stop word list generation Lo et al. [2005] reported that the inverse document frequency approach is one of the best methods for the task. For this reason, a normalised inverse document frequency (idf_{xNorm}) approach based on *Zipf's law* [Zipf, 1949] was used to produce a set of stop words for the 1,500,000 documents in the clean text Wikipedia dump. Equation 4.1 presented in the paper by Lo et al. [2005] was used as the formula for obtaining a normalised IDF for words in the clean text corpus. Each of the documents from Wikipedia was tokenized, case unified, and indexed without removing any tokens so as to obtain the best stop word list possible for the collection. After obtaining the normalised IDF for each word that occurs in the Wikipedia dump, a list of word frequencies versus words was generated. The generated list was sorted by inverse term frequencies in ascending order so that the most frequent term with the lowest normalised IDF was at the top of the list. The terms were then ranked according to their inverse frequency values whereby the term with the lowest frequency got a rank of 1.

Figure 4.2: Zipf's plot for ranked inverse word frequencies in Wikipedia

To choose a threshold frequency for a stop word list, differences between two consecutive inverse document frequencies namely $F(x+1)$ and $F(x)$ are investigated to find a very high value. Then all the terms with an inverse document frequency less than or equal to $F(x+1)$ would then be used in the stop words list.

In our case, a graph of term inverse document frequency versus ranks that obeys Zipf's law was drawn to help in choosing the threshold word frequency. As shown in figure 4.2 it was difficult to locate the threshold frequency because the consecutive inverse frequencies are very close in value and therefore their differences did not give a distinctive very high value. For this reason, we created a stop words list of 634 words by extracting words from an existing standard English stop words list that occur in the list of inverse document frequencies versus words from the Wikipedia dump. The standard English stop words list which is available at <http://www.ranks.nl/resources/stopwords.html> contains 667 words and was created for a web master tool. This standard stop words list was used because it gives a large coverage of frequent English words that have low information value. The 634 words that formed a stop words list for the Wikipedia dump corpus were disregarded while identifying technical terms that occur in an electronic document. These words were also disregarded while performing term sense disambiguation as explained in

section 4.5. A sample of the generated stop words list can be seen in table 4.7

Position	Normalised IDF	Stop Word
1	-1.4624	the
2	-1.1755	in
.	.	.
.	.	.
316	1.5845	look
317	1.5850	is
.	.	.
.	.	.
633	5.6457	et-al
634	5.7918	through

Table 4.7: Sample stop word list for Wikipedia documents

N-grams rank for technical term likelihood

The second criterion used in selecting terms for the technical terms list was the ngrams rank for technical term likelihood. This criterion is motivated by a *keyphraseness* method, which was used by Mihalcea and Csomai [2007]. In this approach a list of all wikilink labels was created to represent terms that occur in Wikipedia. Uni-grams and bi-grams in this list were ranked and selected by their technical term likelihood probability. However, terms that occur as tri-grams through seven-grams were directly included in the terms list with a likelihood probability of one (1), because their instances represented genuine technical terms.

The technical term likelihood rank for a uni-gram or bi-gram term is given by the probability that it occurs often in Wikipedia as a wikilink label and not just as any part of the text. Precisely if a term is often used as a wikilink label among its total number of occurrences in Wikipedia documents, then it is most likely that it will be used in any document as a technical term (tt). Therefore the probability (p) that a term (x) is selected as a technical term in any document is given by the total number of Wikipedia documents where the term was used as a wikilink label ($count(d_{tt,x})$) divided by the total number of documents where the term appeared in the whole of Wikipedia ($count(d_x)$).

$$p(tt \mid x) \approx \frac{count(d_{tt,x})}{count(d_x)} \quad (4.2)$$

The counts were generated from all the articles in the 2010 Wikipedia dump and wikilink label uni-grams and bi-grams with high technical term likelihood rank values ($p(tt|x)$) were considered while identifying technical terms in an existing electronic document. Note that the likelihood ranks are also used in disambiguating the target of ambiguous.

4.5 Technical Term Sense Disambiguation Criteria

In the event that a technical term refers to two or more Wikipedia pages, it is important to decide which page would provide relevant information for that technical term in a particular context. The process of deciding which Wikipedia page best provides relevant information for a technical term in context is what we refer to as technical term sense disambiguation. There are three criteria for disambiguating the senses of ambiguous technical terms and their Wikipedia targets. These criteria depend on the (i) n-grams technical term likelihood rank, (ii) the technical term context overlap counts, and (iii) the Wikipedia page in-links overlap counts.

(i) Disambiguation by technical term target likelihood ranks

The Wikipedia page that is most often used as a target for a particular technical term (wikilink) is selected as the target for that term. The first paragraph in the selected Wikipedia page is then given as the definition of the term in question, which is also its most frequent sense in the Wikipedia encyclopedia. For example, from table 4.1 on page 75 we see that the technical term *autism* would be linked to the Wikipedia page titled *Autism* with *id* (25) because it has the highest term target likelihood rank compared to the other two targets with *ids* (29113700) and (929970) respectively.

(ii) Disambiguation by Lesk-like definition overlap

The approach of technical term sense disambiguation by knowledge-based Lesklike definition overlap resolves term senses and targets by overlapping the words in the paragraph where the term occurs with the words in the term's various context specific definitions. If there is a high overlap of words between the two sets of text, then we assume that their contexts are similar. An intersection computation algorithm was used to disambiguate terms senses in context. Given the two paragraphs, we determined which words are common to both paragraphs (intersection). Each paragraph is tokenized and sorted to

have unique items. Stop words were also removed from the tokenized paragraphs to form a pair of word lists P_1 , and P_2 that contain only the important words within the paragraphs. P_1 is the word list derived from the paragraph where the term occurred in any electronic document and P_2 is a list of words derived from the paragraph containing the term definition. To find the similarity score between the two paragraphs ($SS_{P_1 \text{ and } P_2}$), a count of the words in their intersection ($|P_1 \cap P_2|$) is determined. The definition paragraph (P_2) that gives the highest similarity score is chosen as the sense of the technical term in question and the Wikipedia page title that provides this definition, is chosen as the term's target.

$$SS_{P_1 \text{ and } P_2 \rightarrow n} = \max_{P_1} |P_1 \cap P_2 \rightarrow n| \quad (4.3)$$

Where (n) is the total number of paragraphs containing definition information of the ambiguous technical term in various text contexts.

(iii) Disambiguation by Wikipedia page in-links overlap counts

In this approach, if a technical term links to two or more Wikipedia pages, an overlap count of in-links to the Wikipedia pages are made. Every once in a while, when predicting technical terms, we come across a technical term that is unambiguous. The targets for ambiguous technical terms are taken as content that relate closely to the document in which the technical term occurs. Therefore, the in-links to these target pages are used in disambiguating targets of ambiguous terms. The in-links to a target of an ambiguous term are overlapped with in-links to all the targets of the unambiguous terms, and the ambiguous target with the highest link overlap is selected as the target of the ambiguous term in question.

4.6 TermPedia's Document Enrichment Algorithms

The process of document enrichment in this research involved two distinct steps, namely technical term prediction and disambiguation, and technical term definition and illustrations. The earlier step is responsible for identifying difficult terms in electronic documents and for resolving which Wikipedia pages provide the best information for these terms, while the later step is responsible for providing definitions for the identified terms.

Technical term prediction and disambiguation

One hundred (100) random electronic documents from the plain text files of the Wikipedia dump of 2010 were used for designing five technical term prediction and disambiguation algorithms. The five algorithms designed were: (i) blindfold term prediction (BTP); (ii) longest-string-based term prediction (LTP); (iii) sense-based term prediction (STP), (iv) frequency-based term prediction (FTP), and (v) maximum incoming links overlap term prediction (MILOP). Each of the algorithms attempts to predict technical terms in electronic documents and provide relevant Wikipedia page targets to the predicted terms. Algorithms (i) and (ii) did not consider technical term sense disambiguation while algorithms (iii) through (v) try to disambiguate the senses of technical terms in the context in which they occur.

(i) *Blindfold Term Prediction (BTP)*

The BTP algorithm is the baseline algorithm that uses a string look-up technique which marks all possible strings in an electronic document that match a term from an existing terms list as a technical term. The Last in First Out (LIFO) principle borrowed from the stack data structure [Erdem, 1996] was used for each sentence within the electronic document. We assume that a sentence contains a stack of technical terms and the first term to be recognized in that sentence is then predicted. Multiple terms can be predicted per sentence as long as the string-look up algorithm returns a positive value for the terms in the existing list. This method is expensive because each line of running text is matched against all the technical terms in a terms list.

Once an instance of a term from the existing list of terms is seen in an electronic document, the term is only marked as a technical term if it does not occur in the stop words list. Therefore, the conditions required by BTP algorithm to predict a technical term are that it must occur in the existing list of terms and not in the stop words list.

Example 4.6.1: Sentence in an electronic document

This is the temporal bone.

In example 4.6.1 depending on which term occurs first in the existing list of technical terms, it is possible to have any of the following terms predicted for this sentence: *temporal*, *bone*, or *temporal bone*. If the terms predicted are unambiguous then they have only one target in the terms target database created in section 4.3. Otherwise if the term links to more than one target, then a random target is selected for that particular term.

(ii) *Longest-String-Based Term Prediction (LTP)*

The difference between BTP and LTP is that in LTP the existing terms list was sorted in such a way that the terms having the longest length of characters were stored at a queue head. LTP was thus able to predict the longest string in running text that matched a term from the existing terms list as a technical term. A first-come-first-predicted idea exists for the terms in the list of terms. This idea is biased towards the longest string in a sentence that can be predicted as a technical. This criterion was adopted to avoid splitting up terms that are made of compound nouns or noun phrases. For instance, for the sentence shown in example 4.6.1 if the term *temporal bone* is not first extracted, it may be split into *temporal* and *bone*. We expected the LTP algorithm to predict the term *temporal bone* from text in example 4.6.1. Like BTP, LTP did not take into consideration the ambiguity of term senses in relation to context but disregarded the prediction of stop words as technical terms. The term's target page is selected in the same way as done in the BTP algorithm.

(iii) *Sense-Based Term Prediction (STP)*

The STP algorithm uses the term context overlap counts approach described in section 4.5 (ii), to resolve term sense similarities before they are predicted as technical terms. This algorithm borrowed the idea of term sense disambiguation based on a knowledge-based Lesk-like definition overlap as explained by [Mihalcea and Csomai, 2007]. STP takes into consideration the contextual meaning of a term before it is predicted as a technical term. Here terms are only predicted as technical terms if their context is similar to the context of a term definition, regardless of whether the term is ambiguous or not. A threshold of 2 is given for the similarity score value. Therefore, for a term to be predicted as a technical term, its similarity score value must satisfy equation 4.4.

$$SS_{P_1 \text{ and } P_2} = \arg \max_{P_1} |P_1 \cap P_2|, \quad \text{if } |P_1 \cap P_2| \geq 2 \quad (4.4)$$

Where SS is the similar score, P_1 are a list of words derived from the paragraph where the term occurs in any electronic document P_2 are the list of words derived from the paragraph containing a term's definition. If P_2 gives the maximum number of overlapping words with P_1 compared to other paragraphs that define the term in question.

If a term is ambiguous and its similarity score value satisfies equation 4.4, then the Wikipedia page which provides the highest similarity score value is selected

as a target for the predicted term. We expected STP to have the best recall since it is designed to predict technical terms after disambiguating their senses in relation to context. Moreover, we also expected that a technical term will always be linked to a contextually relevant Wikipedia page thereby improving the value of document enrichment.

(iv) *Frequency-Based Term Prediction (FTP)*

The n-grams term likelihood rank criterion discussed in section 4.4 is used for predicting technical terms by the FTP algorithm. Here a list of term likelihood ranks ($p(tt|x)$) versus terms is generated and sorted in descending order. Therefore, a term with the highest value of ($p(tt|x)$) will be at the top of the list. A term is only predicted as a technical term if it has a high value of ($p(tt|x)$). This algorithm also makes filtering stop words superfluous since they will have very low values of ($p(tt|x)$), therefore they will not be suggested by the algorithm as potential term candidates.

If a term is ambiguous, indicating that it links to more than one Wikipedia page, then the Wikipedia page to which the term links to most frequently is chosen as a target for the technical term. The FTP algorithm therefore performs technical term sense disambiguation by picking the most frequent sense for a predicted term.

Given a list of tokens in an electronic document, at most 10% of the tokens are annotated as technical terms, since Mihalcea and Csomai [2007] noted that on average 6% of the tokens on a Wikipedia page are actually linked to another page. The 10% threshold is used to prevent clogging documents with annotations which may obstruct document legibility. We expected the FTP algorithm to have a high precision in predicting technical terms and their targets since it selects the most frequent in each case.

(v) *Maximum Incoming Links Overlap Counts (MILOP)*

There are three major steps in the incoming links overlap count algorithm which include categorizing terms into unambiguous and ambiguous subsets, finding Wikipedia targets for the terms and incoming links to these targets, and counting the incoming link overlaps between the unambiguous targets and ambiguous targets. The ambiguous target with the most incoming links overlap count is then selected as a target for the ambiguous term. Given an unambiguous term x and an ambiguous term the score (β) for the best target for y was calculated using the formula in equation 4.5. Where X is the collection of incoming links to the target of x and Y are the collection of incoming links to targets of y . As already explained, an unambiguous term is linked to only one Wikipedia page and an ambiguous term can be linked to two or more Wikipedia pages.

$$\beta(x, y) = |X \cap Y| \quad (4.5)$$

In the event that a predicted term is a stop word, it is disregarded by MILOP. Like FTP, we expect this algorithm to provide relevant targets for the predicted terms since it carries out terms meaning disambiguation by counting in-links overlaps. That should result in high recall values.

Conclusion

After technical terms have been predicted and their senses disambiguated in context of the electronic documents in which they occur, the predicted terms are then used to generate hypertexts for document enrichment. The hypertexts integrate the definitions and illustrations of the predicted terms into the document where they occur and create links to a contextually relevant Wikipedia page for additional background information on the predicted terms. See section 4.2 for an example of how documents are enriched by automatic hypertext generation.

In this chapter we indicate that it is possible to use technical terms and available encyclopedic knowledge to automatically enrich documents through information extraction and hypertext generation. In section (4.6) we discussed five document enrichment algorithms, which provide relevant information for technical terms that occur in electronic documents. These algorithms resulted into document enrichment methods of integrated techniques such as technical term prediction, technical term sense disambiguation, automatic hypertext generation, automatic term definition, and technical term illustration. The document enrichment algorithms will be tested in chapter 6 of this thesis for their effectiveness in predicting technical terms and providing contextually relevant definitions for the predicted terms. It is important that the document enrichment algorithms obtain high recall scores during the testing process. If recall scores are high, it will indicate that the algorithms predict most of the terms that the students may find difficult to understand in the contents of their course documents. Contextually relevant definition can be provided if the term sense disambiguation algorithms are effective in resolving the meaning of technical terms in specific document contexts. Therefore high recall scores attained by the document enrichment algorithms in linking predicted terms to contextually relevant Wikipedia pages will be crucial for user satisfaction.

Chapter 5

TermPedia: The DE Approach

Every theory is a self-fulfilling prophecy that orders experience into the framework it provides.

Ruth Hubbard

In this chapter we present TermPedia; the document enrichment (DE) approach to facilitate reading comprehension, as a design product using Sol [1988]'s way of framework. This framework provides concepts and views behind TermPedia by presenting its way of thinking, governance, working, modelling, and supporting.

5.1 Ways of Framework

We use the ways of framework by Sol [1988] to present TermPedia; the document enrichment (DE) approach to facilitate reading comprehension. The ways of framework pays explicit attention to all the important aspects of presenting a design science artifact [Stojanovic et al., 2001]. It has been used by many researchers in the design, development and implementation of information system artifacts [Tumwebaze, 2016]. Sol's way of framework defines a set of concepts and views which we use to characterise TermPedia. As illustrated in figure 5.1, these concepts and views include a way of thinking, governance, working, modelling, and supporting.

5.2 Way of Thinking

TermPedia's way of thinking articulates the underlying principles advanced in facilitating the reading comprehension process. It sets a basis for the ways of governance, working, modelling, and supporting. We argue that document enrichment is an appropriate way to facilitate reading comprehension because it integrates contextually relevant information into electronic course literature used at universities. In the exploration and abstraction stages we noted that university students have insufficient vocabulary and background knowledge in specific subject areas, which

Figure 5.1: Sol [1988]’s Framework for Understanding Design Approaches

creates gaps in their reading process and hinders comprehension. Moreover, in many cases authors do not include graphical support for difficult concepts found in course materials. As a way to address these knowledge gaps we designed TermPedia to provide contextually relevant information for facilitating reading comprehension. The information provided for the content of university course documents included definitions, illustrations and background information for the technical terms that occur in them. The way of thinking of TermPedia is based on the interaction among people, process and technology [Keene, 2002].

People: Reading comprehension is defined as the ability to make meaning from text by integrating prior knowledge to the text that is being read. In this process text is decoded, words recognized and meaning constructed from formed sentences that constitute the overall text content. University students are the people whose reading comprehension ability we seek to facilitate. Our literature review and stakeholders’ survey showed that the reading comprehension process of university students is affected by insufficient background and vocabulary knowledge, ambiguous

vocabulary and lack of graphical support in text. These knowledge gaps motivate the students to look for information outside the contents of their course documents, so that they can understand what they read. On the contrary, this same motive interrupts the reading process, causes distraction and requires extra effort and time from the students. This initiative can be looked at as a search for satisfying solutions [Sol, 1988] to the factors hindering reading comprehension. The main aim of TermPedia is to provide in-text information for filling consequential knowledge gaps so that students stay within the course documents they read. This way, the students may use reasonable time and effort while reading and they shall not get distracted by looking for required information outside their course documents. TermPedia will do this by providing definitions and background information for technical terms from Wikipedia and illustrating the technical terms by using graphical images from Wikipedia info boxes where applicable.

Process: This research shows that the process of reading involves the ability of a student to understand the meaning of individual words in a document, connect these words to build sentences, paragraphs and full text bodies in order for them to understand the text content. Besides understanding individual words, the students also need some prior knowledge from which to infer meaning of the text they read. A gap in both vocabulary knowledge and background knowledge will force the students to look for missing information in order for them to understand what they read. Some things the students do in order to understand the meaning of difficult words and concepts are to look in dictionaries whether print copies or soft copies, look on the internet for the meaning of the words and ask their colleagues or lecturers. Such action consumes energy and time, leads students away from the read material and diverts attention of the reader from the text they are reading. Besides, they will easily forget what they have read if they do not understand the vocabulary used in the content of the documents they read.

TermPedia is designed to support the process of understanding the meaning of technical terms by defining and illustrating them. It also provides easy access to additional background information on the technical terms in case the definitions of such terms are insufficient for the reader to understand them.

Technology: Text wikification has been used as a document enrichment approach that identifies key phrases in text and links them directly to Wikipedia. Much as this approach provides easy access to a Wikipedia article that defines and discusses key phrases, it leads the students away from the read text and does not provide graphical help for the key terms where appropriate. Therefore, the problems of time, effort and distraction from the read material are not addressed. This is why we proposed TermPedia to integrate the definitions and illustrations of technical terms into students' course literature. This way, students will not be led away from the

literature they read in such for meanings and background information on technical terms that occur in such literature. This will help the students to use less time and effort during their reading process. In addition, graphical help given for the technical terms will help the students to retain the knowledge they acquire. Human language technologies such as technical term prediction, term sense disambiguation, technical terms definition and illustration and automatic hypertext generation where used in designing TermPedia. See chapter 2 for brief discussions on how the aforementioned technologies have been used in dealing with technical terms.

5.3 Way of Governance

The way of governance discusses the conditions to be met in the environment for applying TermPedia. It involves activities such as providing assurance to users that the required policies, processes structures and services are in place [Aregu, 2010] and aligned to their reading comprehension needs. We consider the way of governance as the quality control measures regulating the information used for enriching course content read by university students. These control measures were instituted to ensure that TermPedia addresses the main factors hindering reading comprehension by providing contextually relevant information for technical terms that occur in the students' course literature. The measures included term candidate prediction criteria, term sense disambiguation criteria and reading level options.

Term candidate prediction criteria controls which text strings in the students' course literature qualify as relevant technical terms. These criteria are necessary because data from the Wikipedia dump was very "noisy" and may qualify trivial words like "*the*" as potential term candidates. Document enrichment algorithms that were developed to govern the quality of predicted technical terms and the information provided for them are discussed in section 4.6.

Term sense disambiguation criteria were important for resolving the meaning of a term in the context where it occurs. These criteria were important because many technical terms are ambiguous since they carry different meaning in a specific text context. For this reason, term sense disambiguation is an important step in providing contextually relevant information for technical terms such their definitions, graphical representation and background information.

Reading level options help advanced readers not to be presented with relatively common technical terms as predicted candidates in the text they read. This implies that people reading at high levels e.g. post graduate level may be presented with more advanced technical terms compared to students reading at lower levels such as undergraduate.

5.4 Way of Working

The way of working describes the steps followed in using the TermPedia approach to facilitate the reading comprehension process. As already discussed, for TermPedia to facilitate the reading comprehension process, it provides contextually relevant information about technical terms which occur in the content of documents. Steps taken to facilitate reading comprehension can be grouped into two categories, namely way of working of users and way of working of TermPedia.

Way of Working of Users

If a user finds the contents of a document difficult to understand because it contains many unfamiliar technical terms, the user loads such a document into the TermPedia user interface for enrichment. The user then reads the enriched document while having easy access to the definitions of technical terms in the content of the uploaded document. The beauty is that term definitions are integrated into the document content by TermPedia, so users can concentrate on deriving meaning from the contents they read. Having easy access to such information facilitates reading because it allows readers to understand the meaning of difficult terms.

Figure 5.2: The TermPedia Document Enrichment Approach Use-Case Diagram

The use case diagram in figure 5.2 illustrates a list of steps typically defining interaction among university students and TermPedia to achieve the goal of facilitating the process of reading comprehension. University students and TermPedia are the actors involved in this case. TermPedia initiates facilitation of the students' reading comprehension process by requesting a document. After that students import their course documents into the TermPedia user interface and select their reading level preferences. TermPedia then predicts technical terms in the loaded document and disambiguates their meaning with respect to the context where the terms occur. TermPedia then enriches the uploaded document with contextually relevant definitions, illustrations and background information for the predicted terms. Information used for document enrichment is generated from Wikipedia as discussed in chapter 4 of this thesis. Steps taken by TermPedia in facilitating reading comprehension are discussed in subsection 5.4

Way of Working of TermPedia

TermPedia takes specific steps in enriching documents with contextually relevant information so that the process of reading comprehension is facilitated. As already mentioned, TermPedia provides such information by defining technical terms that occur in document content. Therefore, for TermPedia to facilitate reading it requests for a document, predicts technical terms in that document, resolves what the predicted terms mean in the context of the document from which they are predicted and integrates contextually relevant definitions of the predicted terms into the documents from which the terms are predicted. To carry out these tasks towards facilitating reading comprehension, TermPedia uses the following document enrichment algorithms:

Text Extraction Algorithm

The text extraction algorithm generates plain text from the content of a document that has been uploaded onto the TermPedia user interface. As text is extracted, other information such as images, tables and references are disregarded. The text extraction algorithm prepares the content of documents which have been uploaded onto the TermPedia user interface for term prediction.

Technical Term Prediction Algorithm

Technical terms are identified from the extracted plain text using the technical term prediction algorithm. This algorithm tries to find all the terms from a predefined dictionary in the content of a document that has been uploaded into the TermPedia

Figure 5.3: Steps in Applying TermPedia for Facilitating Reading Comprehension

user interface for enrichment. Term strings from the dictionary are compared with text strings in each paragraph of the document and if there is a match, then a technical term is predicted. The double arrows in figure 5.3 shows that there is a back and forth movement between the technical terms dictionary and document content during the step of predicting technical terms. Section 4.6 presents details of the technical terms prediction algorithms experimented with in the task of identifying difficult vocabulary that occur in text documents.

Technical Term Sense Disambiguation Algorithm

After a technical term has been predicted, the meaning of the term in the context of the document where the term occurs is resolved by the technical term sense disambiguation algorithm. The meaning of technical terms are resolved in context so as to provide contextually relevant definitions for the predicted terms. Figure 5.3 shows that the document enrichment algorithm for disambiguating technical term senses in context requires information from a predefined dictionary and the document in which the technical term occurs. Each possible meaning of a technical term from the dictionary is compared to the context of the paragraph where the technical term occurs in a document. If a term definition is similar to the term context, then that term definition is provided as vocabulary knowledge for it. Details

on technical terms disambiguation algorithms where discussed in section 4.5. It is necessary to resolve the meaning of technical terms in context because such terms have different meaning in specific document contexts.

Automatic Hypertext Generation Algorithm

The next task in the way of working of TermPedia is for it to indicate which pieces of text in the content of a document are technical terms and to integrate definitions of those terms into that document's content. This task was carried out by the automatic hypertext generation algorithm. We see from figure 5.3 that this algorithm requires predicted terms and their definitions as resolved by the technical terms sense disambiguation algorithm. This algorithm completes the document enrichment approach because it is responsible for providing easy access to vocabulary knowledge of predicted terms in tool tips within the enriched document.

Automatic hypertext generation is the process of applying web technology to convert plain text to interactive text by using hypertext markup language (HTML). In particular, the HTML `< a >`, and `< /a >` tags are inserted around predicted technical terms (plain text) so as to create hyper-links (interactive text). The uniform resource locator (*url*) of the Wikipedia page that provides relevant information for the predicted term serves as the hypertext reference (*href*) attribute of the `< a >` tag. For example, if the technical term *Precision* is predicted by the term prediction algorithm, the term will become a hypertext after the following transformation:

Example 5.4.1: Hypertext Generation

```
<a href="http://en.wikipedia.org/wiki/Precision and recall">Precision</a>
```

Each automatically generated hypertext has an extended functionality which allows definitions of the predicted technical term i.e., the first paragraph of the target Wikipedia page, to show in a tool tip as soon as a user moves a mouse pointer over the hypertext (mouse hovering action). The functionality of the hypertext is extended using both Java and PHP scripts. The Java script is responsible for generating the tool tip frame while, the PHP script is responsible for generating the content of the tool tip which is the definition of the technical term represented by the active hypertext. Term definitions are retrieved in real-time from the current version of on-line Wikipedia with an excellent speed. For example, in figure 5.4 we see a tool tip displaying the definition of *Precision*. If a user is not satisfied with the definition, he can conveniently link to the relevant Wikipedia page for more explanation by clicking on the hypertext. The information from the first paragraph of a Wikipedia page relevant to the technical term in question enriches electronic documents with text.

Figure 5.4: Document enrichment with Wikipedia text

Figure 5.5: Document enrichment with Wikipedia text and images

A bonus task performed by the automatic hypertext generation algorithm is to provide pictorial illustrations for a predicted term along with the term’s definition. The algorithm uses pictures available in infoboxes of Wikipedia pages to provide the pictorial support. Wikipaida infoboxes include more than ten million pictures that cover nearly all known knowledge domains. This wide coverage offers an opportunity for illustrating technical terms that occur in electronic documents in the absence of manually included pictorial illustrations. It is worth mentioning that this thesis does not look into the relevance of the Wikipedia images used for illustrating the predicted terms. We assume that the term sense disambiguation algorithm identifies relevant Wikipedia pages for predicted terms. Therefore, extracted images from such pages should provide informative pictorial support for predicted terms. A Wikipedia image is extracted by extending the PHP script which is responsible for extracting

the first paragraph of a target Wikipedia page to fetch the first image that occurs in the page’s infobox as well. The tasks of providing pictorial support for predicted technical terms is important because it facilitates reading comprehension and knowledge retention. As an example, in figure 5.5 we see that the first image in the infobox of the Wikipedia page with URL *http://en.wikipedia.org/wiki/Precision_and_recall* is an accurate illustration of the predicted term *Precision*.

5.5 Way of Modelling

The way of modelling refers to suitable representations deployed to model the major components of TermPedia. These models were identified as three algorithms, which provide comprehensive information for facilitating the reading comprehension process through a document enrichment approach.

Figure 5.6: TermPedia’s document enrichment approach sequence diagram

The three document enrichment algorithms identified included (i) term prediction, (ii) term sense disambiguation and (iii) hypertext generation algorithms. The sequence diagram in figure 5.6 demonstrates the TermPedia approach by showing how and in what order the model algorithms work together in enriching documents to facilitate reading comprehension. The approach addresses reading comprehension problems caused by insufficient vocabulary and

5.6. *WAY OF SUPPORTING*

background knowledge through integrating contextually relevant definitions of technical terms into the documents where the terms occur. This approach was applied to solve the reading comprehension needs of university students. Details of how the algorithms work together was outlined in section 5.4 and technical details of the algorithms were discussed in section 4.6 of this thesis.

5.6 Way of Supporting

Way of supporting represents the technologies that support the application of TermPedia, such as human language technologies, Internet technologies and web technologies. Human language technologies supported the application of TermPedia by providing a platform on which the document enrichment algorithms were designed. Specific human language technologies which support the application of TermPedia included term extraction, term prediction, term sense disambiguation and term definition and illustration. Chapter 2 discusses what is implied by the aforementioned technologies and how we intended to apply them in our document enrichment approach. Internet and web technologies support the application of TermPedia by enabling connectivity of Wikipedia data between computers and availing this data for integration into the content of documents. Hypertext markup language is an example of web technology which supports the integration of information into document content. See section 5.4 for details.

Chapter 6

Testing TermPedia's Algorithms

Verification is prudence

Imam Ali (A.S.)

This chapter describes how the performance of TermPedia's document enrichment algorithms was tested. The algorithms were tested to find out if they meet the expectations of predicting technical terms that occur in the content of document that contain such terms. The algorithms were also tested for their performance in providing contextually relevant information for technical terms. The chapter also describes the data that was used in the testing exercise along with presenting and discussing the test results. The test results indicate that the frequency-based document enrichment algorithm performs best at meeting the expectations of document enrichment for facilitating reading comprehension.

6.1 Objectives of Testing TermPedia's Algorithms

We presented a number of document enrichment algorithms in section 4.6. The presented algorithms were generically tested for their performance in predicting technical terms that occur in the content of electronic course documents read by students at universities. The algorithms were also tested for their effectiveness in providing relevant information for the predicted terms. A more specific goal for testing the document enrichment algorithms was to use the best performing algorithm in developing a web-based user interface. The user interface was developed and used for evaluating the concepts of document enrichment for facilitating reading comprehension through a user study and a user survey as discussed in chapter 7 of this thesis.

In principle, the performance of these algorithms was calculated based on their ability and accuracy in predicting, defining, and explaining technical terms that occur in the content of electronic documents. Section 1.5 of this thesis explains that document enrichment is an implementation and application of information extraction. Therefore, methods for testing information extraction systems were found suitable in testing the performance of TermPedia's document enrichment algorithms.

6.2 Methods of Testing the Algorithms

Lehnert et al. [1994] provides detailed explanations on standard methods for testing an information extraction system. These methods were developed over the course of a Message Understanding Competitions and they use sample statistics to compute performance results over the materials used for testing [Lowe and King, 2003]. The methods discussed by Lehnert et al. [1994] include, but are not limited to *precision* and *recall* which we apply for testing the performance of the document enrichment algorithms, including f_score , also known as $f_measure$. These testing methods were selected because they provide a possibility to measure the performance of the document enrichment algorithms in technical term prediction and term target selection. Term target selection is used here to represent term sense disambiguation.

Term target refers to the title of a Wikipedia page to which a predicted term is linked. In other words, a term target is part of a uniform resource locator (URL) of the Wikipedia page to which a document enrichment algorithms links a predicted term. The URL is automatically formed by a document enrichment algorithm that generates hypertext after a term has been predicted and its target selected.

Example 6.2.1: Illustration for term prediction and target selection

LINK depression Major depressive disorder

In example 6.2.1 the predicted term is depression and its selected target is Major depressive disorder. The URL formed in this case will be:

[http://en.wikipedia.org/wiki/Major depressive disorder](http://en.wikipedia.org/wiki/Major_depressive_disorder). This URL is then used as a value of the href attribute for the HTML (<a>) tag while a document enrichment algorithm generates hypertext for the predicted term, like this:

```
<a href="http://en.wikipedia.org/wiki/Major depressive disorder">depression</a>.
```

Since the document enrichment algorithms predict technical terms and select targets for the predicted terms, their testing was done in two steps. First, we tested the algorithms performance at technical term prediction and then we tested their performance in term target selection. By testing performance of the algorithms in predicting technical terms we try to find out how accurate the algorithms are at finding words and word phrases that present difficult concepts in a document. Testing performance of the document enrichment algorithms in selecting targets for predicted terms involves how accurate the algorithms are at resolving the meaning of technical terms in context. In other words, we test how accurate the algorithms are at linking predicted terms to a relevant Wikipedia page and providing contextually relevant information for such terms. The task of term target selection is therefore twice as

difficult as the task of term prediction because in order to select a relevant target, the predicted term has to be acceptable as a difficult concept in a given document. We therefore expect that the performance of the document enrichment algorithms at selecting relevant targets for predicted terms will always be lower than performance of the algorithms at predicting acceptable technical terms.

Some Expressions Explained

Before describing how *precision* and *recall* were used in testing the performance of the document enrichment algorithms, it is important to explain the meaning of some expressions that were used in the description and other parts of the thesis for clarity. The expressions to be explained include: (i) *data and data set*, (ii) *gold data set*, (iii) *testing data set*, (iv) *gold terms or gold technical terms*, (v) *predicted terms or predicted technical terms*, (vi) *gold term targets*, and (vii) *predicted term targets*

(i) Data and Data Set

Data is the content of any electronic document and a data set refers to the collective content of many electronic documents. These documents may contain structured or unstructured data. Structured data refers to data that is identifiable because it is written in some kind of format, like data between Wikipedia link markup. (See example 6.3.1 on page 111 of this thesis)

(ii) Gold Data Set

A gold data set is a collection of electronic documents that contain structured data which identifies technical terms that human experts have identified. Note that the gold data set may contain structured data for other purposes but we consider only those that identify technical terms.

(iii) Test Data Set

Test data set is a collection of an unstructured version of the gold data set. The test data set was created by removing all the indications of the humans' choice of technical terms from the gold data.

(iv) Gold Terms or Gold Technical Terms

Gold terms or gold technical terms are the structured pieces of data in a gold data set, which humans have selected as identifying technical terms.

(v) Predicted Terms or Predicted Technical Terms

Predicted terms or predicted technical terms are structured data which are automatically generated by the document enrichment algorithms from a test data set. These will be compared to the gold terms that human experts have identified.

(vi) Gold Term Targets

Gold term targets are titles of the Wikipedia pages to which structured pieces of data in the gold data set are linked. i.e. They are titles of Wikipedia page to which technical terms in the gold data set are linked.

(vii) Predicted Term Targets

Predicted term targets are titles of the Wikipedia pages to which the document enrichment algorithms automatically link the structured data they generate from the testing data set. i.e. They are titles of Wikipedia pages to which the document enrichment algorithms automatically link the technical terms they predict from the testing data set.

After explaining the above expressions, we will now describe *precision* and *recall* as used in testing the performance of the document enrichment algorithms. In the testing process predicted terms and targets are tested against the terms and targets in the gold data set.

Precision, and recall, for Technical Term Prediction and Target Selection

Precision (P) is a measure of the performance of a system in presenting only relevant items, while *recall* R is the measure of the performance of a system in presenting all relevant items. In the case of the document enrichment algorithms, *Precision* and *recall* can be explained as follows:

Precision for Technical Term Prediction

In the task of predicting technical terms by the document enrichment algorithms, *precision* (P) can be defined as the fraction of predicted terms that are relevant to a specific electronic document. i.e. a fraction of the gold technical terms that were predicted from each testing document by the document enrichment algorithms. *Precision* for predicting technical terms from each of the documents selected for testing is calculated by using equation (6.1).

$$P_{(term\ prediction)} = \frac{|\{(gold\ terms\} \cap \{(predicted\ terms\})|}{|\{(predicted\ terms\})|} \quad (6.1)$$

Precision for Technical Term Target Selection

The task of technical term target selection aims at providing relevant information for a predicted technical term. The testing for this task therefore seeks to measure how well the document enrichment algorithms perform at pointing the predicted term to additional information, which is relevant to the term in context. In this case the *precision* of selecting relevant targets for predicted terms by the document enrichment algorithms is defined as the fraction of predicted terms targets that are relevant to the predicted terms in the context of a given electronic document:

$$P_{(target\ prediction)} = \frac{|\{(gold\ term\ targets\} \cap \{(predicted\ term\ targets\})|}{|\{(predicted\ term\ targets\})|} \quad (6.2)$$

After calculating the precision for the task of predicting technical terms and the task of selecting relevant targets for the predicted terms from each document, average precisions are then calculated for each task. The average precision provides a measure of the overall performance of the document enrichment algorithms on the testing data set. Average precisions are calculated using macro-averaging. A macro-average is calculated by adding up all the different values of precision obtained from each testing document in the testing data set and dividing the sum by the number of documents in that set. Say for instance there are N documents in the testing data set, the macro-average precision *MAP* for this data set will be given by equation (6.3).

$$MAP = \frac{1}{N} * \sum_{i=1}^N P_i \quad (6.3)$$

MAP = macro-average precision over all documents in the testing data set
 N = number of documents in the testing data set
 P_i = value of precision for each document in the testing data set

Macro-averaging gives equal weight to each document in the testing data set. There is also another type of averaging called micro-averaging which would give equal weight to each per-document predicted term and selected target for the predicted term. In micro-averaging large documents dominate small documents because the value of precision greatly depends on the number of correctly predicted terms. For this particular testing micro-averaged results would therefore be a suitable measure of performance of the document enrichment algorithms on the large documents in

the testing data set. We use macro-averaging to calculate average precisions because we want to examine the effectiveness of the document enrichment algorithms on the large documents as well as the small documents [Manning et al., 2008].

Recall for Technical Term Prediction

In the task of predicting technical terms by the document enrichment algorithms, *recall* is defined as the fraction of relevant technical terms that are predicted from a specific electronic document in the testing data set by the algorithms:

$$R_{(term\ prediction)} = \frac{|\{(gold\ terms\} \cap \{(predicted\ terms\})|}{|\{(gold\ terms\}|} \quad (6.4)$$

Recall for Technical Term Target Selection

In the task of selecting relevant targets for predicted terms by the document enrichment algorithms, *recall* is defined as the fraction of relevant term targets that are selected for the predicted terms in the testing data set by the algorithms:

$$R_{(term\ prediction)} = \frac{|\{(gold\ term\ targets\} \cap \{(predicted\ term\ targets\})|}{|\{(gold\ term\ targets\}|} \quad (6.5)$$

After obtaining *recall* for the task of predicting relevant terms and the task of selecting relevant targets for the predicted terms, macro-averaging was used to calculate an average *recall* value for each task. The average *recall* shows the effectiveness of the document enrichment algorithms at predicting relevant terms and selecting relevant targets for the predicted terms over the testing data set. This average is calculated by adding up individual *recall* values obtained from each of the documents in the testing data set and dividing the sum by the number of documents in that set. For example, if there are N documents in the testing data set, the macro-average *recall* MAR would be given by equation (6.6).

$$MAR = \frac{1}{N} * \sum_{i=1}^N R_i \quad (6.6)$$

Where:

- MAR = macro-average recall over all testing data set
- N = number of documents in the testing data set
- R_i = value of recall from each individual document in the testing data

Recall is the vital score for this application, since users will be disappointed if there is no link from a term they don't know, but will not be disturbed when encountering terms they do know which also are linked to definitions.

F – score

F – score particularly measures the accuracy in the performance of a system and it is the harmonic mean of *precision* (P) and *recall* (R). The measure reaches its best value at 1 and worst at 0. Makhoul et al. [1999] provide a rather interesting history of *f – score* when they say that *f – score* was defined as a weighted combination of P and R because the community desired to have “a single figure of merit” to measure performance. They also provide a nice way for deriving the *f – score* equation as follows:

$$F = \left[\frac{\alpha}{P} + \frac{1 - \alpha}{R} \right]^{-1} = \frac{PR}{(1 - \alpha)P + \alpha R}, \quad 0 \leq \alpha \leq 1. \quad (6.7a)$$

The most popular value of $\alpha = 0.5$ and equation 6.7a reduces to equation 6.7b:

$$F = \frac{2PR}{P + R}, \quad \alpha = 0.5 \quad (6.7b)$$

6.3 Description of TestData

Two sets of data were used for the testing of the document enrichment algorithms. One of the data sets was collected from Wikipedia free text related to diseases and the other was free text data collected from the Merck manuals on-line library and Europeans Medicines Agency (EMA) reports. Free text here refers to electronic text documents, which are free from any kind of mark-up. We use data related to diseases, medicine, and drug prescriptions because information related to health care in general, contains many technical terms, most of which are difficult to understand for someone who is not familiar with them. It is also true that people often need information about some health issue because of one reason or the other. For example, someone may contract a rare disease and may want to understand it better by reading some documentation about the disease. If the information he comes across is clouded with unexplained technical terms, then his search for information may turn out to be very frustrating. In the end more questions may be raised after reading a document clouded with technical terms, instead of proving satisfying answers to the questions that the reader had before reading the document. If the document enrichment algorithms can predict and explain difficult terms in such documents then persons looking for information in them may have quick answers to their questions.

Sl. No.	Extracted disease page ids and titles
1	18628 Lambert–Eaton myasthenic syndrome
2	21937 Nitrogen narcosis
.	.
.	.
50	263801 Alien hand syndrome
51	276560 Orf (disease)
.	.
.	.
99	758763 Ice-cream headache
100	768604 Vasovagal response

Table 6.1: Sample titles of disease related articles extracted from the Wikipedia dump

Data Related to Diseases from Wikipedia Dump

There were two main steps in collecting data related to diseases from the Wikipedia dump. The first step was to extract free text document from the Wikipedia dump and then use the document enrichment algorithms to identify technical terms from these free text. The algorithms would also link the predicted terms to relevant Wikipedia articles. In order to achieve the tasks in this step, 200 unique Wikipedia page ids from the *infobox.csv* file that have a template-name Disease were extracted. A brief description of the *infobox.csv* file can be found in section 4.3. The unique ids were then used to extract 100 articles related to diseases from the free text Wikipedia dump. A sample title list for the 100 free text Wikipedia documents related to diseases is given in table 6.1. We refer to these 100 articles related to diseases as the *testing data set*, or testing documents throughout this chapter. Note that we implicitly rely on the authors and editors of Wikipedia as human experts in regarding their links as "gold". Each of the 100 free text documents had to be run through all the three algorithms for document enrichment. The algorithms then predicted link labels (technical terms) and targets for the predicted link labels from the free text documents. The predicted link labels and targets were then tested against the original link labels and targets which exist in the documents that were extracted from a Wikipedia XML dump.

The second step in collecting data related to diseases from the Wikipedia dump was to extract an annotated document from the XML dump of Wikipedia for each of the 100 documents extra from the free text Wikipedia dump. Example 6.3.1 shows a sample of text from the XML dump. The text in double square brackets i.e. `[[Viral hemorrhagic fever|hemorrhagic fevers]]`, represents a wikilink. In this case the

text to the right of the pipe| character is the wikilink label (hemorrhagic fevers) and that to the left of the pipe is the target of the wikilink label. It is important to mention here that this target also represents the title of a Wikipedia article. This example shows that the medical term hemorrhagic fevers is linked to a Wikipedia page with the title Viral hemorrhagic fever. The annotated documents from the XML dump of Wikipedia contained marked-up text indicating original wikilink labels and their targets. We refer to these 100 documents as the *gold data set* throughout this chapter. This data is available in the Wikipedia corpus in a file named *enwiki20101011-pages-articles.xml*, see section 4.3. From this data, it was possible to extract the original wikilink labels that existed in each of the 100 free text documents including their targets.

Example 6.3.1: Content of Wikipedia XML dump annotated text files

Like other [[Viral hemorrhagic fever|hemorrhagic fevers]],
Lassa fever can be transmitted directly from one human to another.

A document from the testing data set is therefore very similar to its counterpart in the gold data set, the difference is that the hypertext in a test document will be automatically generated by the document enrichment algorithms while the hypertext in the gold document are generated by collaborative authors who contribute articles to Wikipedia. The testing exercise is therefore to find out how efficient the document enrichment algorithms are at coming up with exactly the same explanatory links that occur in the gold documents. Easy as it may sound, this task is rather challenging because it tries to unify the opinion of all the collaborative authors on the issue of which words or phrases in any text document represent technical terms or important notions. To illustrate, assume that the same piece of text is given to a number of people from different knowledge backgrounds so that they may identify which part of the text represent technical terms and important notions. No doubt there will be many differences in opinion resulting in many different and similar technical terms and important notions identified from the same piece of text. What the testing process tries to do is, to find out which of the terms or important notions are identified in the same way by a majority of people from different knowledge backgrounds. The testing also hopes that if other people were given a different piece of text containing the same terms and notions, they will still recognize them in the same way.

Another challenge in testing the performance of the document enrichment algorithms against the gold set documents is that the collaborative authors who contribute articles to Wikipedia links pieces of text for many different reasons and not only because they are technical terms or important notions. Because of this, almost

any piece of text can be a hypertext and therefore qualify to be called a technical term or important notion since we use wikilink labels to represent terms and important notions. This also means that if all the hypertext are brought together and used for predicting technical terms and important notions, it is most likely that not all the predicted terms and notions shall be relevant for the document in question. The challenge is therefore to find out if the document enrichment algorithms are able to select wikilink labels, which represent technical terms and important notions across a variety of knowledge domains. Also, there is hardly a limit to the length of text strings that exist as wikilink labels. These vary from singly characters to text string as long as a paragraph. For example, “0.01 km² (1 ha)”, “such as”, “Zymotic Diseases”, and “WHO Disease Staging System for HIV Infection and Disease in Adults and Adolescents”, are all wikilink labels from the English Wikipedia and thus would represent technical terms or important notions in any document.

While creating the testing data set and the gold data set, only data in the text body of a Wikipedia page was considered. Header information indicating revision of articles and meta-data information were disregarded. In addition, information after the main text body like list of references, related pages, and footnotes were also not considered.

Data from EMA Reports and Merck Manuals

The Merck Manuals, are a series of healthcare books for medical professionals and consumers with enhanced on-line versions¹. The European Medicines Agency (EMA) is a decentralised agency of the European Union, responsible for the scientific testing of medicines developed by pharmaceutical companies for use in the European Union. The testing reports can be found at <https://www.ema.europa.eu/en> [EMA, 2005]. Ten (10) random short articles downloaded from the Merck manuals and 10 random EMA medicine testing reports were downloaded and used for testing the document enrichment algorithms.

Since this data was not part of the Wikipedia corpus, they did not have original links therefore their testing did not include term target selection. The data was manually annotated for medical technical terms by humans. Before this annotation, plain text copies of the data were reserved and annotated using the document enrichment algorithms. At the end, testing was made for the annotations discovered by humans against the annotations discovered by the document enrichment algorithms.

¹Can be found at <http://www.merckmanuals.com/professional/index.html> [Merck, 2005].

6.4 Presentation of Test Results

Test results for performance of the document enrichment algorithms will be presented in two subsections with regard to term prediction and term target selection. For an explanation of what is meant by technical term target see section 6.2. The results will be presented for both Wikipedia data, and EMA and Merck data to show the performance of the document enrichment algorithms in term prediction and term target selection on training data and real-world data respectively. Details on term and target selection algorithms are discussed in section 4.6, but for emphasis and clarity, it is wise to briefly mention the idea behind term prediction. For all the three techniques of term prediction, all possible n-grams at a paragraph level were generated from the documents for which the terms were to be predicted. The n-grams were then stored as elements of an array and the array was sorted using different criteria. The order of the n-grams in the sorted array determined the criterion behind a term prediction technique. The following subsection briefly mentions the different criteria behind the term prediction task.

Term Prediction from Electronic Documents

Term prediction was carried out using three algorithms, which used the basic techniques discussed in chapter 4. The baseline technique for term prediction is built on a random selection of candidate n-grams. This is the term prediction technique name *blindfold term prediction* (BTP). In this technique n-grams generated from electronic documents to be enriched are checked against terms in the dictionary created from wikilink labels (anchor texts) without considering any order or frequency of the generated terms. If an n-gram is found in the dictionary, then it is predicted as a technical term in the document from which the n-grams were generated.

The second technique for term prediction is based on a *longest match criterion* (LTP). Here the n-grams generated from a document to be enriched are sorted in descending order of their length at paragraph level and stored as array elements. This means that the longest n-gram from each paragraph of a document is contained at position [0] of the array; in other words it is the first element in that array. Ngrams from the array are then compared with the terms in the dictionary and if the n-gram is found in the dictionary, it is marked as a predicted term in the document of interest. The comparison is done from the first to the last element in the array of n-grams sorted by length.

The third criterion for technical term prediction *frequency based technical term prediction* (FTP) is based on term frequencies and depends on a term likelihood probability as discussed in section 4.6. Term likelihood is a probability calculated for each term that occurs as a uni-gram or a bi-gram in the term dictionary. Terms that occur as tri-grams

Algorithms	Predicted Terms	Term length	Term likelihood
BTP	-developmental	13	0.0103
	-developmental disorders	23	0.0625
	- disorder	8	0.0074
	- grouped into	12	0.0005
	-pervasive	9	0.0026
	-specific	8	0.0012
LTP	-pervasive developmental disorders	33	1.0000
	-specific developmental disorder	31	1.0000
	-grouped into	12	0.0005
FTP	-specific developmental disorder	31	1.0000
	-pervasive developmental disorders	33	1.0000
	-grouped	7	0.0012

Table 6.2: Term prediction by BTP, LTP, and FTP from example sentence 6.4.1

and longer are automatically given a term likelihood probability of 1. In principle term likelihood is interpreted as the probability that an n-gram occurs in Wikipedia as a wikilink label divided by the total number of times the term occurs in the entire Wikipedia corpus. In the array of n-grams generated from this method the elements are sorted in such a way that the n-grams with high term likelihood probabilities come first. During the time of comparing the generated n-grams with the terms in the dictionary, n-grams with high term likelihood probabilities are first considered. If these n-grams are found in the dictionary, they are predicted as technical terms in the electronic document from which the n-grams are generated.

To illustrate how the term prediction algorithms work, assume that we have the sentence in example 6.4.1:

Example 6.4.1: Sentence extracted from a Wikipedia article:

They can be grouped into specific developmental disorder and pervasive developmental disorders.

The terms predicted from the sentence in example 6.4.1 by the three algorithms for term prediction are shown in table 6.2. Notice that the BTP algorithm which depends on a random criterion for term prediction, i.e. first-found-first-predicted, finds more technical terms in the sentence. In most cases this algorithm splits up longer phrases that should occur together as a technical term. For instance, the phrase specific developmental disorder is split up in to three terms i.e. specific, developmental, and disorder. We expect that in the performance testing of the term prediction algorithms, BTP will have lower precision for term prediction compared to LTP and FTP algorithms.

This is because BTP will predict more terms compared to LTP and FTP. Moreover, with the assumption that many technical terms occur as phrases longer than unigrams, the recall of BTP for term prediction will also be comparatively low because most of the terms predicted by this algorithm may not occur in the original Wikipedia documents as wikilink labels or gold technical terms. This also means that LTP and FTP are expected to be better methods for finding terms that exist in the original marked-up Wikipedia documents.

Take for instance the sentence in example 6.4.1, which occurs in the marked-up Wikipedia article as in example 6.4.2 with two gold technical terms, namely specific developmental disorder and pervasive developmental disorder. In the event that a strict pattern-matching criterion is used while testing the technical terms predicted from electronic documents, specific developmental disorder will be the only correct predicted term. (Compare table 6.2 and example 6.4.2) This is true because it is the only predicted technical term that exactly matches its form in the original marked-up document. The term pervasive developmental disorders predicted by both FTP and BTP as shown in table 6.2 will not indicate a correct instance of the term's prediction because of the letter (s) at the end of this term. The comparison between table 6.2 and example 6.4.2 shows that of all the terms predicted using the various algorithms for term prediction; there would be only one correct predicted instance while using a strict pattern-matching criterion during the testing process. We work around this problem by counting a term prediction instance as correct if its singular form occurs as a gold term in the Wikipedia marked up document of interest. In which case both specific developmental disorder and pervasive developmental disorders will be counted as correct and relevant instances of predicted terms.

Example 6.4.2: Sentence extracted from a marked-up Wikipedia article:

They can be grouped into [[specific developmental disorder]] and [[pervasive developmental disorder]]s.

LTP and FTP algorithms are expected to have close values of precision and recall since their performance in predicting terms are very close, see table (6.2). In this example the two algorithms differ in only one instance of term prediction, where LTP predicts “grouped into” as a term and FTP predicts, “grouped” as a possible instance of a technical term. This predictions clearly show that the string grouped into is longer than grouped therefore it is preferred when LTP is used. It is also clear that the string grouped had a higher value of term likelihood probability of 0.0012 compared to that of grouped into which was 0.0005. This is why FTP predicts

grouped as a term and not grouped into.

Results for Terms Prediction from Wikipedia Data

In order to maintain a uniform ground for term prediction, we used a random target selection for the terms predicted using the three criteria of term prediction. The outcome of this experiment shows that LTP performs slightly better than FTP at term prediction with a recall of 79.0% compared to FTP at 78.8% recall. (See table 6.3) The difference in the performance of these two algorithms is rather negligible since they both predict technical terms at a recall of 80% if their recall values are rounded off to one decimal point. This result literally means that both LTP and FTP predict 80% of the gold terms from the original Wikipedia marked-up articles.

Algorithms	#docs	Averages		
		Precision	Recall	f-score
BTP	100	0.118	0.641	0.200
FTP	100	0.152	0.797	0.255
LTP	100	0.152	0.801	0.256

Table 6.3: Over all testing of Wikipedia data technical term prediction

As expected, both LTP and FTP perform better than BTP, which has a recall of 63.2% for term prediction. The values of recall show that on average each criterion for term prediction find more than 60% of the wikilink labels (anchor texts or terms) that exist in the original Wikipedia marked-up document. The higher values of recall obtained by LTP and FTP compared to BTP shows that the later algorithms of term prediction improve the accuracy of predicting technical terms by 16%.

Most of the 20 to 40% missing prediction is accounted for by the weakness in the pattern-matching criteria for testing. Some of the missing precision is accounted for by the process of making plain text. For example, if a wikilink label exists as `[[child]]’s` in the raw Wikipedia data, when clean text is generated by use of the WikiExtractor.py script, the string `child’s` will be formed. While comparing the `child’s` n-gram with the list of terms in the dictionary, it will be marked as a predicted term because it exists in the dictionary as such. Note that even though the string `child` exists in the term dictionary, it will not be recognized at this point because the term prediction criteria matches whole words and phrases. We solve this problem by considering the possessive form of a noun as a relevant prediction of a technical term as long as the noun derived from that possessive form exist in the gold document of interest (gold term). Note that for this to work the string representing

Figure 6.1: Graph showing the effect of reducing the percentage of n-grams permitted to be predicted as technical terms on precision, recall and f-score while using the LTP algorithm.

the possessive form of the noun has to end with “’s” because this is what generates the problem as shown in the case of `[[child]]’s`. In which case the possessive noun `child’s` will be considered a relevant instance of technical term prediction because the noun `child` exists as a gold term.

Examining the low values of precision show that most of the n-grams generated from various disease articles selected for testing exist as wikilink labels in Wikipedia. We are not concerned with the low precision values because it helps to vary the number of n-grams in each document that may be predicted as technical terms. Naturally, reducing the number of n-grams to be predicted as technical terms from each document will increase precision up to a certain maximum value as shown in figure 6.1. The figure shows that the maximum value of precision is 0.385. This value is obtained when (10%) of the candidate n-grams are permitted to be predicted as

technical terms in each testing document. Note that the scores in figure 6.1 are as a result of using the LPT algorithm and they would be different from the scores obtained by using BTP and FTP. Further interpretation of figure 6.1 is that the value of recall reduces as the number of candidate n-grams permitted to be predicted as technical terms are reduced. The lowest value of recall at the point when (10%) of the candidate n-grams are permitted to be predicted as technical terms is 0.222. The LTP algorithm performs best at the point when (20%) of the candidate n-grams are allowed for term prediction from each testing document. This is because the dots representing precision, recall and f-score are closest at this point as can be seen in figure 6.1.

A more important reason for limiting the number of n-grams to be predicted as technical terms has to do with the fact that the idea of document enrichment is designed for people with different background knowledge on various knowledge domains. If someone has a lot of knowledge in a certain knowledge domain, he or she may not need to look up the meaning of most of the terms. A sparsely enriched document would therefore be convenient for such a person in the event that the few predicted terms are the most difficult ones in the given document. In addition, sparsely predicted terms do not clutter documents with hypertext therefore minimally affecting the reading process. Persons with no or little knowledge in the knowledge domain of the documents they read may need to look up the meaning of most of the terms in the document. Increasing the number of terms to be predicted for such a reader with less background knowledge is useful because it provides easy access to necessary information through document enrichment.

The number of n-grams from a document to be predicted as technical terms may be calculated by taking a certain percentage of n-grams from that particular document that exist in the term dictionary. For example, if a document has ten (10) n-grams which exist in the terms dictionary, taking 20% of these n-grams means that only $2 = (0.2 \times 10)$ of the n-grams will be predicted as technical terms in that document. Another way to vary the number of n-grams to be predicted in a document intended for enrichment is by considering the term likelihood probabilities of the n-grams. These probabilities range from 0.0001 to 1 therefore it is possible to allow only n-grams below a certain probability value to be predicted as technical terms. Varying the percentage or term likelihood probability of n-grams to be predicted as technical terms helps to reveal which algorithm for term prediction is stable and has a better performance. The stability of each document enrichment algorithm is measured by their *f-score* given different n-gram percentages or term likelihood probabilities.

By closely studying figure 6.2 we see that the best *f-score* values are obtained by using 30% of the candidate n-grams selected for term prediction from each of the

Figure 6.2: Permitted percentage of terms to be predicted for each document

100 documents selected for testing. The term prediction algorithm that performs a bit best while varying the percentage of n-grams to be predicted as technical terms is FTP which has a maximum f -score value of 50% at the point when 30% of the candidate n-grams are allowed to be predicted as technical terms. It is worth noting that there is a drastic raise in the f -score values for FTP between 20% and 30% selected candidate n-grams. This indicates that many relevant terms are lost by the FTP algorithm if the percentage of candidate n-grams selected for term prediction is below 30. The lowest f -score value of 9% is given by the BTP algorithm at the point when 10% of the candidate n-grams are selected for term prediction.

Figure 6.3 shows that LTP performs better at term prediction while varying the term likelihood probability of candidate n-grams compared to FTP and BTP. The best f -score value for term prediction is 53% given the term likelihood probability of 0.09. The dots in this figure generated by varying the term likelihood probability of candidate n-grams follow a smooth gradient with no raise or fall in the value of

Figure 6.3: Upper limits of term likelihood probability for terms to be predicted

f -score for all the three algorithms. This shows that term likelihood probability is a better method for varying the number of candidate n-grams, which are to be predicted as technical terms in documents intended for document enrichment. As a result, the number of technical terms predicted in a document raises gradually as the term likelihood probability of the candidate n-grams are raised. For this reason, the user is not overwhelmed with hypertext in one instance or presented with terms that are most likely irrelevant for context comprehension.

Comparing figure 6.2 to figure 6.3 it can be concluded that FTP is a better algorithm for term prediction because it maintains high values of f -score in both figures. While varying the percentage of candidate n-grams to be predicted as technical terms, FTP's f -score ranges from 38% to 51% before it starts dropping. The f -score values attained by varying the term likelihood probability of candidate n-grams to be predicted as technical terms for FTP range from 41% to 52% before it starts to drop. If we compare these results to that the LTP

algorithm, we see that varying the percentage of candidate n-grams to be predicted as technical terms, LTP's f -score ranges from 30% to 32% before it starts dropping. The f -score values attained by varying the term likelihood probability of candidate n-grams to be predicted as technical terms for LTP range from 41% to 53% before it starts to drop. These results show that FTP is stable and better at term prediction compared to LTP and BTP because its f -score values from both figure 6.2 and 6.3 are very close in value and relatively high. It can therefore be said that to predict the most relevant technical terms for each document according to the relevance of wikilink labels, it is best to use the FTP algorithm. The best results will be achieved when FTP allows only candidate n-grams with a term likelihood probability greater than 0 and less than 0.09 to be predicted as technical terms.

Terms Prediction from EMA Reports and Merck Manuals

The document enrichment algorithms were developed and tested on data from Wikipedia. It was also important to find out how these algorithms will perform in predicting terms from other data because the document enrichment idea would eventually be applied to electronic education materials for higher learning. We selected data from EMA reports and Merck manuals for this testing. See section 6.3 for a description of these data and the motivation in choosing them for testing the performance of the document enrichment algorithms.

As in the case of experiments with Wikipedia data, the data from EMA and Merck also have testing and gold data sets. The difference is that the gold data sets for EMA and Merck do not have wikilink labels to represent technical terms. Therefore, we had to manually tag the technical terms in the EMA and Merck gold data sets. During the term tagging process from these data sets, only medical terms were put into double square brackets. Double squares brackets were used for tagging the medical terms for uniformity because in the Wikipedia gold data set the wikilink labels are also tagged between these brackets. Example 6.4.3 shows how medical terms were tagged in the EMA and Merck data sets.

Example 6.4.3: Manual medical term tagging for EMA and Merck data sets

[[Polyps]] are not as common as in the [[colon]].

Manual technical term tagging from EMA and Merck data was very tedious and in order to do it well we tagged only one (1) paragraph from each document in the gold data sets. In total we had 20 paragraphs each of an average size of 540 characters and 82 words. For each selected paragraph we preserved a clean text version. LTP term

prediction algorithm was then used to predict terms from the clean data for both EMA and Merck. LTP was used at its best threshold of allowing only candidate ngrams that have a term likelihood probability greater or equal to 0.09 to be predicted as technical terms (See figure 6.3). We hope that the testing results of terms predicted from these paragraphs against the terms that were manually tagged from the same paragraphs shall give a clear clue on how well technical terms can be predicted by the LTP algorithm from electronic text that originate in Wikipedia data. The task of the testing process is thus to find out how accurate document enrichment performs at predicting technical terms that are relevant to the medical knowledge domain. Since technical terms in the gold data set are not linked to Wikipedia, term target selection testing for EMA and Merck data is not necessary.

Doc.I D	#Words / #Chars	Gold	All TTs	Overla p	Misse d	Ne w TTs	P (Overlap) (All TTs)	R (Overla p) (Gold)	F (2RP) (P+R)
454	78 / 403	2	2	1	1	1	0.500	0.500	0.600
458	60 / 439	6	7	6	0	1	0.857	1.000	0.667
490	66 / 458	17	13	11	6	2	0.846	0.647	0.923
564	98 / 542	11	9	7	4	2	0.778	0.636	0.762
606	102 / 578	12	10	7	5	3	0.700	0.583	0.733
616	94 / 585	6	7	6	0	1	0.857	1.000	0.700
638	100 / 610	11	10	8	3	2	0.800	0.727	0.923
643	98 / 643	7	5	4	3	1	0.800	0.571	0.500
784	130 / 754	5	5	3	2	2	0.600	0.600	0.667

Table 6.4: Statistics of terms predicted by LTP against gold set of EMA

Key for table (6.4)

- #Words/#Chars = number of words/number of chars in (10) paragraphs used for testing of LTP algorithm on EMA data
- All TTs = number of terms that were predicted by LTP from EMA testing data.
- Missed = number of terms in EMA gold data set that were not predicted by LTP
- New TTs = number of terms that were predicted by LTP but are not in EMA gold data.
- Gold = number of terms in EMA gold data.
- Overlap = number of terms in EMA gold data set that were predicted by LTP.
- P = precision, R = recall, F = f -score

Testing Results for Term Predict from EMA Reports

From table (6.4) we see that the LTP term prediction algorithm performs at an average *precision* of 73% while predicting technical terms that are relevant for the

EMA data. This *precision* score also indicates that on average only 20% of the terms predicted by LTP may be less important for the data in EMA reports. Looking at the totals in table (6.4), it can be seen that out of the 91 technical terms which are predicted by LTP, 67 of them are relevant for the EMA data. LTP misses to predict a total of 29 terms that are relevant for the EMA data set and predicts 24 new terms that were not identified while manually tagging technical terms from this data. The extra 24 terms are considered less relevant for data from the EMA reports.

On the contrary considering that there can be a difference of opinion in identifying technical terms from documents, some of the terms that do not appear relevant as predicted by LTP may actually be important to someone else. This is especially true because there was only one (1) person involved in tagging terms from the EMA gold data. Additional information for any data set is important because the document enrichment idea is designed for people who have different knowledge backgrounds from various knowledge domains.

Testing results for the EMA data also has a high average *recall* score of 70% as seen from table 6.4. This means that the performance of LTP at predicting all the relevant technical terms for data from the EMA reports is quite good. The totals in table 6.4 show that out of the 96 terms which are relevant for the EMA data, 67 of them were predicted. LTP misses to predict only 29 terms out of the 96 which are relevant for the EMA data.

The average *f-score* value of 71% indicates that LTP performs well above average at predicting only relevant technical terms and predicting all the relevant terms for the EMA data set. This is a good performance compared to the *f-score* value of 25% attained by LTP while predicting technical terms from the Wikipedia data set. (See table 6.3)

The general higher scores of *precision*, *recall*, and *f-score* attained by LTP in predicting technical terms from the EMA data set compared to the scores obtained while predicting technical terms from the Wikipedia data is attributed to the fact that the length of the text in the EMA data set are very sort compared to the length of the text in the Wikipedia data. A technical explanation can be given that LTP was used at its best performance while predicted technical terms from the EMA data and at its lowest performance while predicting terms from the Wikipedia data set. As already mention, LTP was used at a threshold of allowing candidate n-grams which have a term likelihood probability greater or equal to 0.09 to be predicted as technical terms from the EMA data, but while predicting terms from the Wikipedia data, the threshold for candidate n-grams was greater or equal to 0. These threshold probabilities mean that LTP predicted fewer technical terms for the EMA reports, most of which were relevant. Therefore LTP attained higher *precision*, *recall*, and *f-scores* for EMA data compared to the scores from Wikipedia data.

Test Results for Term Predict from Merck Manuals

Doc.I D	#Words / #Chars	Gold	All TTs	Overlap	Missed	New TTs	P (Overlap (All TTs))	R (Overlap (Gold))	F (2RP (P+R))
215	30 / 157	12	10	9	3	1	0.900	0.750	0.818
262	40 / 215	10	6	6	4	0	1.000	0.600	0.750
283	36 / 262	6	3	3	3	0	1.000	0.500	0.667
380	58 / 283	8	7	6	2	1	0.857	0.750	0.800
452	70 / 380	12	10	9	3	1	0.900	0.750	0.818
465	65 / 423	7	5	5	2	0	1.000	0.714	0.833
101	65 / 423	10	17	17	2	0	1.000	0.805	0.884

Table 6.5: Statistics of terms predicted by LTP against gold set of Merck

Key for table (6.5)

- #Words/#Chars = number of words/number of characters in (10) paragraphs used for testing of LTP algorithm on Merck data
- All TTs = number of terms that were predicted by LTP from Merck testing data.
- Missed = number of terms in Merck gold data set that were not predicted by LTP.
- New TTs = number of terms that were predicted by LTP but are not in Merck gold data.
- Gold = number of terms in Merck gold data.
- Overlap = number of terms in Merck gold data set that were predicted by LTP.
- P = precision, R = recall, F = f -score

Table (6.5) contains results for the performance of LTP at predicting technical terms from the Merck data set. Like in the case of EMA data, LTP is used at its best threshold of allowing only candidate n -grams with a term likelihood probability greater or equal to (0.09) to be predicted as technical terms. However, the performance of LTP at predicting technical terms from the Merck data is better than its performance in predicting technical from the EMA data as shown by *precision*, *recall*, and *f-score* values obtained by LTP for both data sets. The totals in table (6.5) show that LTP predicts 104 terms from the Merck data set, a number which is less than the 119 terms that were manually tagged from the same data. Most of the 104 terms predicted by the LTP algorithm are actually relevant for the Merck data that is why the algorithm obtains a high *precision* score of 90%. The *precision* is also high because only 11 of the terms predicted by LTP are less relevant for the Merck data as shown by the total in New TTs column in table (6.5).

High *recall* of 75% shows that the LTP term prediction algorithm is good at

predicting all relevant terms for the Merck data. LTP misses only 26 of the 119 technical terms, which are relevant for the Merck data set as shown in column Missed of table (6.5). The high values of both *precision* and *recall* results in the high *f-score* value of 80%. This indicates that LTP has a good performance in predicting only technical terms that are relevant for the Merck data and predicting all technical terms that are relevant for the Merck data.

LTP term prediction algorithm performs better at predicting technical terms from Merck data compared to its performance at predicting technical terms from the EMA data because of their content. The Merck data provide more general information about HIV/AIDS while the EMA data provides specific information about medicines used for treating HIV/AIDS. Therefore, the data from EMA may have some technical terms that do not exist in the terms dictionary created by using data from Wikipedia. For example, the term *darunavir mylan* exists as a term in the EMA data but by the time of writing this research it did not exist in the dictionary of terms generated from Wikipedia data. Even though this is true, most of the technical terms from the EMA data do exist in the Wikipedia term dictionary.

Term Target Selection for Wikipedia Data

Term target selection is a more difficult task compared to term prediction because it involves providing relevant information for a predicted term. This means that the predicted term has to be genuine for a document in question and the selected target has to be relevant to the context where the term occurs in the document. The LTP algorithm which proved to be the best method for term prediction (see figure 6.3) was used to predict terms from the testing data set before selecting their targets. This algorithm was used at its best threshold, which allows only candidate n-grams with a term likelihood probability greater than or equal to ≥ 0.09 to be selected as technical terms. After the terms had been predicted, the term target selection criteria were then used to select which Wikipedia article provides the most relevant information for the predicted terms in context. In the case when a predicted term has only one (1) possible target there is no challenge presented because the target is not uncertain. In such a case the single possible target is selected as a target for the predicted unambiguous term. A challenge in term target selection is presented when a predicted term has more than one possible target. This is why the criteria for term target selection is developed only for terms which have a possibility to be linked to more than one (1) Wikipedia article, i.e. ambiguous terms. For any of the term target selection criteria described below, if the requirements are not met a random target is selected for the predicted term. But if the requirements are met and there are ties, a random target is selected among the ties and if the requirements are met

and there are no ties, the target with the best disambiguation value is selected for a predicted term. In addition, all term and target patterns are matched in a case insensitive way.

Target selection method	Predicted term	Selected target
<i>Data-driven target selection</i> (DTS)	depression	Long depression <i>Frequency</i>
<i>based target selection</i> (FTS)	depression	Major depressive disorder
<i>Incoming-link target selection</i> (LTS)	depression	Major depressive disorder
<i>Term pattern-match target selection</i> (PTS)	depression	Great depression in the united states
<i>Random target section</i> (RTS)	depression	Depression (geology)
<i>Twin target selection</i> (TTS)	depression	Depression

Table 6.6: Example of target selection for a predicted ambiguous technical term

There are six criteria for target selection used by the document enrichment algorithms to provide the targets of predicted terms, which are ambiguous. All these criteria try to resolve the sense of an ambiguous term in context before selecting a target for the term. The first criterion for ambiguous term target selection is based on word overlap therefore we named it *data-driven term target selection* (DTS). In this criterion words from the terms definition and the term sense databases are overlapped with the words from the term context. For information on the term definition and the term sense databases see section 4.3. The term context is represented by the paragraph in which the predicted term occurs in the testing data set. After words from the databases have been overlapped with words from the term context for each possible target, the target with the highest degree of word overlap is then selected as the target for the predicted term. We can say that this criterion resolves the sense of a predicted ambiguous term by word overlap.

The second criterion for target selection depends on the frequency at which a predicted term is used to refer to a particular target. If the predicted term is most of the time used to refer to some target, then that target is selected for the predicted term. This criterion is named *frequency-based target selection* (FTS). By selecting the target “Major depressive disorder” for the term “depression” as indicated in table 6.6, FTS shows that this term is most of the time used to refer to that target. This term target selection criterion therefore resolves the sense of an ambiguous term by determining how frequently it is used to refer to possible Wikipedia articles.

The third criterion for ambiguous term target selection depends on incoming links overlap. Incoming links are titles of all the Wikipedia articles that refer to a given Wikipedia article. This criterion has the underlining principle that documents with similar incoming links may have related content. For this reason, similar documents refer to them. In this case, the incoming links to all the possible different targets for a predicted ambiguous term are compared with the incoming links to the targets of all

unambiguous terms in the document of interest. The target with the highest number of overlapping incoming links is then selected for the predicted ambiguous term. This criterion is named *incoming-link target selection* (LTS). This criterion resolves senses of ambiguous terms by trying to find documents of similar content as suggested by incoming links. In table 6.6 “depression” is linked to “Major depressive disorder” which is very interesting because the Wikipedia article from which this example is taken talks about child developmental disorders. This suggests that the document in the testing data on child developmental disorders has similar content with the Wikipedia article titled “Major depressive disorder”.

The fourth criterion for target selection depends on a pattern-match criterion. If the pattern of the predicted ambiguous term is found in the target string, then that target is selected for the term. The target selection is done on a first-come-first serve basis, implying that the first target encountered with an occurrence of the technical term pattern in it is selected. We named this target selection criterion *term pattern match target selection* (PTS). Looking at table 6.6, this criterion selects a Wikipedia article with the title “Great depression in the united states” as a target for the ambiguous term “depression” because this term is contained in the target string.

The fifth and baseline term target selection criterion is the *random target section* (RTS), here a random target is selected for a predicted ambiguous term. RTS is the baseline line criterion because it does not involve any kind of sorting or statistics. RTS does not involve any ambiguous term sense disambiguation criteria.

Last but not least, the sixth criterion for ambiguous term target selection tries to find targets whose patterns are the same as the predicted terms. For this reason, we named it the *twin target selection* (TTS). As shown in table 6.6 the ambiguous technical term “depression” is linked to a Wikipedia article with the title “Depression”. Patterns in the predicted ambiguous term and selected target are exactly the same in the event that their character cases are ignored. This criterion resolves the sense of predicted ambiguous terms by selecting a Wikipedia page title, which has the same sting with the predicted term.

It is worth mentioning that the testing results for these methods will not portray the absolute potential of most of the target selection criteria. This is because in many cases some of the ambiguous term target selection criteria do not return a positive value since not all term targets have the necessary data for these criteria. For example, out of the 8, 605, 586 Wikipedia page titles which are used as targets for predicted terms, only 3, 165, 298 of them have incoming links information required by the LTS method and 3, 223, 065 titles have term definition and sense information which is required by the DTS method. We therefore concentrate on finding out whether the methods that do not select random targets for predicted ambiguous terms perform better than the baseline method RTS.

After selecting targets for predicted technical terms, the targets of the terms predicted from the testing data set are then compared against the target of the technical terms which occur in the gold data set for each document pair. The comparison is made on the values of *precision*, *recall*, and *f-score* attained by the document enrichment algorithms for a pair of documents in the gold and test data sets.

In addition to the target testing methods of *precision*, *recall*, and *f-score* there is *disambiguation accuracy* (DA) of target selection for ambiguous terms targets. These methods all test the performance of the target selection algorithms at providing relevant targets for predicted technical terms. *DA* is the percentage of predicted ambiguous technical terms that exist in the gold data set which point to the same targets. Therefore, the *DA* for any given document in the testing set X_{DA} is the total number of predicted ambiguous terms and targets which occur exactly as they do in the gold data set $X_{T \ T \ G}$ divided by the total number of predicted ambiguous terms which occur in the gold data set X_{ig} multiplied by 100.

$$X_{DA} = (X_{TTG} \div X_{TG}) \times 100 \quad (6.8)$$

Results for Target Selection from Wikipedia Data

Methods	#Docs	Averages			
		Precision	Recall	f-score	Disambiguation accuracy
DTS	100	0.319	0.502	0.390	0.435
FTS	100	0.409	0.636	0.498	0.907
LTS	100	0.372	0.576	0.452	0.720
PTS	100	0.344	0.534	0.418	0.548
RTS	100	0.319	0.490	0.387	0.407
TTS	100	0.378	0.585	0.460	0.773

Table 6.7: Over all testing of Wikipedia data term target selection

Unlike in the case of technical term prediction algorithms, the technical term target selection criteria have very higher precision because the LTP algorithm for term prediction was uses at its best performance as mentioned above. These high precisions occur because many of the technical terms predicted by LTP at its best performance exist in the gold data set against which they are tested. Since many of the predicted terms have their equivalent in the gold data set, it is most likely that many of the term's selected targets will also exists in the gold data set, thereby producing relatively high values of precision. This means that most of the targets

selected for the predicted technical terms are relevant in the context of those terms. It is interesting to see that all the other methods achieve better precision for term target selection compared to RTS, which is the baseline method with the lowest precision of 0.096. (See table 6.7) The improvement in target selection means that the suggested criteria for disambiguation are useful for providing relevant targets for technical terms.

The term target criterion with the best precision of 0.409 is FTS. FTS selects a term target based on how frequently the term refers to that target. FTS also turns out to have the highest values of *recall*, *f-score*, and *DA*, which shows that it is the best criterion in selecting relevant targets for technical terms. The *DA* result for FTS tell us that this method has an accuracy of 92% in predicting terms with exactly the same target as in the gold data set. High *DA* scores achieved by all the term target selection criteria show that these methods perform very well at disambiguating the term senses since most of the targets they select for the terms that exist in the gold data are relevant. In general, the target selection criteria disambiguate term senses with accuracy above 60%. Because the baseline criterion also has a high *DA*, we can say that selecting a relevant target for a predicted technical term especially in the event that the term may link to many Wikipedia articles turned out to be a less difficult task than anticipated at the onset of this research. High *DA* scores also mean that appropriate information will be provided for most of the predicted technical terms because the Wikipedia pages which they are linked to will be relevant for their context.

We can compare the performance of DTS and LTS since they have almost the same number of term targets, which have data necessary for their disambiguation criteria. DTS has 3, 221, 668 targets with the necessary disambiguation data while LTS has 3, 165, 298 targets with the necessary disambiguation data. Although the values of their *precision*, *recall*, and *f-score* for selecting relevant targets for technical terms are not too far apart, their *DA* values are quite far apart. Table 6.7 shows that DTS has a *DA* value of 69% while LTS has a *DA* value of 84%. These *DA* scores suggest that LTS is 15% more accurate than DTS in selecting relevant targets for technical terms.

The performance of DTS can also be compared to that of STS because they both predict term targets based on the data-driven approach called word overlap. DTS selects term targets with maximum word overlaps between term definitions and term context while STS selects term targets with maximum word overlaps between the different term senses and term context. It can be argued that DTS performs slightly better than STS at selecting relevant targets for technical terms because it has more targets with data necessary for its disambiguation criterion. While DTS has 3,221,668 targets with the necessary disambiguation data, STS has only 86,158

targets with the necessary disambiguation data. Therefore, DTS performs better by 2.4% at selecting relevant targets for terms. The wide range between 86, 158 and 3, 221, 668 could imply that DTS should perform much better than STS since it has many more data necessary for its disambiguation criteria. But since the improvement is meagre it can be said that in this particular experiment, the word overlap criterion is not the best method for technical term target disambiguation. The statement is further supported by the observation that both DTS and STS improve the baseline method RTS very slightly. i.e. DTS improves the *DA* of RTS by 3.6% while STS improves the *DA* of RTS by 1.2%. An explanation for this low improvement is that most of the time there are few overlapping words between term definition and context and overlapping words between term senses and term context.

The performance of TTS and PTS can be compared because they both depend on pattern-matching. TTS finds term targets which are exactly the same as the term and PTS finds targets which contain the pattern of the term in it. Both methods are case insensitive in that way they are able to find term patterns regardless of their cases. Considering the results in table 6.7 we see that regarding *precision*, *recall*, and *f-score*, TTS performs better than PTS at selecting relevant term targets. The table also shows that TTS is 12% more accurate at selecting relevant targets for technical terms compared to PTS. From the *DA* scores of TTS, it is rather interesting to discover that more than 80% of the time, a term is linked to a Wikipedia article with the same title as the term.

Conclusion

Test results showed that the term prediction algorithm that performed best was FTP. FTP performs best when only candidate n-grams with a term likelihood probability greater than 0 and less than 0.09 are allowed to be predicted as technical terms. Therefore, FTP was used for predicting terms that occur in course documents used by students at universities during the evaluation of TermPedia. The best term likelihood probabilities were considered while applying FTP in TermPedia evaluation. The generally poor precision scores for the technical term prediction algorithms shows that these algorithms predict more terms than those that originally exist in the gold data set. The terms that are newly predicted by the algorithms are not necessarily irrelevant but they are not recognized as technical term in the gold data set. Thus, they lower the precision scores for technical term prediction algorithms.

In the case of disambiguating the sense of technical terms in document context, the test results show that FTS algorithm performed best. Thus, the decision to use it in resolving the meaning of technical terms that occur in course documents read by university students. Generally, all the other algorithms for term target prediction

achieve better precision compared to RTS, which is the baseline method with the lowest precision of 0.096. This shows that suggested criteria for disambiguating term senses in context are useful for providing relevant targets for technical terms. General low precision scores for the term target selection algorithms shows that most of the term targets they resolve do not match those in the gold data set. However, this does not necessarily mean that such term targets are wrong.

Evaluation of TermPedia

Demonstrate works performed
with mildness that comes from

wisdom.

James 3:13 NWT

This chapter provides the results of two user studies which were carried out to evaluate TermPedia, the document enrichment approach for facilitating reading comprehension. The general objective of both user studies was to investigate the usefulness of TermPedia in universities. The study carried out at Gulu University in particular was expected to answer the following questions; 1. Does TermPedia facilitate the comprehension of university students as they read their course documents? 2. Does TermPedia reduce the time needed to acquire knowledge from an educational document? A comparative discussion of the user studies was given to relate the opinions of students at the University of Groningen to those of students at Gulu University, on the idea of document enrichment for facilitating reading comprehension.

7.1 Objectives for Evaluating TermPedia

In chapter 3 it was established that university students have insufficient background and vocabulary knowledge in certain knowledge domains. Therefore, most of their course documents present them with technical terms that they do not understand. If students cannot understand certain terms then their reading comprehension process is hindered. To facilitate their reading comprehension process, the students search for the meaning of difficult terms on the internet or in other reference materials. Searching for the meaning of difficult terms wastes time, costs energy, dulls the reading mood and could derail students from what they are reading. TermPedia was designed based on the need to integrate the meaning of technical terms into course documents used by university students. This would afford the students less time and energy during their reading process. It would also allow them to focus on reading without the need to look outside what they are reading for the meaning of technical terms. The main objectives for evaluating TermPedia were therefore to:

- (i) Find out if integrating the definition of technical terms into course documents used by university students facilitates their reading comprehension process.
- (ii) Find out if providing easy access to definition of technical terms reduces the time needed for knowledge acquisition.

User studies were found appropriate for meeting the objectives of evaluating TermPedia because they allow a researcher to assess the usability and usefulness of an artifact.

7.2 TermPedia User Study at UG

A user study was carried out at the University of Groningen (UG) and particularly, in the Department of Information Science at the Faculty of Arts. This department provides Bachelor's and Master's program of Information Science, which is a branch of science, that deals with the storage, processing and analysis of large amounts of data [Alfa-informatica, 2012]. One of the Bachelor courses offered by the department of Information Science at the University of Groningen is Information Retrieval. We selected students of this course for the TermPedia user study because TermPedia is a practical application of Information Retrieval and also because an electronic copy of the course textbook was available ¹. The user study was specifically carried out for a third year class in which a total number of 13 students participated. TermPedia was used to annotate the electronic textbook by identifying, defining and explaining the difficult terms in it.

In order to present TermPedia to the Information Retrieval students, a webbased user interface was designed and it integrated the annotated textbook into *Nestor*, a digital learning environment of the University of Groningen. A screen shot of the demonstration web-based user interface is displayed in figure 7.1. The user interface allowed the students to read an annotated HTML version of the Information Retrieval textbook with specific emphasis on chapters six to nine. The annotations were hyperlinks automatically generated using technical terms as anchor texts for the (`< a >`) HTML anchor element. URLs of the Wikipedia pages to which the terms are linked formed values of the *href* attribute for `< a >`. For example, in figure 7.1 the generated hyperlink *Information retrieval* was formed using the following HTML text:

```
<a href="https://en.wikipedia.org/wiki/Information_retrieval">
Informaion ritrival</a>
```

¹See, <http://nlp.stanford.edu/IR-book/information-retrieval-book.html>, accessed August, 2011

Each automatically created hyperlink had an added JavaScript functionality, that allowed definitions of technical terms (i.e. the first paragraph of the corresponding Wikipedia page) to show in a pop-up window, as soon as a student moves a mouse over the link (mouse hovering action) as shown in figure 7.1. Definitions are retrieved in real-time from the current version of Wikipedia with an excellent speed. In addition, all generated hyperlinks on the textbook pages were turned into links, which point to the system, so that any new page accessed by the student is also enriched.

Figure 7.1: Screen Sorts of TermPedia User Interface

The students used TermPedia user interface for one month then they answered a study questionnaire that investigated the usefulness and usability of the TermPedia approach. Just before introducing the TermPedia user interface to students, they were given terminology definition exercises in order to prepare them for the possible advantages of the approach in identifying, defining and explaining difficult terms. This part of the user study constituted a manual method of data collection and analysis. An automatic method of data collection and analysis was carried out using Google Analytics, a free web analytics approach offering detailed visitor statistics.

Traffic to TermPedia User Interface

As mentioned above, collection and analysis of data indicating traffic to TermPedia user interface was carried out using Google Analytics. The advantage of using Google Analytics is that detailed visitors' statistics are automatically generated. These statistics ascertain that students used the TermPedia user interface, therefore their responses to the TermPedia study questionnaire could be trusted. From table 7.1 we see that there where a total of 68 visitors during the one month in which the Information Retrieval students interacted with the TermPedia user interface. The table also reveals that 48 out of 68 (i.e. 70.6%) visits to the user interface was from nestor.rug.nl, the server on which TermPedia was integrated into the students' electronic course content. Thus, we can say that students used the user interface with a peak of 12 visits a day during the user study period. (See appendix A.1)

Higher than average bounce rates of 58% and 57% shown in table 7.1 is a positive indication that the students did not have to navigate away from the TermPedia user interface in order to find the definition of predicted technical terms. Google Analytics' bounce rate matrix measures the percentage of site visitors who land on a particular page of a web site and then leave the site without engaging in any way (visiting other pages, clicking hyperlinks or calls to action, filling out forms, etc.) [Dorsey III, 2016]. In this case high bounce rates are a good thing because the TermPedia approach aims to enrich course contents with technical terms definitions so that students do not have to look in reference resources outside their course contents for such information.

Source/Medium	Visits	Pages/ Visit	Avg. Visit Duration	% New Visits	Bounce Rate
nestor.rug.nl/ referral	48	3.52	0:09:07	29.17%	58.33%
(direct)/(none)	14	1.86	0:03:00	35.71%	57.14%
rug.nl/referral	3	3.67	0:01:52	00.00%	00.00%
google/organic	2	4.50	0:02:26	00.00%	00.00%
let.rug.nl/referral	1	3.00	0:00:18	100.00%	00.00%
Total	68	-	-	-	-
Site Average		3.21	0:07:13	29.41%	52.94%

Table 7.1: Summary visits between Oct 19 and Nov 19, 2010 to TermPedia user interface from Google Analytics

Data Collection and Analysis for TermPedia User study

We mentioned above that a questionnaire was used in the manual collection of data during the TermPedia user study. The questionnaire consisted of a quantitative and

a qualitative part. The quantitative part of the questionnaire was divided into three sections A, B, and C. Sections A and B each contained five questions on the TermPedia user interface and the Information Retrieval textbook difficulty respectively. Section C contained eight questions about the usefulness of the TermPedia approach in relation to document enrichment. The qualitative part of the questionnaire had a single section D, which provided an opportunity for the students to express their opinions about TermPedia and give comments and/or suggestions for improving the approach’s techniques and its user interface. In general, the questionnaire investigated the performance of TermPedia’s approach in improving the process of knowledge acquisition from educational materials. The quantitative part of the questionnaire requested ranked responses from 4 to 0 as shown in table 7.2. The ranks imply that high averages are good because then the students agree on the positive attributes of a feature under investigation, for example the extent to which TermPedia provides accurate definitions for difficult terms.

Strongly agree	Agree	Somewhat agree	Disagree	Strongly disagree
4	3	2	1	0

Table 7.2: Ranks of answers to the study questionnaire used at UG

Questions in section B about textbook difficulty concentrated on whether the students found technical terms that they could not easily understand without the help of additional explanation and if these terms hindered their understanding of the textbook content. Section C studied the possibility that TermPedia is a useful approach for easy educational literature comprehension. It included questions on whether the approach was able to identify technical terms from the textbook correctly and accurately provide their definitions and external links to a relevant Wikipedia article. The section also included questions on whether the students were able to easily understand the content of the textbook after accessing the definition of technical terms or if they needed additional explanation of the terms in addition to their definition. Section A (TermPedia user interface) in the quantitative part of the questionnaire had questions on whether it was easy to find information on the user interface, if it was easy to navigate through the interface, and also if the font sizes, colours, and formats were legible. Five (5) questions were selected from each of the three sections in the qualitative part of the questionnaire for analysis of the study results. We received 8 responses to the user study questionnaire out of 13 expected responses from all the students who attended the Information Retrieval class.

Analysis of Quantitative Data from UG TermPedia User Study

Cronbach's alpha reliability coefficient was used to determine if the users' ranks were internally consistent for the responses to questions in the quantitative part to the questionnaire. Alpha was calculated using the following formula:

$$\alpha = \frac{rk}{[1 + (k-1)r]} \quad (7.1)$$

Where: k=number of items considered and r=mean of the inter-items correlation

Cronbach's alpha reliability coefficient normally ranges between 0 and 1. The closer it is to 1 the greater the internal consistency of the items in the scale [Gliem and Gliem, 2003]. Hypothesis for Cronbach's alpha reliability coefficient were set as follows:

- $H_1\alpha = 1$ i.e. Perfect internal consistency of user ranks in this scale.
- $H_0\alpha = 0$ i.e. No internal consistency of user ranks in this scale.

Section	Items	Mean Scores	α	Interpretation
C: (TermPedia usefulness)	5	2.7	0.6	Questionable
B: (Book difficulty)	5	2.4	0.3	Poor
A: (TermPedia user interface)	5	2.5	0.7	Acceptable

Table 7.3: Cronbach's alpha reliability coefficient for RUG questionnaire scores

Table 7.3 indicates alpha values for the three sections of the quantitative part of the questionnaire along with their mean ranks as answered by 8 students who followed the Information Retrieval class. Although the mean rank of the Information Retrieval book difficulty section is slightly above average (i.e. $2.4 > 2.0$), the students' ranks for this section were inconsistent since $\alpha = 0.3$. It can therefore be said that the level to which each student found the Information Retrieval book difficult was subjective to that particular student. An explanation for the students' inconsistency in interpreting the book difficulty could be that they are at different vocabulary levels in the Information Science knowledge domain. Since a total of 8 respondents is rather small, it can be said that $\alpha = 0.6$ is acceptable for internal consistency of students ranks for questions in section C (TermPedia usefulness). Therefore, the mean rank of 2.7 for this section shows that the students tend to consistently agree that the approach is useful for easy comprehension of educational material. The web-based user interface section gave the best results for α with $\alpha = 0.7$ and mean rank of $2.5 \approx 3.0$ indicating that the students unanimously agreed that it was easy to access information from and navigate through TermPedia user interface.

Analysis of Qualitative Data from UG TermPedia User Study

Comments and suggestions given by students of Information Retrieval at University of Groningen (UG) for improving the TermPedia approach can be found in appendix A.4. From the students' comments such as: "*a term like document isn't that relevant*", we learnt that some terms predicted by the TermPedia approach were not useful for the students. This problem is a result of the "noise" in Wikipedia anchor texts that range from simple terms like *document* to complex terms like *k-medoids*. Moreover, depending on knowledge domains and contexts, simple words may denote technical ideas. For example, a word like *truncate* in the field of mathematics and computer science is the term for limiting the number of digits right of the decimal point, by discarding the least significant ones. Much as simple words may appear redundant in some knowledge domains and contexts, it is not always true for all knowledge domains. Comments such as:

"Some definitions are somewhat redundant." and

"It would help if TermPedia linked terms to the correct definitions and Wikipedia articles."

show that TermPedia did not always provide a relevant definition for the terms it predicted. This problem is caused by ambiguity in both the term senses and the definitions retrieved from Wikipedia since a term can have multiple meanings and be linked to more than one Wikipedia article. This therefore directly implies that the TermPedia term sense disambiguation technique needs to be improved. The need to improve TermPedia's term sense disambiguation was supported by a suggestion for the TermPedia approach to "*look for the definitions [of terms] in the IR e-book itself.*" Term definitions retrieved from textbooks where the terms occur would be more relevant to that textbook's context as compared to definitions retrieved from Wikipedia. This would be the right place to start while improving the term sense disambiguation techniques of TermPedia in future. However, it is not always true that technical documents provide definitions of all the difficult terms found in them. We want to make sure that there is an access to the definition of every difficult term in case the definition is not provided in the document content.

The students gave constructive and practical suggestions for improving the TermPedia user interface, for example that the IR book should load on the same website as that of the user interface. This suggestion has been implemented on the current TermPedia user interface website and "*it keeps users from having to switch tabs*". A suggestion was also made to redesign the pop-up window on the user interface so that its size auto grows. This pop-up window is responsible for displaying the term definitions as resolved by TermPedia and currently has a fixed size. Another useful

suggestion was to allow for users to access the definition of terms of their own choice, a function that is lacking in the present TermPedia techniques and user interface but one that could be easily integrated. This function would allow users not to be restricted to the terms predicted by TermPedia.

7.3 TermPedia User Study at Gulu University

To achieve the second objective of evaluating TermPedia, a user study was carried out to verify if TermPedia actually reduces the time for knowledge acquisition. This user study was carried out at Gulu University, a public university located in the north of Uganda. At the university, TermPedia was introduced to a selected category of students. Performances of the students from the selected categories were recorded and analyzed in relation to TermPedia. This analysis provided a measurable value of the approach's usefulness with reference to the research questions in the abstract of this chapter. These questions implied an overall hypothesis that TermPedia is useful to students in higher education institutions, the specific hypotheses implied to guide the study are listed below:

- (i) TermPedia can improve understanding of an educational document by providing contextually relevant information.
- (ii) The time needed to acquire information from an educational document can be reduced through the use of TermPedia.

Answers to the research questions and resolutions about the tested hypotheses are discussed in this section where data collected from the user study was analysed using basic statistics [Freund, 2004].

Selected Category of Students for the User Study

The Department of Computer Science at Gulu University offers a variety of courses among which are Bachelor of Science in Computer Science (BSc.CS), and Bachelor of Information Technology (BIT). BSc.CS enables students to study theories and methods of processing information using computers, the design and operation of computer hardware and software, and the application of computer technology to all fields of knowledge. Similarly, BIT enables students to understand the systems used in digital data handling and retrieval. The course also provides an opportunity for students to develop skills in the techniques necessary to devise, develop and maintain these systems. The objective of BIT is to teach skills applicable to the safe storage of significant amounts of data, easy data modification, and cost-effective information

retrieval. It can therefore be said that both courses have a common interest in developing students who are able to understand and apply the underlying principles of Computer Science and Information Technology to processing information.

Since TermPedia is a document enrichment approach, it provides an opportunity for understanding and applying underlying principles of Information Technology to solving some pedagogy problems in knowledge acquisition and retention. The selected categories of students for the user study included third-year BSc.CS and BIT students of Gulu University in their first semester. These courses consisted of students on day (8:00 a.m. to 5:00 p.m., Monday to Friday) and weekend (7:00 a.m. to 8:00 p.m., Saturday and Sunday) programmes. The user study only involved students following the day programme because they had a relatively relaxed timetable compared to students on the weekend programme. It was assumed that the day students would have ample time to acclimate to the TermPedia user interface and provide reliable results for the study evaluation.

Collection and Analysis of User Study Data at Gulu University

The category of students selected for this user study followed a software engineering course and their primary reading material was a book titled “*Software Engineering*” [Sommerville, 2001]. Both hard and digital copies of lecture slides from this course material were given to the students at the beginning of the semester. It is assumed that the students diligently read the lecture notes in one of the two forms after which they underwent a number of tests (course assessments) according to chapters that had been presented by their lecturer in class. The test results were used as evaluation data for the TermPedia approach. This data was collected before and after the students were introduced to the TermPedia user interface.

Collection and Analysis of Data by Closed-Book Test

Before the students were introduced to TermPedia, they were given a “closed-book” test to find out their level of competence. By closed-book we mean that students were not allowed to reference any textbook, notes, or other knowledge sources while answering the test questions. The test was given on specific sections of the Software Engineering book that the students studied during their course. This test had several questions with multiple-choice answers. After the students had been tested, their results were randomly selected for analysis. The random selection of results was done before post filtering to remove incomplete test results and results from students who did not attempt to answer any test questions. The procedure was taken from Csomai and Mihalcea [2007] who carried out a similar user study evaluation. The mentioned study enriched (or wikified) questions from a quiz for an on-line history course.

On the contrary, TermPedia user study enriches on-line reading material for an off-line course. A summary of the performance of the students after the closed-book test is shown in table 7.4.

	Group <i>A</i>	Group <i>B</i>
Number of students	15	15
Median	15	15
Mean	15.33	15.27
Standard deviation	0.617	0.458

Table 7.4: Summary of students’ closed-book test scores before they used TermPedia

Students’ results were divided into two groups *A* and *B* during the random selection. The summary of these results shows that the students from these two groups were at the same level of competence. Both groups had a mean score of 15 and small standard deviations of 0.478 and 0.617 which shows that the general distribution of students’ scores does not deviate from the mean. The fact that the students were at the same level of competence gave a clear opportunity to notice a change in their performance after they had been introduced to TermPedia. After using TermPedia students in group *A* were given an “open-book” test with the help of TermPedia, while students in group *B* answered the same test without the help of TermPedia.

Data Collected by Open-Book Test

After TermPedia was introduced to the students they were asked to read an electronic copy of chapter six from the Software Engineering textbook, which was the only chapter available with full text in soft copy. This chapter was loaded onto the TermPedia user interface for automatic prediction of technical terms. It was possible to access definitions for the predicted technical terms, as shown in figure 7.2. The predicted terms also created hypertext that linked to Wikipedia articles for additional information. The students read this chapter for a total of eight hours during four sequential lectures in two weeks. After using TermPedia to read chapter six of their course material, they were given an open-book test. By open-book we mean that students were able to consult their course material, notes, and other relevant information during the test session. The test was open-book in the hope that the test results would reveal the efficiency of TermPedia through students’ scores and time required to complete the test. The open-book test consisted of five questions and each question had two parts; (*a*), and (*b*). These questions required short answers where the (*a*) part of each question was a definition question. Question 1 from the test is given below for illustrative purposes.

- Qu. 1(a): What is mean time to failure?
- Qu. 1(b): What nonfunctional requirement of a system does it measure?

	TermPedia Used (A)			TermPedia Not Used (B)		
	Time (Mins.)	Scores	%Scores	Time (Mins.)	Scores	%Scores
Median	48.00	16.00	80.00	43.00	15.0	75.0
Mean	46.87	16.33	81.67	43.87	14.80	74.0
sd(s)	03.98	02.23	11.13	05.46	3.21	16.06

Table 7.5: Summary of students open-book test scores after using TermPedia

Before the open-book test, students were randomly divided into two equal groups of 15. The first group was (A) and students in this group were allowed to consult the course material through the TermPedia user interface only. The the second group was (B) and students in this group could consult any of their reading materials in both hard and digital copies, but they had no access to the TermPedia user interface. If group A performed better than group B, a conclusion could be made that TermPedia improves the process of knowledge acquisition by providing relevant contextual information. One might also conclude that TermPedia reduces the time required for knowledge acquisition if students in group A completed the test earlier than students in group B. Results from the open-book test are shown in table 7.5.

Figure 7.2: Snap shot of TermPedia user interface

Analysis of Students’ Scores from the Open-Book Test

Samples of data collected from the open-book test were investigated for distribution properties through box plots, and quantile-quantile (Q-Q) plots. This was done to test if the samples have a normal distribution in which case, we could then use a t-test for independent samples to investigate the differences in the means of the data. Figure 7.3 displays two box plots that visualize the open-book test scores for students who consulted the course material on TermPedia user interface (A_s) and students who consulted course material from sources not loaded onto TermPedia user interface (B_s).

Figure 7.3: Box plots for open-book test scores of students

The box plots clearly show that sample distributions of A_s and B_s are similar since their inter-quartile ranges overlap. Although this is true, sample A_s has a greater central tendency compared to B_s because the median of A_s is closer to the center of its inter-quartile range, revealing a distribution that is neither sparse nor skewed. A_s box plot also has no outliers, which shows that the data sample is likely

to follow a normal population distribution. The median of box plot for sample data B_s is below the center of its inter-quartile range indicating that the sample may be skewed to the left. Outliers further affect the skewness of sample B_s . Since both sample distributions are not wildly scattered, a Q-Q plot was used to check that sample fit a normal distribution. A p -value=0.5579 from a Shapiro–Wilk test which tests the null hypothesis that a sample comes from a normally distributed population, for test scores (A_s) from students who were helped by TermPedia during the open book test indicated that these scores were normally distributed. On the contrary, the same could not be said about the test scores from students who were not helped by TermPedia during the open book-test. Data from this sample produced a slightly skewed distribution since they do not closely follow the diagonal line of a normal Q-Q plot as shown in figure 7.4.

Figure 7.4: Normal Q-Q plot of open-book test scores for students who were not helped by TermPedia.

The distribution of these data points was greatly affected by the two outliers as earlier mentioned and this causes a non-linear pattern in the plotted data points. In spite of this distribution dispersion, the majority of the data points in the interquartile range lie relatively close to the diagonal line. We therefore calculated a p -value for the Shapiro–Wilk test in order to find out if the data from B_s came from a normally distributed population. The p -value given by this test was 0.01 and it indicated that there was significant evidence to show that B_s did not come from a normally distributed population. With one normally distributed sample A_s and one skewed sample B_s , we carried out a Mann-Whitney-Wilcoxon (MWW) test to see if scores from A_s were significantly higher than scores from B_s . The MWW test gave a p -value=0.12 indicating that there was insignificant evidence to show that students who got help from TermPedia during the open-book test performed better than the students who did not get help from TermPedia during the same test.

It is accepted practice to exclude outliers [Peeck, 1974] in order to focus on the central tendencies of the data because the presence of outliers can lead to inflated error rates and substantial distortions of parameter and statistic estimates when using either parametric or non parametric tests [Osborne and Overbay, 2004]. Therefore, when we ignored one of the outliers in B_s and subjected the sample to another Shapiro-Wilk test, we obtained a p value= 0.184 > 0.05. This p -value showed significant evidence to assume that B_s comes from a normally distribution population. With the outliers ignored, we could therefore assume that both population samples A_s and B_s tend to a normal distribution. A t -test for independent samples was then used to investigate if the mean score for students who got help from TermPedia during the open-book test ($\bar{A}_s$) was significantly greater than the mean score for students who did not get any help from TermPedia during the test ($\bar{B}_s$). The t -test gave a p -value= 0.139 > 0.05 confirming that there is insignificant evidence to show that ($\bar{A}_s$) is greater ($\bar{B}_s$) as indicated by the MWW test above. We therefore accepted the null hypothesis that the means are equal meaning that there is insufficient evidence to show that TermPedia improved the students’ scores for the open-book test.

	Group A	Group B
Number of students	15	14
Median	16.00	15.50
Mean	16.33	15.43
Standard deviation	2.23	2.17

Table 7.6: Summary of students open-book scores with outliers ignored

The mean score ($\bar{A}_s = 16.33$) is greater than the mean score ($\bar{B}_s = 15.43$) by only 0.9, see table 7.6. A value too small to prove that TermPedia can improve the

process of knowledge acquisition and improve understanding of education documents by providing contextually relevant information, in accordance with the t -test results. However, the sample of 15 students from each group is too small to give substantial evidence of the different performances and clearly show the impact of TermPedia on the process of knowledge acquisition. We believe that a larger sample would evidently show that TermPedia has a positive effect on knowledge acquisition by improving students' performance. In addition, we noticed that the difference in the means is more than $0.5sd$ if the open-book test scores are used as obtained and more than $0.25sd$ if the outlier is removed. This indicates potential ability of TermPedia to improve the process of knowledge acquisition, a notion that could be proved with a refined experiment on a larger sample.

Analysis of Time Spent by Students to Complete the Open Book Test

It was also important to look at a box plots for time spent by students in groups A and B to complete the open book text. It is worth mentioning that this data is not reliable since most of the students waited until the last minute to turn in their answer sheets regardless of whether they had completed the test earlier.

User group	Median	Mean	S.D.
A_t	48.00	47.00	3.98
B_t	43.00	43.87	5.46

Table 7.7: Table showing summary statistics of time spent by students on open-book test

Summary statistics in table 7.7 clearly shows that students who got help from TermPedia during the open book test spent more time than students who did not get help from TermPedia during the test. This result is contrary to the hypothesis that TermPedia would reduce the time necessary for students to acquire information from education documents. A likely explanation for this result could be that the students who used TermPedia during the test were more occupied and looked up a lot of information, hence the slight improvement in their performance as discussed above.

Gulu University User Study Questionnaire Data Collection

Another evaluation of the TermPedia approach was realized by using a questionnaire, which can be found in appendix A.6. Through the questionnaire we tried to investigate if TermPedia has a user-friendly interface and also if the course

material of the students had difficult terms which hindered their process of knowledge acquisition. More importantly we investigate if document enrichment by TermPedia helps the process of knowledge acquisition and reduces the time thereof, through technical term prediction, definition and linking the terms to relevant Wikipedia articles. The questionnaire was divided into four sections A, B, C, and D of which sections A to C were quantitative and section D qualitative. Sections A, and B both had 5 questions relating to TermPedia user interface and document difficulty respectively. Section C had 10 questions relating to TermPedia techniques. Ranked scores for the quantitative part of the questionnaire were as follows, indicating that low averages are good, because then the students tend to agree in their opinion:

Strongly agree	Agree	Somewhat agree	Disagree	Not applicable
1	2	3	4	5

Table 7.8: Ranks for Gulu University questionnaire responses

Section D allowed students to write down short comments and/or suggestions on the TermPedia user interface and techniques. (See appendix A.8) A random number of 22 students from the selected category of users answered the questionnaire independently, providing ranked scores of opinions they formed on the three themes as mentioned above. Table 7.9 displays a summary of scores as ranked by the 22 students for all the questions in sections A, B, and C of the questionnaire without missing data.

Questionnaire section	Mean of average scores		
	Median	Mean	SD
A: TermPedia user interface	2.000	1.973	0.791
B: Book difficulty	3.400	3.100	0.976
C: TermPedia techniques	1.850	2.073	0.907

Table 7.9: Summary scores for all the 22 students as ranked for sections A, B, and C

Analysis of GU Data Collected by the Questionnaire

With the assumption that scores given by Gulu University (GU) students while answering the questionnaire are normally distributed, we can make a few remarks about the data in table 7.9 by considering the average means and standard deviations. For instance, responses to section A, which focuses on the TermPedia user interface, presents a low average mean score of $1.973 \approx 2$ and an average standard deviation of 0.79. This shows that although most of the students tend to form a good opinion

about TermPedia's user interface, the difference in some of their opinions is rather wide because more than 16% of the students' average mean score lies in an interquartile (IQ) of scores greater than 2. The table also shows that responses to section B which examines book difficulty, had a high average mean score and standard deviation of $3.1 \approx 3$ and 0.98 respectively. This indicates that the students disagreed that chapter six of the course material which they read with the help of TermPedia was a difficult chapter. In other words, the students thought that the technical terms found in this chapter were not too difficult to understand and therefore did not affect their process of knowledge acquisition. Responses to section C that dealt with TermPedia techniques also had a relatively low average mean score of 2.07 and an average standard deviation of 0.907 indicating that 68% of the students' average mean scores to this section lies between 2.07 ± 0.907 . The positive boundary of $2.977 \approx 3$ leaves more than 16% of the average scores above 2 which is an ideal score for showing that most of the students formed a good opinion about the usefulness of TermPedia techniques in this case. A detailed summary of these scores can be found in appendix A.7.

To dive deeper into the analysis of results from the questionnaire we consider each section separately since they discuss different aspects that sum up to investigate the requirement and usefulness of TermPedia. The main goal for this analysis was to investigate the performance of TermPedia in document enrichment and find out whether it is a necessary aid to knowledge acquisition. Another goal for this analysis was to find out if the book chapter given to the students to read by the help of TermPedia had many technical terms, which hindered their understanding of the content. If this was the case then we hoped that they would need TermPedia to help them to understand the chapter content given that TermPedia performs well at providing relevant information for the technical terms. A final goal of the results analysis was to find out if the TermPedia user interface is user friendly. This is to say, the students could easily navigate through and find information on the user interface and that the user interface was legible and simply designed. To carry out the results analysis, we first calculated correlation coefficients (r) for scores to questions of each section in order to find out which questions may be considered together during the analysis discussion. A Chronbach's alpha (α) was then calculated for questions with high positive values of r . α values are considered because they provide a measurement of the internal consistency or reliability of scores given by the 22 students. Therefore, the preliminary condition for determining if scores to different questions in the same section could be analysed together was that they must have a high value of both r and α . Questions that did not highly correlate were discussed separately. Details about r and α for each section can be found in appendix A.3. We first analysed section C of the questionnaire because it was tailored to investigate the performance and requirement of TermPedia, which were the main goals of this analysis.

Analysis of Section C of Gulu User Study Questionnaire: TermPedia Techniques

Questions in this section were tailored to investigate the performance and benefits of our document techniques. For explanations of the document enrichment techniques see chapter 4. Document enrichment involved the use of TermPedia to predict difficult terms in chapter six of the students' course material and provide the definition of these terms. The predicted terms were then linked to relevant Wikipedia articles that provide additional enrichment of the course material. The main objective of this section was therefore to find out how well TermPedia performed in predicting and defining terms, and linking them to relevant Wikipedia pages. GC1 to GC10 referred to mean scores of the 22 students to questions 1 to 10 in this section. After calculating r values for these scores, it was revealed that GC3, GC4, GC5, GC8, and GC9, which are mean scores to questions 3, 4, 5, 8, and 9 in this section had high positive correlations, see table 7.10 for references of the mean scores. The highest r value among these mean scores was 0.7 between GC5 and GC9. A high α value of 0.8 for the scores indicated an internal consistency by the students while ranking their opinions about TermPedia techniques related to these questions. For these reasons this set of mean scores were analysed together.

Questions	Median	Mean	S.D.
GC3: TermPedia gave relevant term definitions	2.00	1.727	0.631
GC4: Difficult term understood through definitions	2.00	1.952	1.024
GC5: Definitions helpful for understanding chapter	2.00	1.905	0.831
GC8: Correct definitions predicted	2.00	2.333	0.577
GC9: TermPedia helpful for e-learning	1.00	1.455	0.671
Avg. of Means		1.874	0.747

Table 7.10: Summary for user scores to questions 3, 5, 8, and 9 in section C

In table 7.10 we see that the average of mean scores to questions 3, 4, 5, 8, and 9 is $1.874 \approx 2$. This average leads us to the interpretation that the students agreed positively about the techniques of TermPedia in relation to these statements:

- TermPedia always gave relevant definition of terms,
- It was possible to understand difficult terms by their definitions,
- The term definitions helped in understanding the book chapter,
- TermPedia always predicted correct difficult terms, and that

- TermPedia can be a helpful approach for e-learning.

We see that most of the students either strongly agreed or agreed that TermPedia was good at predicting technical terms and the term definitions given were helpful in understanding the book chapter. We also see that the students found most of the predicted terms relevant and that TermPedia could be a helpful e-learning approach.

Questions	Median	Mean	S.D.
GC6: Terms linked to relevant Wikipedia articles	2.00	2.381	0.590
GC7: Wikipedia articles improved understanding	2.00	2.286	0.717
Avg. of Means		2.334	0.653

Table 7.11: Summary for user scores to questions 6 and 7 in section C

Another set of mean scores that we considered together were GC6, and GC7, which had a relatively high r value of 0.53. An α value of $0.685 \approx 0.7$ for these means was also high enough to show internal consistency in the scores the students gave to questions 6 and 7 of section C. A relatively low average of $2.334 \approx 2$ for the mean scores of this set indicated that the students agreed that TermPedia linked predicted terms to relevant Wikipedia articles, see table 7.11. This low average mean score also indicated that understanding the book chapter was improved by reading widely from linked Wikipedia articles.

Questions	Median	Mean	S.D.
GC1: A link for each difficult term	1.50	1.727	0.883
GC2: TermPedia finds term definition in a short time	2.00	1.810	0.928
GC10: Illustrate terms by graphics	2.00	2.238	0.831

Table 7.12: Summary for user scores to questions 1, 2, and 10 in section C

The set of mean scores with low r values were GC1, GC2, and GC10 as shown in table 7.12. Since these mean scores did not correlate highly with any of the other mean scores in section C of the questionnaire, we discussed them individually. For instance, the low mean score of $1.727 \approx 2$ showed that the students agreed that TermPedia provided a link for all predicted terms in chapter six of their course textbook. The mean score of $1.810 \approx 2$ for question 2 indicated that the students agreed that it took a short time to find term definitions by using TermPedia. If this is the case then it can be said that TermPedia reduced the time necessary for the students to find additional information. Finally, a mean of $2.238 \approx 2$ pertaining to average scores of question 10 in this section showed that the students agreed that it would help to illustrate difficult terms with graphics. This indication supports the interest of illustrating technical

terms found in a variety of reading material with graphics from Wikipedia info-boxes, a concept that was under consideration during this user study.

Generally, because of the low mean scores to questions 1, 3, and 9 in section C, we can say that the students were pleased with the fact that TermPedia provided relevant definitions for technical terms, and provided a link to a Wikipedia article for these technical terms. The students were also pleased by the idea of using TermPedia as an e-learning approach. However, the students were not satisfied because TermPedia did not always predicted correct technical terms (question 8, section C). This indicates that sometimes TermPedia left out important technical terms and that not all the technical terms predicted by the approach were actually considered difficult by the students. The variation in opinions about the difficulty of terms may occur because of different background knowledge levels the students had in the field of Software Engineering.

Analysis of Section B of Gulu User Study Questionnaire: Document Difficulty

The document referred to here is chapter six of the book that the students used for their course material, see section 7.3. GB1 to GB5 refer to mean scores given by the 22 students to questions 1 to 5 in section B of the study questionnaire. Correlation coefficient values (r) of mean scores GB2 to GB5 were high indicating an existing linear relationship among the average scores to these questions. We discuss these average scores together since they also had a high α value of 0.844. For instance, the highest value of r was 0.441 between GB2 (unfamiliar terms in document) and GB4 (document above reader intellect level) and the α for these scores was 0.411 a value too low to indicate internal consistency of the mean scores. It is possible that the students' scores do not correlate neither are they consistent because the questions in this section investigated different attributes of document difficulty. For example, the lowest r value of 0.041 was between GB1 and GB4, where question 1 investigated if the book chapter contained technical terms and question 4 investigated if the chapter was written above the students' intellectual level.

In table 7.13 GB2 to GB5 refer to the mean of scores given by the 22 students to questions 2 to 5 in section B of the Gulu University user study questionnaire. These scores had high correlation coefficient values (r) indicating that there exists some kind of linear relationship between them. The highest r value of 0.68 was between GB2 (unfamiliar terms in document) and GB4 (document above readers' intellect level), see appendix A.3. This high correlation showed that students tended to disagree that the book chapter had unfamiliar terms and also that the book was written

Questions	Median	Mean	S.D.
GB2: Unfamiliar terms in document	3.00	3.095	1.044
GB3: Understand text around unfamiliar terms	4.00	3.227	1.020
GB4: Document above reader's intellect level	4.00	3.200	1.056
GB5: Bad and boring document writer	4.00	3.579	0.902
Average of Means		3.275	1.006

Table 7.13: Summary for user scores to questions in section B of Gulu user study questionnaire

above their level of intellect. Altogether GB2 to GB5 had a high Cronbach's alpha coefficient of 0.844 prompting us to discuss the resultant average scores together. The average mean of $3.275 \approx 3$ as indicated in table 7.13 indicate that the students tend to disagree on the following statements in section B of the user study questionnaire:

- The technical terms were unfamiliar and not easy to understand.
- It was not easy to understand text around difficult terms.
- Material in the e-book chapter was written above my intellectual level.
- The e-book writer was bad at explaining things and boring.

The highest mean score of $3.579 \approx 4$ for GB5 shows that the students found the book chapter well written and interestingly explained. GB3 had a mean of $3.227 \approx 3$ indicating that the students tend to disagree that it was not easy to understand text around difficult terms. If this is the case then it means that the difficult terms did not hinder their process of learning.

The only mean score that did not correlate with the other mean scores was GB1 which represents mean scores to question of this section by the 22 students, see appendix A.3. GB1 had a value of $1.955 \approx 2$ indicating that the students agreed that there were technical terms in the e-book chapter.

In summary we may say that the students agreed that the book chapter contained technical terms and that the chapter was relatively easy to read and understand. This came as no surprise since the user study dealt with students in their third year. It was revealed that these students were familiar with the subject area and therefore may not have had problems understanding technical documents in their field of study.

Analyzing Section A of Gulu User Study Questionnaire: TermPedia User Interface

Correlation coefficient for mean scores of the 22 students to the 5 questions in section A shows that only questions 1 (easy to locate information on the interface) and 5 (enjoy using the interface) have a high r value of 0.556, the rest do not. Therefore, we calculated Cronbach's alpha for these two questions to investigate internal consistency of scores by the students and obtained an α value of 0.7. This α indicated internal consistency estimate of reliability of the scores given by the 22 students to questions 1, and 5 in Section A. We could therefore combine these two mean scores in discussing a summary of their basic statistic results as indicated in table 7.14.

Questions	Median	Mean	S.D.
GA1: Locate information on interface	2.00	2.00	0.69
GA5: Enjoy using the interface	2.00	1.77	0.87
Average of Means	2.00	1.89	0.78

Table 7.14: Table showing summary statistics of scores given to questions GA1 and GA5

As such, an average mean score of $1.89 \approx 2$ for questions in figure 7.5 clearly shows that most of the students gave a response of either 1 or 2 for 1 and 5 shows that majority of the students found it easy to locate information on TermPedia user interface and that they agreed that the interface was enjoyable to use. The bar plots in questions 1 and 5 in section A of the user study questionnaire. This confirms that most of the students found it easy to locate information on the TermPedia user interface and also that they enjoyed using it.

Responses to questions 2, 3, and 4 did not have high correlation coefficients (r) therefore we analysed these responses separately. There is a uniform median of 2 for the mean scores to questions 2, 3, and 4 as shown in table 7.15.

Questions.	Median	Mean	S.D.
GA2: Navigate through interface	2.000	2.000	0.535
GA3: Colours and design of interface	2.000	2.227	0.869
GA4: Legibility of text on interface	2.000	1.714	0.717

Table 7.15: Table showing summary statistics of mean scores for questions 2, 3, and 4 in section A of Gulu user study questionnaire

The mean of 2 and standard deviation of 0.535 shows that the students' scores to question 2 (navigation through interface) hardly vary. This explains the low

Figure 7.5: Bar plot for frequencies of ranked responses to questions 1, and 5 in section A of the user study questionnaire

correlation of mean scores to question 2 with mean scores to other questions in this section. The general low r values among U2, U3, and U4 was influenced by the fact that questions 2, 3, and 4 investigated different aspects of TermPedia’s user interface. For example, question 2 investigated if it was easy to navigate through the interface while question 3 investigated if the students found the colours and design on the interface appealing.

Considering means and standard deviations, we can say that question 2 has a low average score of 2 from all the 22 students. This implies that most of the students agreed that it was easy to navigate through TermPedia user interface. GA4 (legibility of text on the interface) could have had a better performance but the standard deviation of 0.99 is rather high for a mean score of 1.864. Thus, although the mean score indicates that most of the students agreed that the font size, and colour of text on

TermPedia user interface was easy to read, their opinions vary from strongly agree (0.874) to disagree (2.854).

Analysis of Section D of User Study Questionnaire: Comments and suggestions

As mentioned in subsection 7.3, section D of the user study questionnaire was qualitative and it gave an opportunity for the students to give comments and/or suggestions regarding the TermPedia user interface and techniques. In the analysis of this qualitative data, we seek to provide understanding from the respondent's perspective [Taylor-Powell and Renner, 2003] about the benefits and weakness of TermPedia. Since we are more interested in the techniques of TermPedia, i.e. predicting and defining technical terms, and linking the terms to relevant Wikipedia articles, we first discuss the comments and/or suggestions given by the students on this aspect. For details on the comments and suggestions please see appendix A.8. We learnt that there is a lack in the coverage of the definition of technical terms retrieved from Wikipedia. This was evident from the suggestion such as, "link technical terms to Wikipedia and other resources." This problem may be caused by the term sense disambiguation technique of TermPedia in that, this technique does not always link a predicted term to a contextually relevant Wikipedia page, revealed by a suggestion such as "give relevant definition of terms." It is therefore necessary to improve the term sense disambiguation technique of TermPedia and consider the use of other knowledge bases like web dictionaries and glossaries. We also learnt that TermPedia did not define all the technical terms it predicted from the comment that, "some difficult terms were not explained." This problem could have been caused by Wikipedia disambiguation pages and a variation in the structure of the first paragraph of Wikipedia articles. We therefore have to modify the term definition technique to exclude all Wikipedia disambiguation pages and provide a way to extract only definition information from the first paragraphs of Wikipedia articles. A slow network problem, which obviously does not stem from techniques, was also revealed by the comment, "network was slow and not interactive." This problem is mentioned because it influenced the user study since the students relied on the local area network (LAN) to be able to use TermPedia. Besides these low sides of TermPedia techniques we got some positive comments to show that TermPedia could be necessary and useful in an educational setting. For example, "useful technique for e-Learning. It saves time and other resources".

From the comments and suggestions on the TermPedia user interface, we learnt that although it was easy to navigate and find information, the students saw a lack of graphics and animation. For example, the suggestions, "add graphical illustrations,"

and “more animations would be appealing”. Apparently, some students also thought the interface needed more colour, as shown by the comment “improve interface colour.” As much as these suggestions may be useful, the current focus of TermPedia is on document enrichment and the simple user interface was designed to present the approach to users and enable a user study evaluation of its techniques. Generally, there was a need to improve both TermPedia techniques and the user interface.

Challenges During the User Study

In this section we shall discuss challenges that were encountered during TermPedia user study at Gulu University, the biggest of which was lack of free time in computer laboratories. For this reason, students only accessed TermPedia user interface during their lecture session which time was inadequate for reading the entire chapter six of their course material on the interface. The fact that only this chapter was available in full-text as a soft copy also limited the time the students used TermPedia for thorough reading. Unreliable local area network (LAN) at Gulu University limited access to TermPedia user interface by the students, because on days when the LAN was not functional, they could not use the interface. As the TermPedia user interface was installed on a local server at the University and not on the Internet, it was impossible for the students to access the interface outside of the University LAN. To be able to install TermPedia on the local server a database for a Wikipedia dump was created and indexed as used in Information Retrieval [Manning et al., 2009] so that the approach did not have to depend on an Internet connection. This sacrifice of computer storage space for mirroring Wikipedia was worthwhile because Internet service at Gulu University is quite unreliable. Unfortunately, this sacrifice overlooks the need to use current Wikipedia articles that may contain important corrections, amendments, and references. Using an older version of Wikipedia may affect the quality of information given by TermPedia during its document enrichment process.

Conclusion

In conclusion, the user study did not provide significant proof to indicate that TermPedia can improve understanding of an educational document by providing contextually relevant information. This conclusion was reached in section 7.3 where the *t*-test of two independent samples provided an indication that students who took the open-book test with the help of TermPedia performed as well as students who took the test without any help from TermPedia. Similarly the alternative hypothesis that time needed to acquire information from an educational document can be reduced through the use of TermPedia was rejected. Although the results are not

in favour of TermPedia, we believe that the sample size was rather small and that the time the students spent using TermPedia was reduced because of the technical difficulties at Gulu University. Happily, the questionnaire gave good results on TermPedia user interface and TermPedia techniques as the average mean for the scores in sections A and C of the questionnaire were not above 3. We can therefore say that although the students found it easy to read chapter six of their course material, they also thought that TermPedia was useful for improving the process of knowledge acquisition.

7.4 Comparative Discussion of Evaluation Results

In this section we present a comparison of the user study and user study findings as discussed in subsections 7.2 and 7.3. We compare these results because they indicate the users opinions on the performance of TermPedia in document enriched. The questionnaire was also the only method used in both the user study and study to get an insight of what the intended users of TermPedia feel about the approach. The comparative discussion will look at similarities and difference in the opinions of users from the University of Groningen, located in a developed country to that of users from Gulu University, located in a developing country. Students from both universities provided opinions which suggest that TermPedia has a good potential of improving the process of knowledge acquisition despite its weakness in term prediction, definition, and linking the predicted terms to relevant Wikipedia pages. Concerning book difficulty, the students did not find their course materials difficult to understand, this could be because in both cases we used students in their third year of study and therefore they were familiar with the terms in their fields of study. The students also had no problem with accessing information from TermPedia user interface and some of them actually liked the idea. Despite these similarities, the students also had differences in their opinions concerning the improvement of TermPedia techniques and user interface.

A student from the University of Groningen provided a useful suggestion for improving the term definition technique of TermPedia by retrieving the definitions from the document in question, while a student from Gulu University suggested retrieving the definitions from web dictionaries as a solution for the same. These are very good suggestions and it shows that some of the students took a keen interest in the idea of document enrichment for facilitating the process of knowledge acquisition. As a suggestion for improving the TermPedia user interface, some student from Gulu University suggested the inclusion of graphics and links to down-loadable electronic books. On the contrary some student from the University of Groningen thought that including a term search function to look for the definition of specific words would improve the user interface. The variations in these opinions indicate the user need

for a approach like TermPedia. For example, some students in Gulu would require the approach to provide electronic course materials perhaps because they do not have access to these due of various reasons including poverty. This need is different from that of some of the students at the University of Groningen who would like to be a little independent on deciding which terms they would like TermPedia to define. In general, although some of the students' opinions differ, they indicate that TermPedia would be a good approach for knowledge acquisition.

Chapter 8

Epilogue

The knowledge of word meanings and the ability to access the knowledge efficiently are recognized as an essential factor in reading comprehension

H.M. Sidek and H. Ab. Rahim

In this chapter, we reflect on the thesis and provide conclusions on the results obtained through the research on facilitating the process of reading comprehension. The chapter discusses contributions made by our research to the knowledge and practice of facilitating reading comprehension through document enrichment. The generalizability of the document enrichment approach for facilitating reading comprehension is also discussed. The chapter further outlines recommendations for the document enrichment approach and suggests future research in the direction of how such recommendations can be implemented.

8.1 Thesis Overview

This research was motivated by the reading comprehension needs of university students who normally have a lot of time-bound reading tasks. Some of the tasks include preparing for lectures, presentations, assignments and examinations. Consequently, students need to read and quickly understand their course content so as to successfully accomplish their reading tasks. However, during the literature review and stakeholders survey, it was established that university students have challenges in understanding their course content. The literature survey carried out in chapter 1 indicated that one major challenge was inadequate reading comprehension skills. This challenge manifests itself as insufficient vocabulary and background knowledge in specific subject domains. We referred to vocabularies that are specific to certain subject domains as technical terms. We also took into account that words in such vocabularies are ambiguous because these words have different meanings in specific

knowledge domains. Insufficient knowledge of technical terms makes course content difficult for students because they can neither recognize terms for which they do not have prior knowledge, nor understand what such terms mean. Thus, unfamiliar technical terms interfere with the process of deriving meaning from what the students read. This in turn demoralizes and frustrates the students as indicated in the stakeholders survey reported in chapter 3 of this thesis. The stakeholders' survey also indicated that students look for the meaning of unfamiliar technical terms from sources outside their course contents, in order to fill knowledge gaps caused by such terms. Doing so distracts the students away from their course contents and demands from them extra time and effort.

Therefore, the key research question was “*How can reading comprehension be facilitated among university students without distracting the student from the contents of course documents?*”

Design science research was identified as an appropriate methodology for investigating a way to facilitate reading comprehension without distracting from the content of course documents. Design science research provided the tools for understanding the reading process and identifying factors that hinder reading comprehension. It also helped in identifying ways of addressing the factors that hinder reading comprehension, which provided functional requirements for a design artifact. The functional requirements pointed to a document enrichment approach as an artifact for addressing factors that affect reading comprehension without distracting from document contents. The document enrichment approach modified text wikification techniques so that knowledge could be seamlessly integrated into document contents. Specific research questions that helped to investigate the problem in reading comprehension among university students and assist to devise a design artifact that facilitates the grounded problem were:

Question 1: What are the key processes involved in reading comprehension?

Identifying the key processes involved in reading comprehension was vital in facilitating the reading comprehension process. In chapter 1 various models were used to explain the key processes involved in reading comprehension. A key process identified by the top-down model was the ability for people to relate what they read to what they already know. This process emphasized the importance of prior knowledge in reading comprehension. Vocabulary knowledge was identified as a major indicator of prior knowledge. Therefore, readers with sufficient vocabulary knowledge can easily comprehend what they read.

Another key process in reading comprehension identified by the situations models is that readers construct mental images from the text they read. For readers to easily

create these mental images, authors need to provide sufficient clues to text meaning. The reading comprehension models showed that sufficient background knowledge, sufficient vocabulary knowledge and sufficient clues to text meaning facilitate the process of reading comprehension.

Question 2: What are the key factors that hinder the reading comprehension processes?

In chapter 3, a stakeholders' survey was carried out to understand which factors hinder the reading comprehension process of university students in particular. The stakeholders' survey results agreed with the literature review carried out in chapter 1. Most importantly, the results indicated that if students have insufficient background and vocabulary knowledge in specific subject domains, then it is difficult for them to understand documents containing those subjects. The students indicated that reading difficult course contents consumes time and energy. They also showed that unfamiliar subject-specific vocabularies or technical terms can lead to a loss of interest in reading and may cause students to misinterpret their course content. (See section 3.4) The students' suggestions for facilitating their reading comprehension process pointed to a document enrichment approach. They indicated that providing easy access to the correct definition of technical terms that occur in their course content could facilitate their reading comprehension process. The students also indicated that providing pictorial illustrations for technical terms and easy access to background knowledge on difficult concepts would facilitate their reading comprehension process.

TermPedia, a document enrichment approach was suggested and implemented as a means of keeping students from looking for the meaning of technical terms that they do not understand from sources outside their course contents. The idea of document enrichment promotes integrating contextually relevant information into the contents of electronic documents. Based on the indication that technical terms hinder the reading comprehension process because of their obscurity, TermPedia sought to provide definitions and background information for such terms. This was partly an initiative from the stakeholders' view on facilitating the reading comprehension process. Easy access to technical terms definitions promotes knowledge acquisition. To promote knowledge retention, TermPedia provides contextually relevant graphic support for the technical terms whenever available.

Question 3: What are the key tasks in a document enrichment approach for facilitating reading comprehension?

This question was vital for investigating the necessary functions of a document enrichment approach that allows it to facilitate reading comprehension. As already

mentioned, our idea of document enrichment involves integrating contextually relevant information about technical terms into document content. The literature review carried out in chapter 2 indicated that this idea would lead to a complex venture because computers do not process world knowledge but rely on algorithms and electronic data to carry out document enrichment. Integrating relevant information about technical terms into document contents therefore required a combination of several techniques from human language technologies. The key tasks in a document enrichment approach for facilitating reading comprehension were thus identified as follows:

- (i) Predicting difficult terms that occur in document contents
- (ii) Resolving what the predicted terms mean in document context
- (iii) Providing easy access to context-relevant definitions for the predicted terms
- (iv) Providing easy access to background information for the predicted terms
- (v) Providing graphical support for the predicted technical terms whenever possible.

In chapter 4, document enrichment algorithms were designed for TermPedia, an approach that accomplishes the tasks identified in (i) and (v) above. Human language technology techniques and data from a Wikipedia dump were used in designing TermPedia's document enrichment algorithms. The algorithms were designed to provide services for predicting technical terms, resolving the meaning of technical terms in document context (sense disambiguation), providing context-relevant definitions of technical terms, providing background information for technical terms and automatically generating hypertexts. Hypertexts served to integrate relevant information about technical terms into document contents. We experimented with various criteria for technical terms prediction and technical terms sense disambiguation. The best criteria for predicting technical terms and disambiguating their senses in context were based on frequencies. On the one hand, the criteria for technical terms prediction considered how often a term was used for creating links in the Wikipedia dump as an indication of potential technical terms. On the other hand, the criteria for disambiguating the sense of technical terms in context considered the technical terms' most frequent sense as used in the Wikipedia dump. The document enrichment algorithms showed that it is possible to use Wikipedia data for predicting technical terms, resolving their meaning in context and integrating contextually relevant information for such terms into document content. Integrating contextually relevant information about technical terms into document contents facilitates reading comprehension because it helps students to have easy access to the meaning of terms that they do not know.

Therefore, students are able to derive accurate meaning from their course contents without extra time and effort needed to search for the meaning of such terms from other sources.

Question 4: What is the performance of TermPedia's algorithms?

TermPedia's algorithms were tested to establish if they could be effectively used to enrich document contents with contextually relevant information about technical terms. A specific goal for this test was to find out the most effective algorithms and use them in developing a web-based user interface for evaluating TermPedia. Methods for testing information extraction systems were found suitable in testing the performance of TermPedia's document enrichment algorithms because document enrichment is the application and implementation of information extraction technologies. Such methods include precision and recall.

Test results showed that the frequency based technical term prediction (FTP) algorithm performed best at predicting technical terms in document content. Therefore, FTP was used for predicting terms that occur in course documents used by students at universities during the evaluation of TermPedia. High recall scores (80%) for the term prediction algorithms show that these algorithms predict most of the terms that occur in the gold data set. Therefore if students want to look up the meaning of particular technical terms, it is likely that a link for this purpose will be available in the enriched document.

General poor precision scores for the technical term prediction algorithms show that these algorithms predict more terms than those originally exist in the gold data sets. The terms that are newly predicted by the algorithms are not necessarily irrelevant but they are not recognized as technical terms in the gold data sets. Thus, they lower the precision scores for technical term prediction algorithms. In the case of resolving the meaning of technical terms in document context, the test results show that the frequency based target selection (FTS) algorithm performed best. Thus, we decided to use it in resolving the meaning of technical terms that occur in course documents read by university students. Generally, all the other algorithms for term target prediction achieve better precision compared to the random target selection (RTS) algorithm, which was the baseline method. This shows that the suggested criteria for resolving the meaning of terms in document context have the potential of providing relevant targets for technical terms. General low precision scores for the term target selection algorithms show that most of the targets the terms resolve to do not match those in the gold data set. However, this does not necessarily mean that such term targets are wrong. Details of the methods and data used for testing the performance of TermPedia's document enrichment algorithms are discussed in chapter 6 of this thesis.

Question 5: What is the perceived usefulness and usability of the document enrichment approach?

A user evaluation of TermPedia was carried out as indicated in chapter 7. The main objectives for evaluating TermPedia were to find out if:

- (i) By integrating the definitions of technical terms into course documents used by university students, the student's reading comprehension process is facilitated.
- (ii) By providing easy access to definitions of technical terms, the time needed for university students to acquire knowledge is reduced.

To achieve the first objective for TermPedia evaluation, 13 undergraduate students of Information Retrieval from the University of Groningen were selected to participate in a user survey. TermPedia was used to integrate the definition of technical terms found in an electronic textbook used by the students into the textbook, a process we refer to as document enrichment. After the students had read the enriched textbook, they answered a questionnaire that investigated if integrating the definition of technical terms into course documents facilitates their reading comprehension process. Analysis of the survey data indicates that the students tend to consistently agree that TermPedia is useful for easy comprehension of educational material.

The number of students from Gulu University who participated in the user study was 30. 15 of these students carried out an open-book test with the help of TermPedia and the other 15 carried out the same open-book test without the help of TermPedia. The user study evaluated if providing easy access to definition of technical terms through TermPedia, reduces the time needed by the students for knowledge acquisition. A reduction in time for knowledge acquisition would indicate that TermPedia facilitates the reading process. The study results did not provide supporting evidence to draw this conclusion. However there was a slightly better performance of students who carried out the open-book test with the help of TermPedia, compared to those who did the test without help of TermPedia. It is anticipated that students performed better in the open-book test with the help of TermPedia because they had easy access to contextually relevant definitions of technical terms and background information. Students also found the TermPedia approach usable because they could easily access information from the user interface that was developed for the user study.

8.2 Reflection on the Thesis Approach

This thesis had a two-part research trajectory in its approach. In the beginning the thesis took a technical approach and later on it took a design science approach. In

the technical approach, the researcher was interested in investigating what might usefully be done with human language technologies in facilitating the process of reading comprehension. This approach helped in realizing that university students in general have problems with reading and understanding technical knowledge. The approach also helped in identifying technical terms as a factor that hinders the reading comprehension process and led to a suggestion of document enrichment as a way of addressing this problem. As work on the thesis progressed, the researcher preferred design science because this perspective led to examining the development process from the point of view of potential users. Design science also broadened the researcher's intellectual horizons, leading her to reflect more deeply on why she was developing a service for students who were having trouble with their technical reading.

The design science trajectory adopted design science principles within an engaged scholarship philosophy, while following a Singerian inquiry strategy in a pragmatic framework of abduction. This research approach was useful for grounding relevant problems that affect society and searching for useful and usable solutions to such problems, through involvement of the society members. Applying the problem-grounding principles of design science research allowed the researcher to follow a rigorous course in confirming that insufficient vocabulary knowledge affects the reading comprehension process of university students. The problem-grounding principles were applied through literature and stakeholders' surveys. Results from these surveys supported the suggestion that a document enrichment approach would be an appropriate solution in facilitating the reading comprehension process. Other design science research principles provided appropriate guidance for designing the document enrichment approach as a solution artifact for the grounded problem. The principles also provided guidance for implementing, testing and evaluating the solution artifact.

Involving university students throughout the research was a way to follow the engaged scholarship philosophy. Reference points at which the students were involved were during problem-grounding and artifact evaluation. At the stage of problem grounding, students' opinions were solicited to learn about the problem domain of factors hindering reading comprehension at universities. TermPedia; an innovative solution for facilitating reading comprehension, was supported by applying an engaged scholarship philosophy in finding solutions to factors hindering reading comprehension. TermPedia evaluation results show that it has a good potential for facilitating the reading comprehension process.

A Singerian inquiry research strategy was carried out in a pragmatist framework of abductive reasoning. This inquiry research strategy provided the principles for

arriving at the observation that university students struggle with understanding course contents, which discuss unfamiliar subjects. This finding was arrived at after a number of students' opinions indicated that they have a real struggle in understanding some of their course contents which contain unfamiliar technical terms. Applying the pragmatist framework of abductive reasoning the poorly defined problems that hinder reading comprehension were dealt with. The problems related to reading comprehension occur in a complex environment where students have to apply their cognitive and linguistic processing capacities in understanding their course content. Abductive reasoning helped the researcher to infer through the students' experiences that, technical terms which occur in their course contents are hindrances to reading comprehension. It also helped in generating new knowledge so that we may conclude that document enrichment is a suitable approach for facilitating the reading comprehension process.

8.3 Research Contribution

This research made a number of contributions to various knowledge domains and practices. In a quest to establish requirements for an ideal solution, design science research guided the researcher to understand contemporary pertinent approaches to facilitating reading comprehension. In understanding these approaches, we found that they have both strengths and drawbacks. Some of the strengths in our interest included extraction of technical terms that occur in document content and linking such terms to contextually relevant Wikipedia pages [Mihalcea and Csomai, 2007]. The extracted technical terms are linked directly to Wikipedia so that readers may have easy access to the vocabulary and background knowledge for the terms. The drawback of this approach is that it does not directly provide a contextually relevant definition of the extracted term. A reader interested in knowing the meaning of an extracted term has to open a Wikipedia page, thus linking away from the content of the document being read. While in Wikipedia, the reader has to look for which portions of text provide a definition for the extracted technical term. Linking away from the content of documents disrupts the reading comprehension process and requires extra time and effort from the students since they have to search in the linked Wikipedia document for the definition of technical terms. Contributions to knowledge and practice of extracting technical terms and providing their meaning as presented in chapter 4 of this thesis, were as follows:

- (i) Using wikilink labels in extracting technical terms as opposed to using Wikipedia page titles. Wikilink labels represent technical terms better, compared to page titles because they occur within the text body in Wikipedia articles.

- (ii) Experimented with Wikipedia page in-links and the length value of wikilink label text strings in predicting technical terms that occur in document content.
- (iii) Using document enrichment techniques to integrate contextually relevant definitions of technical terms into document content.
- (iv) Using document enrichment techniques to integrate pictorial illustrations for technical terms into document content.

The above contributions were founded on the theory that the integration of contextually relevant vocabulary and background knowledge of technical terms into document content facilitates reading comprehension. Evaluation results of this document enrichment approach indicate that it can potentially facilitate reading comprehension because of the easy access to term definitions it provides. In turn, readers stick to the contents of the documents they are reading thereby reducing the time and effort required for their reading comprehension process.

In order to provide contextually relevant vocabulary and background knowledge for technical terms, our research was interested in term sense disambiguation, a technique that resolves the meaning of technical terms in context of the documents where the terms occur. A drawback of contemporary approaches to term sense disambiguation was that they apply complex mechanisms in natural language processing techniques, such as syntactic and lexical features, semantic representations and semantic inference. To bypass these complicated techniques, our idea of term sense disambiguation uses frequencies of linking wikilink labels to various Wikipedia pages and the frequency of in-links from various Wikipedia pages to a specific Wikipedia page as discussed in section 4.5.

In the field of vocabulary instruction Sedita [2005] suggested that substituting easy words for more difficult words could assist low-achieving students' reading comprehension. Replacing difficult words with easier ones denies students the opportunity to adequately and steadily grow their vocabulary knowledge, however students therefore acquire insufficient vocabulary knowledge, which affects their reading comprehension process [Chall and Jacobs, 2003]. Moreover, if information technology is applied to automatically substitute difficult words that occur in document content with easier ones, sentence meanings could be distorted if text context is ignored because both the original difficult word and its replacement may be ambiguous, making automatic replacement a complex task. Our approach to facilitating reading comprehension provides vocabulary knowledge for difficult words without a replacement. This allows students to understand their course content and gives them the opportunity to grow their vocabulary knowledge.

At the time of this research, the text wikification document enrichment approach did not take into consideration the importance of integrating definitions of predicted technical terms into document content. TermPedia integrates not only the definition of technical terms into document content but also provides graphical illustrations for such terms. Integrating definitions of technical terms into document content provides easy access to vocabulary knowledge while saving time and effort. In some cases these definitions are sufficient to quench a reader's thirst for comprehension, without the need for referring directly to a knowledge base. Nevertheless, these terms are linked to contextually related Wikipedia pages through various word sense disambiguation techniques as explained in section 6.4.

Dictionaries have been known for ages as one of the major sources of information about words and their definitions but we manipulated Wikipedia data using information technology and used it like a dictionary. We extracted technical terms and their definitions from this reference work. Unlike using Wikipedia data for generating dictionaries, which provide the definition of all single words [Akalya and Sherine, 2010], our approach provides definitions for only technical terms. We bore in mind that some technical terms are made up of multiple words. Besides, we do not only provide term definitions for technical terms but also their background information by linking the terms to Wikipedia. In providing the definition of technical terms, the context in which the terms occur is taken into consideration. Hence, contextually relevant definitions of the technical terms are provided, which is unlike the case of looking up the definition of terms in a traditional dictionary which provides all possible definitions for terms. With this comes the contribution of easy access to relevant definitions of technical terms there by facilitating the reading comprehension process and reducing the time needed for looking up background information on terms that are unfamiliar to a reader.

In the field of term recognition, we introduced term prediction using wikilink labels. [Section 4.2 explains wikilink labels] This is contrary to the work carried out by [Grineva et al., 2009] that used Wikipedia page titles for term recognition. The advantage of using wikilink labels is that they are a better representation of terms because they occur within the text body.

At the time of carrying out this research, contemporary works in word sense disambiguation resolved the meaning of single words in document context. Our approach of word sense disambiguation uses wikilink labels which occur both as single words and multi-word strings. Therefore, the notion of disambiguating the meaning of technical terms which are formed by combining words is added to the knowledge base and practice of word sense disambiguation.

In our document enrichment approach, pictorial illustrations were provided for enhancing reading comprehension without the cumbersome work of creating a

pre-defined image data set. We used images in Wikipedia where applicable. This work thus, enhances the practice of defining technical terms by providing pictorial illustrations along with the text definitions of the terms. Images were used without applying complex similarity statistics and other graphical technique for dealing with images and their relevance to context, unlike in the work done by Joshi et al. [2006].

8.4 Generalizability of TermPedia

Liesaputra and Witten [2012] argued that design science supports generalizability and utility of cases similar to the investigated research problem. In line with this argument Nunamaker et al. [1990] pointed out that design science artefacts must be generalizable. It was therefore noted that design science research generalizes beyond the case level [Wieringa, 2014] of an investigated problem. TermPedia was designed with the principles of generalizability in mind. For example, the technical term prediction technique uses a dictionary that contains terms from all known knowledge domains contained in Wikipedia. This dictionary can be used by the term prediction algorithm to predict technical terms from a majority of subject domains as represented by Wikipedia the content.

There are many aspects of generalizability but the important ones are whether a research artefact can deliver the same results when used by different people and, what design and technology aspects of the artefacts can be changed to suite the generalized scenarios [Namatovu, 2018]. TermPedia artefacts for this research was designed to help university students in their technical reading. However, other people can use this artifact with the same results. For example, document enrichment can provide practical solutions for facilitating reading comprehension for learners in secondary schools. TermPedia can be used for predicting and defining technical terms which occur in the course content of students in secondary schools. This will allow the students to have easy access to the meaning of unfamiliar terms, thereby facilitating their reading process. The idea of document enrichment can also be beneficial for readers outside the academic circle. For example, patients with peculiar medical conditions like morgellons, (a mysterious skin disorder [Virginia and Raphael, 2007]) who want to read about and understand their illnesses but do not have sufficient background and vocabulary knowledge for that task. When documents discussing such peculiar illnesses are enriched by TermPedia, the patients can easily access the definition of unfamiliar technical terms contained in the documents. TermPedia also provides easy access to background information about the peculiar illnesses by linking the technical terms to Wikipedia. Generalizing TermPedia for secondary students and patients concerns background and vocabulary knowledge as aspects of facilitating reading comprehension.

Other aspects of reading comprehension to which document enrichment can be generalized include word identification, sight words, vocabulary development and supporting the use of both mother tongue and the English language. People who can benefit from these generalization aspects are students and teachers. Document enrichment highlights technical terms in course content, therefore it can help teachers in identifying and presenting main course concepts and terminologies to their students. As the concepts and terms are introduced, it becomes easy for the students to identify and sight new words, which in turn helps the students in developing their vocabulary. There is increasing support for the use of mother tongues in the classrooms [Gacheche, 2010]. Document enrichment can help teachers to engage students in activities that support the use of mother tongues in the classrooms because Wikipedia has similar content in multiple languages. It is possible to present predicted terms and their meaning in the mother tongue of students in a specific classroom if content for that tongue exists in Wikipedia.

The term prediction algorithm is the one design aspect of the artefact that needs to change in order to suite other scenarios. This algorithm requires adjustment so that it predicts terms that are relevant to particular document contents since technical terms vary from one subject area to another. This will allow the term prediction algorithms to predict domain specific terms that will be more useful in the context of any given document content. Generalization of the TermPedia algorithms design was based on the following conditions:

- *Factors hindering the reading comprehension process.* Insufficient vocabulary and background knowledge are some of the factors that hinder the reading comprehension process of university students. Some scholars [Rayner and Reichle, 2010; Onwuegbuzie and Collins, 2002] indicate in their work that these factors do not only affect university students but the majority of people involved in reading in some way. In light of this, TermPedia's algorithms may be generalized beyond the scope of university students.
- *Practice in facilitating reading comprehension.* University students look up the meaning of terms that they do not understand from the internet, dictionaries and other reference sources like encyclopaedias. This practice is experienced by people reading in different context, for example secondary students, patients and teachers. Therefore the design of TermPedia's algorithms may be generalized to facilitate the reading comprehension process of people in such contexts.
- *Stakeholder involvement.* Reading comprehension is a multi-stakeholder task. These stakeholders include all people especially academicians, researchers and

professionals. Therefore the application of document enrichment fits very well in all areas where reading comprehension needs to be facilitated.

Although the design of TermPedia can be generalized to the contexts of secondary students, patients and teachers among others, there is need to first evaluate the applicability of the design to these contexts to ascertain its usefulness, usability and usage in the generalized scenarios.

8.5 Directions for Future Research

Chapter 6 reported low precision scores for the document enrichment algorithms in predicting technical terms. An explanation was given that the algorithms predicted many more technical terms compared to those that originally existed in the Wikipedia articles. A way improving the precision scores would be to classify both documents and technical terms according to specific knowledge domains. Classification would exclude the prediction of technical terms that do not belong to a given knowledge domain. This would reduce the number of technical terms predicted for specific document contents, thereby improving precision scores. Classification should also improve recall because technical terms relevant to specific knowledge domains would be predicted. Such terms are expected to be identical to the terms that originally existed in the Wikipedia articles. Since classification is expected to reduce the number of terms predicted while improving their relevance to specific knowledge domains, it should naturally improve recall scores for the document enrichment algorithms at predicting technical terms.

A classified terms dictionary would also enhance the term sense disambiguation techniques of TermPedia since the terms would be classified according to their subject domains. Therefore, terms for specific subject domains would be selected while predicting terms from contents of documents in that domain. For example, there could be a list of medical terms that are used for predicting technical terms in contents of documents that discuss issues related to medicine. Specialized dictionaries could also support the automatic generation of glossaries of technical terms, which can provide off-line access to technical terms definitions. Such glossaries can be important in instances when readers cannot access the definition of technical terms through the Internet.

Currently, the TermPedia document enrichment algorithms predict technical terms. The functions of these algorithms could be extended to include learning. In this way, students should be able to indicate which terms they find difficult to understand and TermPedia should be able to learn the difficult terms indicated by students. In that case, TermPedia should also be able to provide contextually relevant definitions

and background information for the terms learnt from students. This further research suggestion could allow TermPedia to be able to predict technical terms that are absent in the terms dictionary generated using data from Wikipedia.

During the stakeholders' survey, some students suggested that synonyms helped them to understand unfamiliar technical terms. Experimentation with technical term synonyms in cases where the technical terms are made of one word would help to confirm or disregard this claim. Therefore, TermPedia could be extended to provide term synonyms along with the term definitions, which should also facilitate the reading comprehension process. As mentioned in section 8.4 synonyms could also be used in automatically generating thesauri for document contents in an effort to facilitate the reading comprehension process.

Bibliography

- Abbas, P. G. and Narjes, B. S. (2016). A study of factors affecting efl learners' reading comprehension skill and the strategies for improvement. *International Journal of English Linguistics*, 6:180.
- Ács, N., O'Brien, C., Jiang, P., Burke, J., Jimenez, R., Garner, E., and Chwalisz, K. (2015). Treatment of endometriosis-associated pain with elagolix, an oral gnRh antagonist: Results from a phase 2, randomized controlled study. *Journal of Endometriosis*, 7(2):56–62.
- Adekoya, H. O. (2013). Schema theory: A conceptual review. *Journal of Research and Development*, 1(3):1–7.
- Akalya, B. and Sherine, F. (2010). Term recognition and extraction based on semantics for ontology construction. *International Journal of Computational Science Issues (IJCSI)*, 9(2).
- Alfa-informatica (2012). Alfa-informatica (information science). <http://www.rug.nl/let/onderwijs/afdelingen/informatiekunde/index>. Accessed: 16-May2012.
- An, S. (2013). Theory in reading. *Theory & Practice in Language Studies*, 3(1):130– 134.
- Ananiadou, S. (1994). A methodology for automatic term recognition. In *Proceedings of the 15th conference on Computational linguistics*, volume 2. Association for Computational Linguistics.
- Andres, M.A.C. (2014). Theories in reading instruction. <https://www.slideshare.net/netnetnetten/three-reading-models>. Accessed: 31-August-2017.

- Androutsopoulos, I. and Galanis, D. (2005). A practically unsupervised learning method to identify single-snippet answers to definition questions on the web. In *In HLT/EMNLP*, pages 323–330.
- Aramini, S. A., Ardizzone, E., and Mazzola, G. (2015). Automatic illustration of short texts via web images. In Braz, J., Kerren, A., and Linsen, L., editors, *IVAPP 2015 Proceedings of the 6th International Conference on Information Visualization Theory and Applications, Berlin, Germany, 11-14 March, 2015.*, pages 139–148. SciTePress.
- Aregu, R. (2010). *Market and Price Enhancement Decision Services for Farmers in Uganda*. PhD thesis, University of Groningen, The Netherlands.
- Barbier, F., Dop, P., Vermeent, S., and Fiedeldij Dop, P., e. (1972). *Winkler Prins medische encyclopedie*. Elsevier.
- Barrón-Cedeño, A., Sierra, G., Drouin, P., and Ananiadou, S. (2009). An improved automatic term recognition method for spanish. In Gelbukh, A., editor, *CICLing*, volume 5449, page 125–136, Berlin Heidelberg. Springer-Verlag.
- Barth, A. E., Vaughn, S., Capin, P., Cho, E., Stillman-Spisak, S., Martinez, L., and Kincaid, H. (2016). Effects of a text-processing comprehension intervention on struggling middle school readers. *Topics in Language Disorders*, 36(6):368–389.
- Bazerman, C. (2015)). *Involved: Writing for College, Writing for Your Self*. The WAC Clearinghouse, Fort Collins, CO.
- Berkow, R., Beers, M. H., and J., F. A., editors (2005). *Merck manual medisch handboek*. Merck Publishing, 2nd edition.
- Binotti, J. A., Hamilton-Gunkel, J. A., and Sipple, D. (2001). Reading deficiencies: A roadblock to success. Master’s thesis, Saint Xavier University and IRI/SkyLight Professional Development.
- Boggs, R. and Hammer, B. (2010). Formal inquiry systems: An inquiry profile for project management. *Issues in Information Systems (IIS)*, 6:199–207.
- Bosma, W., Theune, M., van Hooijdonk, C., Krahmer, E., and Maes, F. (2008). Illustrating answers: an evaluation of automatically retrieved illustrations of answers to medical questions. *Journal of Chemical Physics*.
- Bower, G. H. and Morrow, D. G. (1990). Mental models in narrative comprehension. *Science*, 247(4938):44–48.

- Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.
- Bron, M., Huurnink, B., and de Rijke, M. (2011). Linking archives using document enrichment and term selection. In Gradmann, S., Borri, F., Meghini, C., and Schuldt, H., editors, *Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2011, Berlin, Germany, September 26–28, 2011. Proceedings*, pages 360–371. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Brophy, M. C. (2009). Reading comprehension: Process and strategies. <https://www.ntid.rit.edu/sea/processes/comprehension/intro>. Accessed: 30-April2018.
- Busch, K. L. (2004). A glossary for mass spectrometry. <http://www.waters.com/webassets/cms/library/docs/720000807en.pdf>. Accessed: 22-April-2017.
- Cabré, C. M. T., Bagot, R. E., and Palatresi, J. V. (2001). Automatic term detection: A review of current systems. In Bourigault, D., editor, *Recent Advances in Computational Terminology*, chapter 3, pages 53–87. John Benjamins Publishing Company, Philadelphia, PA, USA.
- Cai, Z., Zhao, K., Zhu, K. Q., and Wang, H. (2013). Wikification via link cooccurrence. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management: San Francisco, California, USA, CIKM '13*, pages 1087–1096, New York, NY, USA. ACM.
- Cain, K. (2007). Deriving word meanings from context: Does explanation facilitate contextual analysis? *Journal of Research in Reading*, Vol.30:347–359.
- Cain, K., Oakhill, J., and Byant, P. (2000). Investigating the causes of reading comprehension failure: The comprehension-age match design. *Reading and Writing*, 12:31–40.
- Casper, M., Catton, J., and Westfall, S. (1998). Comprehension: Theories and strategies. <http://www.dominican.edu/academics/education/about/madaliennepeters/comprehension>. Accessed 22-March-2017.
- Chall, J. and Jacobs, V. (2003). Poor children’s fourth-grade slump. *American Educator*.

- Chang, C.-H., Lui, S.-C., and Wu, Y.-C. (2001). Applying pattern mining to web information extraction. In *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD '01*, pages 4–16, London, UK, UK. Springer-Verlag.
- Chervenick, E. C. (1992). *Schema Theory [microform]: Teaching U.S. History to Beginning Amnesty Students/Ellen Cecelia Chervenick*. Distributed by ERIC Clearinghouse [Washington, D.C.].
- Chongtay, R., Last, M., Verbeke, M., and Berendt, B. (2013). Summarize to learn: Summarization and visualization of text for ubiquitous learning. In *Proceedings of the 3rd IEEE Workshop on Interactive Visual Text Analytics at VIS*, pages 1–4, Atlanta, GA.
- Creswell, J. (2013). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications.
- Cromley, J. and Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, 99(2):311–325.
- Cruz, R. I. S. and Escudero, M. D. P. (2012). Models of reading comprehension and their related pedagogical practices: A discussion of the evidence and a proposal. *Mextesol journal*, 36(2):1–18.
- Csomai, A. and Mihalcea, R. (2007). Linking educational materials to encyclopedic knowledge. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED 2007)*, Los Angeles, CA.
- Davoudi, M. and Moghadam, H. R. H. (2015). Critical review of the models of reading comprehension with a focus on situation models. *International Journal of Linguistics*, 7(5):172–187.
- Desrist.org (2015). Science research in information systems. <http://desrist.org/desrist/content/design-science-research-in-information-systems.pdf>. Accessed 13-November-2017.
- Domingue, J., Franes, N., Motta, E., Shum, S. B., Vargas-Vera, M., and Kalfoglou, Y. (2001). Supporting ontology-driven document enrichment within communities of practice. In *The 1st International Conference on Knowledge Capture (K-Cap) proceedings*, Victoria, BC, Canada.

- Dorsey III, B. H. (2016). Understanding and improving website bounce rates. <https://www.yourspear.com/wp-content/uploads/2016/11/SH-WP-BounceRate.pdf>. Accessed 02-November-2019.
- Duckworth, M. and Spyrou, A., editors (1995). *Law Words*. Centre for Plain Legal Language.
- Duffy, G., Rochler, L., and Mason, J., editors (1984). *A schema-theoretic view of the reading process as a basis for comprehension instruction*, New York. Longman.
- Dworkin, S. L. (2012). Sample size policy for qualitative studies using in-depth interviews. *Archives of Sexual Behavior*, 41(6):1319–1320.
- Edmonds, P. and Agirre, E. (2008). Word sense disambiguation. *Scholarpedia*, 3(5):4358.
- Elwér, Ä . (2014). *Early predictors of reading comprehension difficulties*. Linköping University, Linköping, 1:1 edition.
- EMA (2005). European medicines agency: Science medicines with health. <https://www.ema.europa.eu/en>. Accessed: 16-May-2005.
- Engineer, S. (2005). 21st century english vocabulary.
- Erdem, E. (1996). Data structures and algorithms. <http://www.cmpe.boun.edu.tr/~akin/cmpe223/chap2.htm>.
- Fahmi, I., Bouma, G., and der Plas, L. V. (2007). Learning to identify definitions using syntactic features. In *EACL 2006 Workshop on Learning Structured Information in Natural Language Application*, pages 64–71.
- Fairbairn, G. and Fairbairn, S. (2001). *Reading At University: A Guide for Students*. Open University. All publications. McGraw-Hill Companies Incorporated.
- Fender, M. (2008). Spelling knowledge and reading development: Insights from arab esl learners. *Reading in a Foreign Language*, 20(1):19–42.
- Ferragina, P. and Scaiella, U. (2010). Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management: Toronto, ON, Canada, CIKM '10*, pages 1625–1628, New York, NY, USA. ACM.
- Fiester, L. (2010). Early warning! why reading by the end of third grade matters. Technical report, Annie E. Casey Foundation. Accessed: 23-November-2017.

- Fisher, D. (2009). Vocabulary and background knowledge. In *Important Factors in Reading Comprehension*. McGraw Hill Education.
- Fisher, D. and Frey, N. (2013). Background knowledge. In *The Overlooked Factor in Reading Comprehension*. McGraw-Hill Education.
- Flowers, P. (2009). Research philosophies—importance and relevance. *European Journal of Information Systems*, 3(2):112–126.
- Freund, J. E. (2004). *Mathematical statistics with applications*. Pearson Education, Inc., 7th. edition.
- Gacheche, K. (2010). Challenges in implementing a mother tongue-based languagein-education policy: Policy and practice in kenya. *POLIS J.*, 4.
- Gao, D., Li, W., Ouyang, Y., and Zhang, R. (2012). Lda-based topic formation and topic-sentence reinforcement for graph-based multi-document summarization. In Hou, Y., Nie, J.-Y., Sun, L., Wang, B., and Zhang, P., editors, *Information Retrieval Technology: 8th Asia Information Retrieval Societies Conference, AIRS 2012, Tianjin, China, December 17-19, 2012. Proceedings*, pages 376–385. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Gelbukh, A. F., editor (2014). *Computational Linguistics and Intelligent Text Processing 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part I*, volume 8403 of *Lecture Notes in Computer Science*. Springer.
- Gilakjani, A. P. and Ahmadi, M. R. (2011). A study of factors affecting efl learners’ english listening comprehension and the strategies for improvement. *Journal of Language Teaching & Research*, 2(5).
- Giles, J. (2005). Internet encyclopedias go head to head. *Nature*, Vol. 438:900–901. Published online 14 December 2005.
- Gliem, J. A. and Gliem, R. R. (2003). Calculating, interpreting, and reporting cronbach’s alpha reliability coefficient for likert-type scales. Presented at the Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, The Ohio State University, Columbus, OH, October 8-10, 2003.
- Goffinet, L. and Noirhomme-Fraiture, M. (1995). Automatic hypertext link generation based on similarity measures between documents. Unpublished research paper.

- Gonzalez, R. and Sol, H. (2012). Validation and design science research in information systems. In Mora, O., Steenkamp, G. A., and Raisinghani, M., editors, *Methodologies, Innovations and Philosophies in Software Systems Engineering and Information Systems*, pages 403–426. PA: IGI Global, Hershey.
- Goodman, K. S. (1967). Reading: A psycholinguistic guessing game. *Journal of the Reading Specialist*, 6(4):126–135.
- Grabe, W. (1988). *Interactive Approaches to Second Language Reading: Reassessing the term “interactive”*. Cambridge University Press.
- Graham, L. and Bellert, A. (2005). Reading comprehension difficulties experienced by students with learning disabilities. *Australian Journal of Learning Disabilities*, 10(2):71–78.
- Graves, M. and Graves, B. (2003). *Scaffolding Reading Experiences: Designs for Students Success*. Christopher-Gordon Publishers, Inc., 2nd edition.
- Griffiths, G., Sohlberg, M. M., and Biancarosa, G. (2011). A review of models of reading comprehension with implications for adults with mtbi and the campus reader. http://www.campusreader.org/dev/resources/pdfs/CampusReadertheory_final0401.pdf. Accessed:10-April-2017.
- Grineva, M., Grinev, M., and Lizorkin, D. (2009). Extracting key terms from noisy and multitheme documents. In *Proceedings of the 18th International World Wide Web Conference (WWW '09): Madrid, Spain*, pages 661–670, New York, NY, USA. ACM.
- Guba, E. G. and Lincoln, Y. S. (1994). Competing paradigms in qualitative research. *Handbook of qualitative research*, 2:163–194.
- Gunning, T. G. (1996). *Creating Reading Instruction for All Children*. Allyn and Baconeds, Boston.
- Hameed, H. T. (2008). A comparison of l1 and l2 reading: Cultural differences and schema. *Diala, Jour*, 28.
- Haslam, T. J. (2017). Wikipedia and the humanities in higher education: Past time to renegotiate the relationship. *International Journal of Information and Education Technology*, 7(4):246–251.
- Hearst, M. (1991). Noun homograph disambiguation using local context in large text corpora. In *Proceedings of the Annual Conference of the UW Center for the New OED and Text Research*, pages 1–22.

- Hendricks, K. M. (2013). *Reading and Test Taking in College English as a Second Language Students*. Psychology dissertation, Syracuse University. Paper 181.
- Herlina, U. (2016). Increasing reading comprehension ability through visualizing media of 4th grader students at labschool primary school setiabudi south jakarta. *American Journal of Educational Research*, 4(18):1265–1270.
- Hevner, A. (2007). The Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, 19(2):87.
- Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research. *MIS Q.*, 28(1):75–105.
- Hiroshi, N. and Tatsunori, M. (2000). A simple but powerful automatic term extraction method. Unpublished research paper.
- Hobson, E. (2004). Getting students to read: Fourteen tips. In *Idea Paper 40*. The Idea Center.
- Hoeft, M. E. (2012). Why university students don't read: What professors can do to increase compliance. *International Journal for the Scholarship of Teaching and Learning*, 6(2).
- Hoover, W. A. and Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal*, pages 127–160.
- Horton, S., Beattie, L., and Bingle, B. (2015). *Lessons in Teaching Reading Comprehension in Primary Schools*. SAGE Publications Inc.
- IRA (1990). A position statement from the international reading association on secondary school reading. *Journal of Reading*, 33(4):285–287.
- Jabbar, A. (2015). How to structure your research approach. <https://researcholic.wordpress.com/2015/03/10/how-to-structure-your-research-approach-within-qualitative-research/>. Accessed: 17-January-2017.
- Jennings, J. H., Caldwell, J. S., and Lerner, J. W. (2014). *Reading Problems: Assessment and Teaching Strategies*. Pearson, 7th edition.
- Johnson (2011). A brain-buster from america's book of tax-law. https://www.economist.com/blogs/johnson/2011/04/legal_language. Accessed: 23-November0-2017.

- Johnson-Laird, P. (1981). Comprehension as the construction of mental models. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 295(1077):353–374.
- Jones, M. S. (2007). An evaluation of different term extraction techniques for searching ontologies. Unpublished research paper.
- Joshi, D., Wang, J. Z., and Li, J. (2006). The story picturing engine—a system for automatic text illustration. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):68–89.
- K12Reader (2016). What is reading comprehension? <http://www.k12reader.com/what-is-reading-comprehension/>. Accessed: 14-March-2017.
- Kamarudin, K. M., Ridgway, K., and Ismail, N. (2016). Abstraction and generalization in conceptual design process: Involving safety principles in triz-sda environment. *Procedia CIRP*, 39(Supplement C):16 – 21. Structured Innovation with TRIZ in Science and Industry: Creating Value for Customers and Society.
- Kasture, N. R. and Agrawal, A. (2012). A supervised word sense disambiguation method using ontology and context knowledge. *Computer Engineering and Intelligent Systems*, 3(8).
- Katz, L., Brancazio, L., Irwin, J., Katz, S., Magnuson, J., and Whalen, D. (2012). What lexical decision and naming tell us about reading. *Read Writ*, 25(6):1259– 1282.
- Keene, E. O. (2002). Comprehension strategies. <http://www.readinglady.com/mosaic/tools/COMPREHENSIONSTRATEGIESbyElli2002.pdf>. Retrieved: February 16, 2011.
- Kendeou, P., van den Broek, P., Helder, A., and Karlsson, J. (2014). A cognitive view of reading comprehension: Implications for reading difficulties. *Learning Disabilities Research & Practice*, 29(1):10–16.
- Kharismawati, A. (2015). Improving the reading comprehension ability of the tenth grade students of sma n 6 yogyakarta in the academic year of 2013/2014 using collaborative strategic reading. <http://eprints.uny.ac.id/17803/>. Accessed:10April-2017.
- Klapaftis, I. and Manandhar, S. (2005). Google & wordnet based word sense disambiguation. In *Proceedings of the Workshop on Learning and Extending Ontologies by using Machine Learning methods*, Bonn, Germany. held at the 22nd International Conference on Machine Learning (ICML05).

- Knoth, P., Schmidt, M., Smrž, P., and Zdráhal, Z. (2009). Towards a framework for comparing automatic term recognition methods. In *Conference Znalosti 2009*.
- Korkontzelos, I., Klapaftis, I. P., and Manandhar, S. (2008). *Reviewing and Evaluating Automatic Term Recognition Techniques*, pages 248–259. Springer, Berlin, Germany.
- Krovetz, R. and Croft, W. B. (1992). Lexical ambiguity and information retrieval. *ACM Trans. Inf. Syst.*, 10(2):115–141.
- Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E., and Soderland, S. (1994). Evaluating an information extraction system.
- Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, New York, USA. ACM.
- Liesaputra, V. and Witten, I. H. (2012). Realistic electronic books. *International Journal of Human-Computer Studies*, 70(9):588–610.
- Lindsey, R. V., Shroyer, J. D., Pashler, H., and Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological Science*, 25(3):639–647. PMID: 24444515.
- Liu, F. (2010). Short analysis of the nature of reading. *Language Teaching*, 3(3). Liu, Z. (2006). Print vs. electronic resources: A study of user perceptions, preferences, and use. *Inf. Process. Manage.*, 42(2):583–592.
- Lo, R., He, B., and Ounis, I. (2005). Automatically Building a Stopword List for an Information Retrieval System. In *Proceedings of the Fifth Dutch-Belgian Information Retrieval Workshop*.
- López, A. R. (2013). *Greening the Media Literacy Ecosystem: Situating Media Literacy for Green Cultural Citizenship*. PhD thesis, Prescott College.
- Lowe, W. and King, G. (2003). Some statistical methods for evaluating information extraction systems. In *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?:Budapest, Hungary, Evalinitatives '03*, pages 19–25, Stroudsburg, PA, USA. Association for Computational Linguistics.

- Lui, A.-F., Ng, V.-C., Tsang, E. K., and Ho, A. C. (2011). Generation of hypertext for web-based learning based on wikification. In Kwan, R., McNaught, C., Tsang, P., Wang, F., and Li, K., editors, *Enhancing Learning Through Technology. Education Unplugged: Mobile Technologies and Web 2.0*, volume 177 of *Communications in Computer and Information Science*, pages 280–290. Springer Berlin Heidelberg.
- Makhoul, J., Kubala, F., Schwartz, R., and Weischedel, R. (1999). Performance measures for information extraction. In *In Proceedings of DARPA Broadcast News Workshop*, pages 249–252.
- Mancilla-Martinez, J. and Lesaux, N. K. (2010). Predictors of reading comprehension for struggling readers: The case of spanish-speaking language minority learners. *Journal of Educational Psychology*, 102(3):701–711.
- Manning, C. D., Raghavan, P., and Schütze, H. (2008). *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.
- Manning, C. D., Raghavan, P., and Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.
- Maulizan, Z. A. (2015). Teaching reading comprehension to efl junior high school students. *ENGLISH EDUCATION JOURNAL (EEJ)*, 6(3):370–382.
- Mayer, R. (2005). *The Cambridge handbook of multimedia learning*. Cambridge University Press.
- McNamara, D. S. (2007). *Reading Comprehension Strategies: Theories, Interventions, and Technologies*. Lawrence Erlbaum Associates.
- Medialab (2012). Wikipedia extractor. http://medialab.di.unipi.it/wiki/Wikipedia_Extractor. Cited October 1, 2012.
- Merck (2005). Merck manual professional version: The trusted provider of medical information since 1899. <https://www.merckmanuals.com/professional>. Accessed: 16-May-2005.
- Mihalcea, R. (2007). Using wikipedia for automatic word sense disambiguation. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, April 22-27, 2007, Rochester, New York, USA*, pages 196–203. The Association for Computational Linguistics.
- Mihalcea, R. and Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on information and knowledgemanagement*, pages 233–242, Lisbon, Portugal. ACM.

- Milios, E., Zhang, Y., He, B., and Dong, L. (2003). Automatic term extraction and document similarity in special text corpora. In *proceeding of the 6th Conference of the Pacific Association for Computational Linguistics*, pages 275–284.
- Miller, G. A. and Johnson-Laird, P. B. (1976). *Language and Perception*. Harvard Univ Press, Cambridge MA.
- Miller, S. and Schwartz, R. (1994). Statistical language processing using hidden understanding models. In *Proc. ARPA Speech and Natural Language Workshop*, pages 278–282, Germany. Morgan Kaufmann Publishers, Inc.
- Milne, D. and Witten, I. (2009). An open-source toolkit for mining Wikipedia. In *Proc. New Zealand Computer Science Research Student Conf., NZCSRSC*, volume 9.
- Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management: Napa Valley, California, USA, CIKM '08*, pages 509–518, New York, NY, USA. ACM.
- Mirembe, D. P. (2015). *The Threat Nets Approach to Information System Security Risk Analysis*. PhD thesis, University of Groningen.
- Mohammad, D. and Hamidreza, H. M. (2015). Critical review of the models of reading comprehension with a focus on situation models. *International Journal of Linguistics*, 7:172.
- Moodley, P. and Singh, R. (2015). Addressing student dropout rates at south african universities. *Alternation*, (17):91–115.
- Moore, A. L. (2016). A research review of cognitive skills, strategies, and interventions for reading comprehension. <http://www.book.dislib.info/b1-other/752991-1-reading-comprehension-amy-moore-reading-comprehension.cognitive-s.php>. Accessed: 28-February-2017.
- Moore, J. (2013). Vocabulary instruction – should it be a priority? <https://prezi.com/qpvsp0espbb/vocabulary-instruction-should-it-be-a-priority/>. Accessed: 13-April-2017.
- Motta, E., Shum, S. B., and Domingue, J. (2000). Ontology-driven document enrichment: principles, tools and applications. *International Journal of HumanComputer Studies*, 52(6):1071–1109.
- Mulholland, P., Zdrahal, Z., Domingue, J., and Hatala, M. (2000). Integrating working and learning: A document enrichment approach. <http://projects.kmi.open.ac.uk/enrich/BIT-working-and-learning.pdf>. Accessed: 4-November-2017.

- Murnane, R., Sawhill, I., and Snow, C. (2012). Literacy challenges for the twenty-first century: Introducing the issue. *The Future of Children*, 22(2):3–15.
- Murphy, S. (2013). *Assessing Text Difficulty for Students*. Number 44 in What Works? Research into Practice. Ontario Ministry of Education.
- Murray, R. W. (1996). Historical linguistics: the study of language change. In O’Grady, W., Dobrovolsky, M., and Katamba, F., editors, *Contemporary Linguistics. An introduction*, chapter 8, pages 313–371. Longman, London and New York, 3rd. edition.
- Nagy, W. E. (1988). Technical report no. 431. In *Vocabulary instruction and reading comprehension*. University of Illinois at Urbana Campaign.
- Namatovu, H. K. (2018). *Enhancing antenatal care decisions among expectant mothers in Uganda*. PhD thesis, University of Groningen, SOM research school, Groningen, The Netherlands.
- Navigli, R. (2009). Word sense disambiguation: a survey. *ACM COMPUTING SURVEYS*, 41(2):1–69.
- Nsamenang, A. and Tchombé, T. (2012). *Handbook of African Educational Theories and Practices: A Generative Teacher Education Curriculum*. Human Development Resource Center (HDRC).
- Nunamaker, J. J. F., Chen, M., and Titus, D. P. (1990). Systems development in information systems research. *Journal of Management Information Systems*, 7(3):89–106.
- Onwuegbuzie, A. J. and Collins, K. M. (2002). Reading comprehension among graduate students. *Psychological Reports*, pages 879–882.
- Orlikowski, W. J. and Baroudi, J. J. (1991). Studying information technology in organizations: Research approaches and assumptions. *Information Systems Research*, 2:1–28.
- Osborne, J. and Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Assessment, Research & Evaluation*, 9(6):1–8. Available online: <http://PAREonline.net/getvn.asp?v=9&n=6>.
- Oxford Living Dictionaries (2017). <https://en.oxforddictionaries.com/definition/fitpiration>. Accessed 22-November-2017.
- Pal, A. R. and Saha, D. (2015). Word sense disambiguation: a survey. *CoRR*, abs/1508.01346.

- Palmer, B. C., Bilgili, E. M., Gungor, A., Taylor, S. H., and Leclere, J. T. (2008). Reading comprehension, figurative language instruction, and the turkish english language learner. *Reading Horizons*, 48(4).
- Palmer, D. (2016). Theories of comprehension. http://dlpalmer.weebly.com/uploads/3/5/8/7/3587856/theories_of_comprehension.docx. Accessed: 04April-2017.
- Pardo, L. S. (2004). What every teacher needs to know about comprehension. *International Reading Association*, page 272–280.
- Parker, K. and Chao, J. (2007). Wiki as a teaching tool. *Interdisciplinary Journal of Knowledge and Learning Objects*, 3:57–72.
- Patajoki, U. (2013). Towards a successful contractual relationship: Public service procurement from a small business perspective. Master's thesis, Aalto University.
- Peeck, J. (1974). Retention of pictorial and verbal content of a text with illustrations. *Educational Psychology*, pages 880–888.
- Peppers, K., Tuunanen, T., Rothenberger, M., and Chatterjee, S. (2007). A design science research methodology for information systems research. *J. Manage. Inf. Syst.*, 24(3):45–77.
- Perfetti, C. (1985). *Reading ability*. Oxford University Press, Oxford.
- Perfetti, C. and Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, 18(1):22–37.
- Pontus, P.-S., Matheson, G. J., Schiffler, B. C., and Thompson, W. H. (2017). The readability of scientific texts is decreasing over time. <https://doi.org/10.1101/119370>.
- Pope, C. and Mays, N. (1995). Qualitative research: Reaching the parts other methods cannot reach: an introduction to qualitative methods in health and health services research. *BMJ*, 311(6996):42–45.
- Pressley, M., Yokoi, L., van Meter, P., Van Etten, S., and Freebern, G. (1997). Some of the reasons why preparing for exams is so hard: What can be done to make it easier? *Educational Psychology Review*, 9(1):1–38.
- Radvansky, G. A. and Copeland, D. E. (2001). Working memory and situation model updating. *Memory and Cognition*, 29(8):1073–1080.

- Rauha, N. H. (2011). Investigating reading difficulties in english second language of grade 3 learners in one primary school in the khomas education region of namibia. Master's thesis, THE UNIVERSITY OF NAMIBIA.
- Rayner, K. and Reichle, E. (2010). Models of the reading process. *Wiley Interdisciplinary Reviews*, 1(6):787–799.
- Reutzel, D. R. and Cooter Jr., R. B. (2015). *Teaching Children to Read: The Teacher Makes the Difference*. Pearson, 7th edition.
- Ricciardi, F. (2009). Epistemology of information systems: Time for something new? positivism, interpretivism, and beyond. In D'Átri, A. and Saccá, D., editors, *Systems: People, Organizations, Institutions, and Technologies*. Physica-Verlag HD.
- Rieber, L. P. (1990). Animation in computer-based instruction. *Educational Technology Research and Development*, 38(1):77–86.
- Roehl, K. and Shiue, C. (2003). Developing reading comprehension skills in efl university level students. *Foreign Language Learning and Teaching Conference*, 1(2).
- Ross, J. M. (2010). Informatics creativity: a role for abductive reasoning? *j-CACM*, 53(2):144–148.
- Rotard, M., Gieretha, M., and Ertl, T. (2007). Semantic lenses: Seamless augmentation of web pages with context information from implicit queries. *Computers and Graphics*, Vol.31(No.3):361–369.
- Rumelhart, D. E. (1977). Toward an integrative model of reading. In S. Dornic (Ed., *Attention and performance*, pages 573–603.
- Sandra, S., Dobson, B. K., and Clarke, M. A. (2008). *Reader's Choice*. University of Michigan, 5th edition.
- Sarawagi, S. (2008). Information extraction. *Found. Trends databases*, 1(3):261–377.
- Schaefer, D., editor (2014). *Product Development in the Socio-sphere: Game Changing Paradigms for 21st Century Breakthrough Product Development and Innovation*. Springer.
- Schwartz, M. (2012). Getting students to do their assigned readings. http://www.ryerson.ca/content/dam/lt/resources/handouts/student_reading.pdf. Accessed: 23-November-2017.

- Sebastiani, F., editor (2003). *Advances in Information Retrieval, 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14-16, 2003, Proceedings*, volume 2633 of *Lecture Notes in Computer Science*. Springer.
- Sedita, J. (2005). Effective vocabulary instruction. *Insights on Learning Disabilities*, 2(1):33–45.
- Shafie, L. A. and Nayan, S. (2011). The characteristics of struggling university readers and instructional approaches of academic reading in malaysia. *International Journal of Human Sciences*, 8(1).
- Shahnazari, M. T. and Dabaghi, A. (2014). A critical overview of models of reading comprehension with a focus on cognitive aspects. *Research In English Language Pedagogy*, 2.
- Shehu, I. (2015). Reading comprehension problems encountered by foreign language students, case study: Albania, croatia. *Academic Journal of Interdisciplinary Studies*, 4(1 S1).
- Shen, Y. (2008). An exploration of schema theory in intensive reading. *English Language Teaching*, 1(2):104–107.
- Shihab, I. A. (2011). Reading as critical thinking. *Asian Social Science*, 7(8):209–218.
- Smith, R. (1996). What clinical information do doctors need? *BMJ*, 313(7064):1062–1068.
- Snow, C. E. (2002). *Reading for understanding: toward a research and development program in reading comprehension*. RAND, Santa Monica, CA.
- Snow, C. E. (2010). Reading comprehension: Reading for learning. *Harvard Graduate School of Education*.
- Sol, H. (1988). Information system: A problem solving approach. In *In Proceedings of 1988 INTEC Symposium Systems Analysis and Design: A Research Strategy*, Atlanta Georgia.
- Sol, H. G. (1982). *Simulation in Information Systems Development*. PhD thesis, University of Groningen.
- Sommerville, I. (2001). *Software Engineering*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 6th edition.
- Sousa, D. (2011). *How the ELL Brain Learns*. Calif.
[dio:https://books.google.co.ug/books?id=9loznQAACAAJ](https://books.google.co.ug/books?id=9loznQAACAAJ).

- Spiro, R. J., Bruce, B. C., and Brewer, W. F. (1980). *Theoretical Issues in Reading Comprehension: Perspectives from Cognitive Psychology, Linguistics, Artificial Intelligence, and Education*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Sreedhar, J., Raju, S. V., Babu, A. V., Shaik, A., and Kumar, P. P. (2012). Word sense disambiguation: An empirical survey. *International Journal of Soft Computing & Engineering*, 2:494–503.
- Stanovich, K. (1988). The language code: Issues in word recognition. In Yussen, S. and Smith, M., editors, *Reading across the lifespan*, pages 87–94. Springer-Verlag, New York.
- Stojanovic, Z., Dahavayake, A., and Sol, H. (2001). A methodology framework for component-based systems development support. In *Proceedings of the 6th CAISE/IFIP8.1 International Workshop on Evaluation of Modeling Methods in Systems Analysis and Design*.
- Sullivan, G. and Sargeant, J. (2011). Qualities of qualitative research: Part i. *Journal of Graduate Medical Education*, 3(4):449–452.
- Svinicki, M. (1993-1994). What they don't know can hurt them: The role of prior knowledge in learning. *Essays on Teaching Excellence: Toward the Best in the Academy*, 5(4).
- Talanquer, V. (2013). *How Do Students Reason About Chemical Substances and Reactions?*, pages 331–346. Springer Netherlands, Dordrecht.
- Tan, J. P. J. G. (2015). Using outline to enhance reading comprehension in a high school english language classroom. *American Journal of Educational Research*, 3(7):893–896.
- Taylor-Powell, E. and Renner, M. (2003). Analyzing qualitative data. *Screen*, 25(3):217–20.
- Thuring, M., Hannemann, J., and Haake, J. M. (1995). Hypermedia and cognition:. In *Design for Comprehension*, volume 38(8). Community of the ACM.
- Torralbo, R., Alfonso, E., Moreno-Sandoval, A., and Guirao, J. M. (2005). Automatic generation of term definitions using multidocument summarisation from the web. In *RANLP '05 Workshop on Crossing Barriers in Text Summarization Research proceedings*.
- Tumwebaze, R. P. (2016). *Decision Enhancement for Poultry Farmers in East Africa*. PhD thesis, University of Groningen.

- Turdakov, D. and Velikhov, P. (2008). Semantic relatedness metric for wikipedia concepts based on link analysis and its application to word sense disambiguation. In *Proceedings of the SYRCODIS 2008 Colloquium on Databases and Information Systems, Saint-Petersburg, Russia, May 29-30, 2008*, volume 355 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Tweheyo, R. (2018). *Indigenous Knowledge and Food Security: Enhancing Decisions of Rural Farmers*. PhD thesis, University of Groningen, The Netherlands.
- Unruh, S. (2015). Struggling international students in the us: Do university faculty know how to help? *Athens Journal of Education*, 2(2).
- Ur, P. (2004). *A course in Language Teaching. Practice and Theory*. Cambridge University press.
- Vande Ven, A. (2007). *Engaged Scholarship: A Guide for Organizational and Social Research*. OUP Oxford.
- Van de Ven, A. (2009). Reflections on engaged scholarship. <https://www.iei.liu.se/kite/dokument/1.117786/ESReflections.pdf>. Accessed 08-November-2017.
- Vekiri, I. (2002). What is the value of graphical displays in learning? *Educational Psychology Review*, 14(261).
- Virginia, R. S. and Raphael, S. (2007). Morgellons disease: the mystery unfolds. *Expert Review of Dermatology*, 2:585–591.
- Vivaldi, J. and Rodríguez, H. (2010). Finding domain terms using wikipedia. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Wang, Y. and Gafurov, D. (2003). The cognitive process of comprehension. In *Proceedings of the 2Nd IEEE International Conference on Cognitive Informatics, ICCI '03*, pages 93–, Washington, DC, USA. IEEE Computer Society.
- Wanja, K. (2012). *and Comprehension in the African Context: A Cognitive Enquiry*. Zapf Chancery Publisher, African Ltd., Limuru, Kenya.
- Wharton, C. and Kintsch, W. (1991). An overview of construction-integration model: A theory of comprehension as a foundation for a new cognitive architecture. *SIGART Bull.*, 2(4):169–173.

- Wieringa, R. (2014). *Design science methodology for information systems and software engineering*. Springer. 10.1007/978-3-662-43839-8.
- Wijnhoven, A. (2013). *Enabling the collective brain for organizations: a quickstart in management software skills*. University of Twente, Faculty Management and Governance, Department.
- Wikgren, M. (2005). Critical realism as a philosophy and social theory in information science? *Journal of Documentation*, 61(1):11–22.
- Wilfrid Laurier University (2008). Active reading: Getting the most out of your text. https://web.wlu.ca/learning_resources/pdfs/Active_Reading_Note_Taking.pdf. Accessed: 02-May-2018.
- William, E. (1988). Instruction and reading comprehension. Technical Report 431, University of Illinois.
- Wills, S. T. (2004). Barriers: Enhanced comprehension of patient education material. *Theory, Research, Education, and Training*, pages 140–144.
- Woolley, G. (2010). Developing reading comprehension. *Australian Journal Of Language & Literacy*, 33(2):108–125.
- Woolley, G. (2011). *Reading Comprehension*. Springer, Dordrecht, Heidelberg, London, New York.
- Yu, L., Pajouhi, H., Nelis, M., Rhoads, J., and Mohammadi, S. (2012). Tunable, dualgate, silicon-on-insulator (soi) nanoelectromechanical resonators. *Nanotechnology, IEEE Transactions on*, 11(6):1093–1099.
- Zachariadis, M., Scott, S. V., and Barrett, M. I. (2010). Exploring critical realism as the theoretical foundation of mixed-method research: evidence from the economies for is innovations. working paper, Judge Business School, Cambridge, UK.
- Zhang, C. (2010). The teaching of reading comprehension under the psychology schemata theory. *Journal of Language Teaching and Research*, 1(4):457–459.
- Zhang, M., Qin, B., Liu, T., and Zheng, M. (2014). Triple based background knowledge ranking for document enrichment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 917–927, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- Ziemke, T. (1999). Rethinking grounding. In Riegler, A., Peschl, M., and von Stein, A., editors, *Representation in the Cognitive Sciences*. Springer, Boston, MA.

- Zipf, H. (1949). *Human Behaviours and the Principles of Least effort*. AddisonWesley, Cambridge, MA.
- Zwaan, R. A. (2015). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic bulletin & review*, 23(4):1028–34.
- Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162–185.

Appendix A

TermPedia Evaluation Tools and Data

A.1 Traffic to TermPedia User Interface

Google Analytics

http://www.rug.nl/~olango/TermP...
www.rug.nl/~olango/TermP...

All Traffic

100.00% of total visits

Explorer

Site Usage

Visits

Visits	Pages/Visit	Avg. Visit Duration	% New Visits	Bounce Rate
68	3.21	00:07:13	29.41%	52.94%
% of Total: 100.00% (68)	Site Avg: 3.21 (0.00%)	Site Avg: 00:07:13 (0.00%)	Site Avg: 29.41% (0.00%)	Site Avg: 52.94% (0.00%)

Source/Medium	Visits	Pages/Visit	Avg. Visit Duration	% New Visits	Bounce Rate
1. nestor.rug.nl / referral	48	3.52	00:09:07	29.17%	58.33%
2. (direct) / (none)	14	1.86	00:03:00	35.71%	57.14%
3. rug.nl / referral	3	3.67	00:01:52	0.00%	0.00%
4. google / organic	2	4.50	00:02:26	0.00%	0.00%
5. let.rug.nl / referral	1	3.00	00:00:18	100.00%	0.00%

Rows 1 - 5 of 5

A.2 TermPedia User Survey Questionnaire for RUG

Your contribution to TermPedia user evaluation is much appreciated. Please provide the following information by completing this questionnaire. All information will be kept confidential and only used for the purpose of this user evaluation.

Personal Data

Age: []

Gender, (male [M]/female [F]): []

Student number: [.....]

Did you use TermPedia at all? (yes [Y]/no [N]): [] If

yes, on how many occasions?: []

Usability Data

On a scale of 4 to 0 give your impression on the following questions where:

4 ≡ Strongly

agree 3 ≡ Agree

2 ≡ Somewhat

agree 1 ≡ Disagree

0 ≡ Strongly disagree

A. TermPedia User Interface

- (i) I enjoy reading on-line educational literature. []
- (ii) It was easy to locate information on TermPedia user interface. []
- (iii) It was easy to navigate on TermPedia user interface. []
- (iv) The colors and design on TermPedia user interface are appealing. []
- (v) The text font, size, and color on TermPedia user interface are easy to read. []

B. Document Technical Difficulty

- (i) There were many technical terms in the IR e-book. []
- (ii) The technical terms were familiar and easy to understand. []
- (iii) It was possible to understand text context around unfamiliar technical terms without help. []
- (iv) Sentences in the IR e-book were short enough to understand. []
- (v) The IR e-book was written in easy English grammar. []

C. TermPedia Techniques

- (i) TermPedia linked only technical terms found in the IR e-book. []
- (ii) TermPedia linked all technical terms found in the IR e-book. []
- (iii) Finding technical terms definitions is fast with TermPedia. []
- (iv) TermPedia defined technical terms in the IR book correctly. []
- (v) Technical terms could be understood through their definitions. []
- (vi) TermPedia linked technical terms to relevant Wikipedia articles. []
- (vii) Definitions of the difficult terms helped in understanding the IR e-book. []
- (viii) Reading largely from Wikipedia about the difficult terms helped in understanding the IR e-book. []

D. Comments and Suggestions

Please give short useful comments and suggestions that you feel may be applied to improve these two features of TermPedia:

- (i) TermPedia User Interface
.....
.....
.....
- (ii) TermPedia Techniques
.....
.....
.....

A.3 r and α of Mean Scores for TermPedia User Survey

Section A: TermPedia User Interface

```
> cor(dat[,c("GA1", "GA2", "GA3", "GA4", "GA5")],
use="pairwise.complete.obs")
```

	GA1	GA2	GA3	GA4	GA5
GA1	1.00000000	0.1290994	0.15879200	0.02977457	0.55577200
GA2	0.12909944	1.0000000	0.00000000	0.38030437	0.00000000
GA3	0.15879200	0.0000000	1.00000000	-0.22821773	0.07163324
GA4	0.02977457	0.3803044	-0.22821773	1.00000000	-0.01141089
GA5	0.55577200	0.0000000	0.07163324	-0.01141089	1.00000000

```
# GA1 and GA5 have a high correlation coefficient
> cronbach(dat[,c("GA1", "GA5")]) # cronbach's alpha for scores in
Sec. A
$Alpha [1] 0.7023945
```

Where: GA1 to GA5 are mean scores of questions 1 to 5 in section A of questionnaire.

Section B: Document Difficulty

```
> cor(dat[,c("GB1", "GB2", "GB3", "GB4", "GB5")],
use="pairwise.complete.obs")
```

	GB1	GB2	GB3	GB4	GB5
GB1	1.00000000	0.2215143	0.08771182	-0.06590671	0.2557919
GB2	0.22151426	1.0000000	0.54199212	0.68011105	0.4903830
GB3	0.08771182	0.5419921	1.00000000	0.43983626	0.6402878
GB4	-0.06590671	0.6801110	0.43983626	1.00000000	0.6141412
GB5	0.25579186	0.4903830	0.64028777	0.61414123	1.0000000

```
GB2, GB3, GB4, and GB5 have high correlation coefficients

> cronbach(dat[,c("GB2", "GB3", "GB4", "GB5")])
$Alpha [1] 0.8442016
```

Where: GB1 to GB5 are mean scores of questions 1 to 5 in section B of questionnaire.

Section C: TermPedia Techniques

```
> cor(dat[,c("GC1", "GC2", "GC3", "GC4", "GC5", "GC6", "GC7",
"GC8", "GC9", "GC10")], use="pairwise.complete.obs")
```

	GC1	GC2	GC3	GC4	GC5
GC1	1.00000000	0.17052337	0.45848153	-0.06802648	0.16194149
GC2	0.17052337	1.00000000	-0.21765958	-0.14803625	-0.29174914
GC3	0.45848153	-0.21765958	1.00000000	0.37226460	0.69446598
GC4	-0.06802648	-0.14803625	0.37226460	1.00000000	0.66479420
GC5	0.16194149	-0.29174914	0.69446598	0.66479420	1.00000000
GC6	0.30878743	0.04784780	0.24480912	-0.18729016	-0.05057217
GC7	-0.04481107	0.05904185	0.07737306	0.22739926	0.13485913
GC8	0.28791446	0.10084855	0.40364717	0.37134416	0.38214458
GC9	0.21926450	0.23053940	0.41915367	0.39361409	0.70407124
GC10	0.08056974	0.11600794	0.11463549	0.13157474	0.10342572

	GC6	GC7	GC8	GC9	GC10
GC1	0.308787428	-0.04481107	0.2879145	0.219264505	0.08056974
GC2	0.047847800	0.05904185	0.1008485	0.230539399	0.11600794
GC3	0.244809123	0.07737306	0.4036472	0.419153670	0.11463549
GC4	-0.187290161	0.22739926	0.3713442	0.393614091	0.13157474
GC5	-0.050572174	0.13485913	0.3821446	0.704071244	0.10342572
GC6	1.000000000	0.53182160	-0.1198712	-0.006849315	0.23708852
GC7	0.531821600	1.00000000	0.1101087	0.219828893	0.08814765
GC8	-0.119871208	0.11010874	1.0000000	0.339799820	0.43627873
GC9	-0.006849315	0.21982889	0.3397998	1.000000000	0.05480794
GC10	0.237088516	0.08814765	0.4362787	0.054807941	1.00000000

High correlation coefficients among GC3, GC4, GC5, GC8 and GC9 and between GC6 and GC7

```
> cronbach(dat[,c("GC3", "GC4", "GC5", "GC8", "GC9")])
$Alpha [1] 0.8059006
```

```
> cronbach(dat[,c("GC6", "GC7")])
$Alpha [1] 0.6854839
```

Where: GC1 to GC10 are mean scores of questions 1 to 10 in section C of questionnaire answered by RUG students.

A.4 Reactions to TermPedia by RUG Students

TermPedia Feature	Category	Edited Comments and/or suggestions of user from Gulu Universiteit
User Interface	Pop-up	The size of the pop-up could auto grow. May be displaying the on click Wikipedia entry in a 'fancybox' layout would keep users from having to switch tabs
	Layout	<ul style="list-style-type: none"> - May be change the layout of the text of the book itself. For example it is now written in Times New Romans, this is not a screen font and makes it harder to read - It would help if the IR book and Wikipedia articles loaded on TermPedia user interface
	Links	The presentation is very basic, which is not bad in this case. personally like the style on the main page more (bold hyper-links instead of underlined); however it may be more consistent with web standards to underline links
	Others	<ul style="list-style-type: none"> - First of all I don't like reading educational literature from the screen. - So I'm a bit biased. While using it, I find it a bit annoying that terms and references are underscored and things pop-up. Call me old-fashion, but I just love to read from a book and feel the pages. - <i>Lets aantrekkelijker design. Zoekfunctie om te zoeken naar bepaalde definities.</i> [In English: Attractive design. Search Function to look for certain definitions]
TermPedia Techniques	Term prediction and definition	<ul style="list-style-type: none"> - Terms like 'document' isn't that relevant, may be some kind of stop list? I would like to add: Nice project! - Good, but not flawless. Some definitions are somewhat redundant ("document"), some are the same as exactly meant in the text - It would help if TermPedia linked only technical terms and linked these terms to the correct definitions and Wikipedia articles. A good place to look for the definitions may be in the IR e-book itself - I understand the text most of the time and when I don't understand a term its probably explained somewhere in the book. The advantage is that the explanation in the book is the one they want you to know. A wiki explanation can be a bit different. So I most of the time try to stick to information from the book itself.

Table A.1: Users comments and suggestions to improve TermPedia user interface and techniques

A.5 Test Scores of Gulu University Students
Gulu University Students Test Scores Before TermPedia

Group A		Group B	
Students Index	Test 1 Scores	Students Index	Test 1 Scores
1	16	16	15
2	15	17	16
3	16	18	15
4	16	19	15
5	16	20	16
6	15	21	15
7	16	22	15
8	15	23	15
9	14	24	16
10	15	25	15
11	15	26	15
12	15	27	16
13	15	28	15
14	15	29	15
15	16	30	15
Basic statistics for test 1 scores			
Median	15.00		15.00
Mean	15.33		15.27
sd(s) d	0.617		0.458

Table A.2: Students closed-book test scores before TermPedia

Gulu University Students Test Scores After TermPedia

	TermPedia Used (A)			TermPedia Not Used (B)		
	Time (Mins.)	Scores	%Scores	Time (Mins.)	Scores	%Scores
	35	13	65	40	19	95
	43	18	90	40	15	75
	53	19	95	53	06	30
	51	14	70	48	17	85
	48	16	80	43	14	70
	48	13	65	48	15	75
	48	17	85	53	17	85
	48	17	85	33	16	80
	48	18	90	48	17	85
	46	15	75	43	10	50
	46	16	80	43	15	75
	48	19	95	43	17	85
	47	20	100	43	15	75
	48	14	70	37	13	65
	46	16	80	43	16	80
Summaries						
Median	48.00	16.00	80.00	43.00	15.0	75.0
Mean	46.87	16.33	81.67	43.87	14.80	74.0
sd(s)	03.98	02.23	11.13	05.46	3.21	16.06

Table A.3: Students open-book test scores after using TermPedia

A.6 TermPedia User Study Questionnaire for GU

Your contribution to TermPedia user evaluation is much appreciated. Please provide the following information by completing this questionnaire. All information will be kept confidential and only used for the purpose of this user evaluation.

Personal Data

Age: []

Gender, (male [M]/female [F]): []

Student number: [.....]

Did you use TermPedia at all? (yes [Y]/no [N]): [] If

yes, on how many occasions?: []

Usability Data

On a scale of 1 to 5 give your impression on the following questions where:

- 1 ≡ Strongly agree
- 2 ≡ Agree
- 3 ≡ Somewhat agree
- 4 ≡ Disagree
- 5 ≡ Not applicable

A. TermPedia User Interface

- (i) It was easy to locate information on the interface. []
- (ii) It was easy to navigate through the interface. []
- (iii) The colors and design on the interface are appealing. []
- (iv) The font, size, and color of text on the interface are easy to read. []
- (v) I enjoy using the interface. []

B. Document Technical Difficulty

- (i) There were technical terms in the e-book chapter. []
- (ii) The technical terms were unfamiliar and not easy to understand. []
- (iii) It was not easy to understand text around difficult terms. []
- (iv) Material in the e-book chapter was written above my intellectual level. []
- (v) The e-book writer was bad at explaining things and boring. []

C. TermPedia Techniques

- (i) There was always a link for all difficult terms. []
- (ii) It takes a short time to find term definitions by TermPedia. []
- (iii) TermPedia always gives relevant definitions of terms. []
- (iv) Difficult terms could be understood through their definitions. []
- (v) Definition of difficult terms helped in understanding the e-book chapter. []
- (vi) Technical terms are always linked to relevant Wikipedia articles. []
- (vii) Understanding the e-book chapter was improved by reading from linked Wikipedia articles. []
- (viii) TermPedia always predicted correct difficult terms. []
- (ix) TermPedia can be a helpful tool for e-learning. []
- (x) It would help to illustrate the difficult terms through graphics. []

D. Comments and Suggestions

Please provide useful comments and suggestions that you feel may be applied to improve these two main features of TermPedia:

- (i) TermPedia User Interface
.....
.....
.....
- (ii) TermPedia Techniques
.....
.....
.....

A.7 Summary Scores for GU TermPedia User Study

Section	Qns.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	S.D.
A: TermPedia user interface	GA1	1.00	2.00	2.00	2.000	2.00	4.00	0.690
	GA1	1.00	2.00	2.00	2.000	2.00	4.00	0.690
	GA2	1.00	2.00	2.00	2.000	2.00	3.00	0.535
	GA3	1.00	2.00	2.00	2.227	3.00	4.00	0.869
	GA4	1.00	1.00	2.00	1.714	2.00	3.00	0.717
	GA5	1.00	1.00	2.00	1.773	2.00	4.00	0.869
	Avg. of Means (Sec. A)	1.00	1.60	2.00	1.943	2.20	3.60	0.736
B: Document difficulty	GB1	1.00	2.00	2.00	1.955	2.00	3.00	0.653
	GB2	1.00	3.00	3.00	3.095	4.00	4.00	1.044
	GB3	1.00	3.00	4.00	3.227	4.00	4.00	1.020
	GB4	1.00	2.75	4.00	3.200	4.00	4.00	1.056
	GB5	1.00	4.00	4.00	3.579	4.00	4.00	0.902
	Avg. of Means (Sec. B)	1.00	2.95	3.40	3.011	3.60	3.80	0.935
C: TermPedia techniques	GC1	1.00	1.00	1.50	1.727	2.00	4.00	0.883
	GC2	1.00	1.00	2.00	1.810	2.00	4.00	0.928
	GC3	1.00	1.00	2.00	1.727	2.00	3.00	0.631
	GC4	1.00	1.00	2.00	1.952	2.00	4.00	1.024
	GC5	1.00	1.00	2.00	1.905	2.00	4.00	0.831
	GC6	1.00	2.00	2.00	2.381	3.00	3.00	0.590
	GC7	1.00	2.00	2.00	2.286	3.00	3.00	0.717
	GC8	2.00	2.00	2.00	2.333	3.00	4.00	0.577
	GC9	1.00	1.00	1.00	1.455	2.00	3.00	0.671
	GC10	1.00	2.00	2.00	2.238	3.00	4.00	0.831
	Avg. of Means (Sec. C)	1.10	1.40	1.85	1.981	2.40	3.60	0.768

Table A.4: Sections A, B, and C summary scores for Gulu University user study questionnaire

Key for table A.4

- GA1: Locate information on interface
- GA2: Navigate through interface
- GA3: Colours and design of interface
- GA4: Legibility of text on interface
- GA5: Enjoy using the interface
- GB1: Document contains technical terms
- GB2: Unfamiliar terms in document
- GB3: Understand text around technical terms
- GB4: Document above reader intellect level
- GB5: Bad and boring document writer

- GC1: A link for each difficult term
- GC2: TermPeida finds term definition in short time
- GC3: Relevant term definitions
- GC4: Difficult term understood through definitions
- GC5: Definitions helpful for understanding
- GC6: Terms linked to relevant Wikipedia articles
- GC7: Reading Wikipedia articles improved understanding
- GC8: Correct definitions predicted
- GC9: TermPedia helpful for e-learning
- GC10: Illustrate terms by graphics

A.8 Comments and Suggestions for TermPedia by GU Students

TermPedia Feature	Category	Edited Comments and/or suggestions of user from Gulu University
User interface	Graphics and navigation	<ul style="list-style-type: none"> - The interface was easy to use and navigate - Easy to navigate - Needs simple navigation buttons and more graphics - Interface should be more interactive through the use of graphics. It looks beautiful - Add graphical illustrations - More animations would be appealing
	colour	<ul style="list-style-type: none"> - Improve interface colour - Colour should be appealing to the user - Add more colours for attractiveness. - Improve interface colour - Highlight strong points on interface with different colour
	Others	<ul style="list-style-type: none"> - Provide user interactions and allow for e-book downloads - Font should be improved because they are too tiny - Nice but improve on usability goals - The interface was nice - Friendly - The interface needs Improvement - Interesting and appealing but lacks prompts - Not appealing and not attractive
Techniques	Links	<ul style="list-style-type: none"> - Improve links - Needs improvement (include links to other learning sources and data dictionary) - Link technical terms to Wikipedia and other sources - Links to diagrammatic illustrations are lacking - Links had inadequate information about required information
	Terms	<ul style="list-style-type: none"> - Some difficult terms were not explained - Labour to provide definitions to technical terms - Gives relevant definition of terms - Include all technical terms - It is sufficient and good enough. System suffers from technical terms being defined by technical terms - Good reference mechanism. Difficult terms are sometimes linked to difficult terms
	Others	<ul style="list-style-type: none"> - Useful technique for e-Learning. It saves time and other resources - Add search functionality for ordinary users to make it simpler - Word search is good - Techniques should be simple for both wizards and ordinary users - The technique is tricky but not bad - Pop up menu for copying is lacking - Difficult to connect to TermPedia - Network was slow and not interactive - Technique used was not very good - Use simple English - Improve URL and make it short

Table A.5: Users comments and suggestions to improve TermPedia user interface and techniques

List of Abbreviations and Acronyms

BSc.CS	Bachelor of Science in Computer Science
BIT	Bachelor of Information Technology
DE	Document Enrichment
DSR	Design Science Research
GU	Gulu University
HTML	Hypertext Markup Language
HREF	Hypertext Reference
IR	Information Retrieval
IQ	Inter-Quantile
MWW	Mann-Whitney-Wilcoxon
PHP	Hypertext Preprocessor
Q-Q plots	Quantile-Quantile plots
URL	Uniform Resource Locator
UG	University of Groningen
WWW	World Wide Web

Summary

Each year hundreds and thousands of students in Uganda and many other countries around the world join universities. While at the university students have to read and quickly comprehend large volumes of course content [Hoeft, 2012; Fairbairn and Fairbairn, 2001] in order to accomplish various time bound assignments. Document content comprehension is an important component of reading without which the goal for learning from documents would be futile. Therefore, students at universities need to have strong comprehension skills in order to quickly understand the course contents they read. Comprehension skills help students to develop good qualities in decoding, analysing, and explaining text. The skills also help the student in expressing their own ideas about written materials [Gilakjani and Ahmadi, 2011]. Such qualities are important for furthering academic development. Since universities are higher institutions of learning, it is normally assumed that university students have sufficient reading comprehension skills for understanding their course content. However, there is evidence that a number of university students struggle to understand their course contents because of poor reading comprehension skills [Unruh, 2015; Shafie and Nayan, 2011; Roehl and Shiue, 2003]. Some indications of poor reading skills are when students have insufficient vocabulary and background knowledge in specific subject domains. We refer to domain specific vocabulary as technical terms. Such terms may make it difficult for students to understand their course content if they cannot recognize the terms or understand what the terms mean.

The problem of technical terms may be addressed using reader intuitions or by use of information technology. For example, when university students use their intuition to address the problem of technical terms, they may choose to predict the meaning of terms by considering the words that surround the terms in document content. Students may also try to find meanings of terms from reference sources like the Internet, dictionaries or encyclopedias.

from reference sources leads students away from document contents. Leaving the document contents disrupts the reading process and it requires more time and effort for students to comprehend their course contents. As students try to find term definitions on the internet, they may be lost in a “sea” of information and forget to return to reading the contents of their course documents.

Applying information technology to facilitate reading comprehension by addressing technical terms is more challenging compared to intuitive approaches. For example, supervised approaches of dealing with technical terms are labour intensive tasks because such approaches require complex mechanisms in natural language processing techniques such as knowledge bases, semantic feature representations, or semantic inference. Supervised methods generally have important drawbacks for instance, they can only be applied to a limited number of words for which sense-tagged data is available. Moreover, the accuracy of such methods depends on the availability of labelled data. Supervised methods also suffer from a data scarcity problem due to intense manual labour required in tagging words with senses. It is equally hard to acquire sufficient contextual information about senses of a large number of words in natural languages. Unsupervised methods tend to provide solutions for the drawbacks in supervised methods by not depending exclusively on large-scale resources which are manually annotated with word senses, but these incur other special difficulties. For instance, they do not use any dictionary and therefore cannot rely on a shared reference inventory of senses [Navigli, 2009].

Design science research was identified as an appropriate methodology for investigating a way to facilitate reading comprehension without distracting from the content of course documents. Design science research provided the tools for understanding the reading process and identifying factors that hinder reading comprehension. It also helped in identifying ways of addressing the factors that hinder reading comprehension, which provided functional requirements for a design artifact. The functional requirements pointed to a document enrichment approach as an artifact for addressing factors that affect reading comprehension without distracting from document contents.

To save the readers’ time and elude complex algorithms and intense labour for sense-tagging, our research set out to facilitate reading comprehension through this document enrichment approach. The approach aimed at providing in-text contextually relevant information for technical terms by using an integration of techniques from human language technologies and data from Wikipedia. Data from Wikipedia is used because it has more coverage compared to dictionaries. Moreover, the data in Wikipedia is constantly updated and corrected therefore, it captures recent developments in language. Some of the developments include new technical terms and their definitions. Wikipedia data is a rich source of information from which machine-readable dictionaries can be easily generated and used in recognizing and defining technical terms. The encyclopedia is also a good source of text cues such as background knowledge and pictorial illustrations for various technical terms. Note that, we do not directly deal with the complexities of finding appropriate pictorial illustrations for text, but we rely on a term sense

language technologies. The techniques include term prediction, term sense disambiguation, term definition and automatic hypertext generation. Term prediction helps in identifying technical terms which occur in document contents. This technique anticipates which words or group of words in document contents may present readers with unfamiliar vocabulary or jargon. The term sense disambiguation techniques resolves the meaning of predicted terms in contexts of the documents where the terms occur. This is important because many terms have multiple meanings. Resolving what the terms mean in specific document context supports the term definition technique in providing contextually relevant definitions for predicted technical terms. The automatic hypertext generation technique actualizes the document enrichment approach by using the predicted terms to create links that integrate information into document contents. The links created provide in-text term definitions and pictorial illustrations where possible. Thereby, enriching document contents with contextually relevant information about the predicted technical terms. Through document enrichment, readers have easy access to the definition, illustration and background information on technical terms. Thus, they do not have to leave contents of the documents they read in search for the meaning of difficult technical terms. Therefore, the reading comprehension process is facilitated and less time and effort may be used by readers in understanding document contents.

TermPedia document enrichment algorithms were designed to instantiate the techniques for facilitating reading comprehension that are based on human language technologies. TermPedia's document enrichment algorithms were tested and found to have a strong potential in providing contextually relevant information for technical terms. This in turn indicates that the algorithms can potentially facilitate readers in understanding document contents. The TermPedia approach was evaluated and the evaluation results show that the approach can be potentially useful and usable in facilitating the reading comprehension process of university students. Since insufficient vocabulary and background knowledge is common to people reading at all levels, the TermPedia approach can be used to facilitate reading comprehension for students in primary schools up till universities. The approach may also be used for readers outside the academic circle to facilitate the understanding of documents whose contents are cluttered with jargon. Ideas of the document enrichment techniques can be applied in computer assisted learning or eLearning for facilitating reading comprehension. The semantic web could use the automatic hypertext generation technique for generating anchor texts that link to web pages. Knowledge processing and information extraction applications could use features of the term prediction technique for improving text annotations.

There are many aspects of generalizability but the important ones are whether a research artefact can deliver the same results when used by different people and, what design and technology aspects of the artefacts can be changed to suite the generalized scenarios [Namatovu, 2018]. TermPedia was designed to help university students in their technical reading. However, other people can use this artifact with the same results. For example, document enrichment can provide practical solutions for facilitating reading comprehension for learners in secondary schools. TermPedia can

Samenvatting

Ieder jaar gaan er in Oeganda en veel andere landen ter wereld duizenden studenten naar de universiteit. Daar moeten de studenten snel grote hoeveelheden studiemateriaal lezen en kunnen begrijpen [Hoeft, 2012; Fairbairn en Fairbairn, 2001] om de deadlines van hun opdrachten te halen. Tekstbegrip is een belangrijk onderdeel van leesvaardigheid. Zonder dat zou het leren via teksten zinloos zijn. Daarom moeten de studenten aan de universiteit zeer begripsvaardig zijn zodat ze snel het studiemateriaal tot zich kunnen nemen. Begripsvaardigheid helpt studenten om teksten goed te kunnen doorgronden, analyseren en verklaren. Het helpt hen ook om zich over een tekst te kunnen uitspreken [Gilakjani en Ahmadi, 2011]. Dergelijke vaardigheden zijn belangrijk voor de verdere academische ontwikkeling. Aangezien universiteiten instellingen voor hoger onderwijs zijn wordt gewoonlijk aangenomen dat studenten over voldoende leesvaardigheid beschikken om hun studiemateriaal te begrijpen, echter blijkt dat dit voor sommige studenten niet het geval is [Unruh, 2015; Shafie en Nayan, 2011; Roehl en Shiue, 2003]. Een ontoereikende woordenschat en achtergrondkennis over specifieke onderwerpen kunnen op een matige leesvaardigheid duiden. Wij gebruiken voor domeinspecifieke woordenschat de omschrijving technische termen. Als studenten dergelijke termen niet kunnen herkennen of interpreteren, dan is het lastig om het studiemateriaal te begrijpen.

Het probleem van technische terminologie kan worden aangepakt door middel van intuïtief lezen of met behulp van informatietechnologie. Als studenten technische termen bijvoorbeeld intuïtief benaderen, kunnen ze de betekenis ervan voorspellen door in het document naar de woorden te kijken die om een technische term heen staan. Ze kunnen ook de betekenis proberen te achterhalen uit naslagwerken, zoals het internet, woordenboeken of een encyclopedie. Het zoeken naar de betekenis van een technische term in een naslagwerk voert de student weg van waar een document over gaat. Dit verstoort het leesproces en het kost de student meer tijd en moeite om het studiemateriaal te begrijpen. Wanneer een student de betekenis van een term op internet opzoekt, kan hij of zij verstrikt raken in een woud van informatie en daardoor vergeten terug te gaan naar de inhoud van het studiemateriaal.

De leesvaardigheid met betrekking tot technische termen vergroten met behulp van informatietechnologie is ingewikkelder dan met een intuïtieve aanpak. Zo is de gecontroleerde benadering van technische termen een arbeidsintensieve taak, omdat zo'n benadering complexe technieken van natuurlijke taalverwerking vereisen, zoals het gebruik van kennisbanken, semantische kenmerken of semantische interferentie. Gecontroleerde methoden hebben vaak belangrijke nadelen omdat ze bijvoorbeeld alleen maar kunnen worden gebruikt voor een beperkt aantal woorden waarvoor betekenistags beschikbaar zijn. De juistheid van dergelijke methoden hangt bovendien af van de aanwezigheid van gelabelde data. Gecontroleerde methoden hebben ook te kampen met een tekort aan data als gevolg van het vele handmatige werk dat nodig is om betekenistags aan woorden toe te kennen. Net zo lastig is het om voor de betekenissen van een groot aantal woorden in natuurlijke talen voldoende contextuele informatie te verzamelen. Ongecontroleerde methoden zouden tegemoet kunnen komen aan deze tekortkomingen doordat zij niet uitsluitend afhankelijk zijn van grootschalige bronnen waarvan de woorden handmatig van betekenistags zijn voorzien, maar deze methoden kennen weer andere problemen. Zo maken ze bijvoorbeeld geen gebruik van woordenboeken en kunnen zich daardoor niet verlaten op een gedeelde referentielijst van betekenissen [Navigli, 2009].

Design science research werd gezien als een passende methode om te onderzoeken hoe leesvaardigheid ondersteund kan worden zonder afleiding van de inhoud van studiemateriaal te veroorzaken. Design science research bood de hulpmiddelen om het leesproces te doorgronden en vast te stellen welke factoren de leesvaardigheid in de weg staan. Het heeft ook geholpen bij de aanpak van deze factoren, hetgeen functionele eisen voor het ontwerp van een ontwerp-artefact heeft opgeleverd. De functionele eisen wezen op een documentverrijkingsaanpak als een artefact waarmee factoren kunnen worden aangepakt die de leesvaardigheid beïnvloeden zonder afleiding van de inhoud te veroorzaken.

Teneinde lezers tijd te besparen, en om complexe algoritmen en het arbeidsintensief taggen van betekenissen te vermijden, wilden we in ons onderzoek met een documentverrijkingsaanpak de leesvaardigheid verbeteren. De aanpak richtte zich op het geven van contextrelevante informatie bij technische termen binnen een tekst door technieken uit natuurlijke taaltechnologieën met data van Wikipedia te combineren. Er is voor Wikipedia gekozen omdat het meer gebieden bestrijkt, vergeleken met woordenboeken. Bovendien wordt Wikipedia steeds bijgehouden en gecorrigeerd waardoor het de nieuwste ontwikkelingen op taalgebied weergeeft. Een aantal van deze ontwikkelingen betreft nieuwe technische termen en de definities ervan. Wikipedia is een rijke informatiebron waarmee eenvoudig door computers leesbare woordenlijsten kunnen worden gegenereerd en gebruikt voor het herkennen en definiëren van technische termen. De encyclopedie is ook een goede bron voor tekstsuggesties voor bijvoorbeeld informatie over en illustraties van diverse technische termen. We pakken op deze manier niet direct het vraagstuk van het vinden van geschikte illustraties aan, maar we verlaten ons op een algoritme dat polysemie vermijdt om zo relevante definities en illustraties bij de technische termen in een document te vinden.

De grootste bijdrage van dit onderzoek bestaat uit TermPedia, een documentverrijkingsaanpak die werd ontwikkeld om leesvaardigheid te vergemakkelijken.

Deze techniek anticipeert op woorden en woordgroepen die voor de lezer onbekende betekenissen of jargon kunnen bevatten. Ambigüiteitsvermijding bepaalt de betekenis van voorspelde termen op basis van de context van de documenten waar deze termen in voorkomen. Dat is belangrijk omdat veel termen meerdere betekenissen kunnen hebben. Het bepalen van de betekenis van een woord in een document ondersteunt de definitievorming door bij voorspelde technische termen binnen de context relevante definities aan te bieden. Automatische generatie van hypertext actualiseert de documentverrijkingsaanpak door voor de voorspelde termen hyperlinks te maken die informatie in een document integreren. Waar mogelijk geven deze links binnen de tekst een definitie of illustratie weer. Daardoor wordt de inhoud van een document verrijkt met informatie die binnen de context relevant is. Via documentverrijking heeft de lezer eenvoudig toegang tot de definitie van een technische term, een illustratie ervan of achtergrondinformatie erover. Zo hoeft de lezer de tekst niet te verlaten om de verklaring van een moeilijke term te kunnen vinden. Dat bevordert de leesvaardigheid en zorgt ervoor dat de lezer sneller en makkelijker de inhoud van een document begrijpt.

De tekstverrijking algoritmen van TermPedia zijn ontworpen om leesvaardigheidstechnieken die op natuurlijke taaltechnologieën zijn gebaseerd te concretiseren. De tekstverrijking algoritmen van TermPedia zijn getest en blijken een krachtig instrument te zijn om bij technische termen informatie aan te reiken die binnen de context relevant is. Dit wijst er dan weer op dat de algoritmen lezers kunnen helpen bij tekstbegrip. De werking van TermPedia is geëvalueerd en de testresultaten laten zien dat de methode zeer nuttig en bruikbaar kan zijn voor de leesvaardigheid van studenten. Een ontoereikende woordenschat en achtergrondkennis speelt een rol bij lezers op elk niveau en daarom kan TermPedia de leesvaardigheid vergroten voor leerlingen vanaf het basisonderwijs tot studenten aan de universiteit. TermPedia kan ook buiten de academische setting van nut zijn bij het lezen van een tekst die volstaat met jargon. De principes van documentverrijking kunnen worden toegepast bij computerondersteund leren of eLearning om leesvaardigheid te vergemakkelijken. Het semantische web zou geautomatiseerde generatie van hypertext kunnen gebruiken voor tekstblokjes die naar webpagina's verwijzen. Toepassingen voor kennisverwerking en informatie-extractie kunnen elementen van tekstvoorspelling gebruiken om tekstannotaties te verbeteren.

Generaliseerbaarheid kent veel aspecten, maar de belangrijkste zijn of een onderzoeksartefact dezelfde resultaten kan opleveren als hij door andere personen wordt gebruikt en welke aspecten van ontwerp en technologie van de artefacten veranderd zouden kunnen worden om aan algemene scenario's te voldoen [Namatovu, 2018]. TermPedia is ontworpen om studenten te helpen technische teksten te doorgronden. Ook anderen kunnen dit artefact met dezelfde resultaten gebruiken. Documentverrijking kan bijvoorbeeld praktische oplossingen bieden bij het verhogen van de leesvaardigheid van scholieren in het voortgezet onderwijs. TermPedia kan worden gebruikt voor het voorspellen en definiëren van technische termen die in hun schoolboeken voorkomen.

Curriculum Vitae

Proscovia Olango was born on 22nd November, 1973 in Gulu district, Uganda. She obtained her Bachelor's degree in Library and Information Science from Makerere University in 1998. In 2006, she obtained a Master of Arts in Information Science from The University of Groningen in The Netherlands. During the pursuit of her masters, she worked on events and relations extraction for a semantic web project on early American history. Proscovia is currently a full-time lecturer in the Department of Computer Science at Gulu University, Uganda.