

A CORPUS-BASED SEMANTIC ANALYSIS OF SEMI-TECHNICAL VOCABULARY IN THE FIELD OF MEDICINE



A thesis submitted in fulfilment of the requirements for the degree of
DOCTOR OF PHILOSOPHY
School of Education
Faculty of Arts, Business, Law, and Economics
The University of Adelaide

Chinh Ngan Nguyen Le

A1690064

Supervisors:

Dr Julia Miller

Dr Stephen Kelly

A/Prof Edward Palmer

November 2023

THESIS DECLARATION

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

The author acknowledges that copyright of published works contained within the thesis resides with the copyright holder(s) of those works.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Chinh Ngan Nguyen Le

November 2023

ABSTRACT

The multiplicity of meaning has long been a central issue in lexical semantics, lexicography and corpus linguistics. In lexical semantics, multiple meanings contribute to lexical ambiguity—polysemy and homography, where a word form has multiple related and unrelated meanings. Semantic studies have yielded contradictory findings about methods of distinguishing polysemy and homography and approaches to the mental representation of polysemy. In lexicography, there has been little agreement on distinguishing and presenting different senses of the same polysemous word and its homographs, and the conventional format of numbered word senses in dictionary entries does not fully depict relationships between polysemous words, potentially confusing users. In corpus linguistics, polysemy and homography present a challenge to the task of word sense disambiguation (WSD) in corpus-derived wordlists. Continuous attempts have been made to achieve more precise and satisfactory corpus-derived outcomes when assigning appropriate senses to given words. This thesis revisits multi-meaning challenges in lexical semantics, lexicography and corpus linguistics. The focus is on semi-technical medical vocabulary, which has yet to be adequately addressed in dictionaries and corpus-derived wordlists. The study is divided into three journal articles.

Paper 1 examines Hsu's (2013) 595-word Medical Word List (MWL) and uncovers the lack of indication of polysemy and homography resulting from its corpus-based automatic analyses of semi-technical medical word forms regardless of their meanings. The examination of the MWL entailed a core meaning-based analysis which reconciles different lexical semantic methods (etymology and native speaker judgement) and approaches (monosemy and polysemy) to identify and distinguish polysemes and homographs. The examination of the MWL resulted in 302 words whose polysemes and homographs are anticipated to pose pedagogical difficulties.

Paper 2 presents SemiMed, a lexicographic resource for semi-technical medical vocabulary, as an alternative to word form-based lists. SemiMed is based on a semantic analysis underpinned by lexical semantic theories (Lakoff's (1987) radial categories and Tyler and Evans's (2004) Principled Polysemy). A corpus-based analysis employs the WSD method of one-sense-per-collocation to validate the semantic analysis. SemiMed's headword templates transfer corpus-based WSD results into semantic networks, thus visualising relations between polysemes and homographs in MWL words. This non-conventional format aims to minimize confusion associated with the traditional entry-structured layout.

Paper 3 analyses SemiMed's practicality and usefulness in a pilot study in which a 40-word e-version of SemiMed was introduced to 18 medical students with English as a foreign language to use while role playing medical scenarios. Student participants' feedback on their experience in using SemiMed was gathered through focus groups. The students' preference for SemiMed over conventional dictionary entries highlighted the benefits of SemiMed's non-conventional format in facilitating the understanding of polysemy and homography in semi-technical medical vocabulary.

The main implication of the study's findings is that some of the challenges of learning and teaching words with multiple meanings may be resolved by an interdisciplinary approach, where theoretical and methodological frameworks from lexical semantics and corpus-based WSD inform lexicographic practices to better manage polysemy and homography.

ACKNOWLEDGEMENT

I commenced my PhD just before the onset of the Covid-19 pandemic and went through an unprecedented period full of challenges and uncertainties. I still find it hard to believe that I have made it to the finish line. I would like to express my deepest gratitude to all those individuals whose support and guidance enabled me to complete my PhD despite unforeseen obstacles that arose.

First and foremost, I am indebted to my principal supervisor, Dr. Julia Miller, who spent lots of time giving enlightened comments and constructive suggestions for every piece of my work. I most heartily appreciate her professional guidance and affectionate encouragement during the writing of this thesis. Without her support, I might never have completed my work.

I would also like to thank my co-supervisors, Dr. Stephen Kelly and Associate Professor Edward Palmer, for their immeasurable contributions to enriching the quality of my work. Thanks to Stephen's thought-provoking questions, my arguments were considerably sharpened. Edward's timely and invaluable feedback played a pivotal role in refining my academic writing.

I am grateful to the A.S. Hornby Educational Trust for granting me an A.S. Hornby Dictionary Research Award (ASHDRA). Without their financial support, a pilot phase of this study would not have been possible. I highly value their generosity in offering me an opportunity to present my research at the 20th EuraLex International Congress.

A special appreciation goes to the ASHDRA panel of experts, Dr. Michael Rundell, Professor Hilary Nesi and Ms. Julie Moore, who provided their valuable expertise and guidance for the funded pilot study from the initial stage of the funding proposal to the final submission of the report.

I would like to extend my appreciation to Hugh Dellar and Andrew Walkley, founders of Lexical Lab, who kindly sponsored me for my presentation at the 56th IATEFL Conference,

and my thanks to the AustraLex Committee, who kindly sponsored me for my presentation at the AustraLex 2021 Conference.

I am also thankful to Dr. Liz Pridham for the time and effort she devoted to part of the data analysis and to Ms. Maureen McInroy for her proofreading and feedback on previous drafts of the thesis.

Thanks should go to Ms. Maureen Goldfinch, my research assistant, and IT support staff from a University of Medicine and Pharmacy in Vietnam (UMP), who designed electronic formats of SemiMed headwords and set up the Moodle platform.

I would like to acknowledge the Associate Head and officers of the Department of Science, Technology and International Relations, UMP, who helped me liaise with participants, and all 18 UMP medical students who made themselves available after hours for Zoom meetings.

Last but not least, I am grateful to my beloved family, officemates, senior colleagues and friends who gave me so many best wishes, sound comments, enormous motivation and inspiration to keep me on the right track and complete my PhD thesis.

LIST OF ABBREVIATIONS

AWL: Academic Word List

BNC: British National Corpus

CEFR: Common European Framework of Reference for Languages

CF: Consent form

COCA: Corpus of Contemporary American English

DSTIR: Department of Science, Technology and International Relations

GSL: General Service List

EAP: English for Academic Purposes

EFL: English as a foreign language

EMP: English for Medical Purposes

ESL: English as a second language

ESP: English for Specific Purposes

IE: Invitation email

IELTS: International English Language Testing System

LC: Lexical Constellation

L1: First language

MWL: Medical Word List

OED: Oxford English Dictionary

PIS: Participant Information Sheet

TOEFL: Test of English as a Foreign Language

UMP: University of Medicine and Pharmacy

WSD: Word sense disambiguation

LIST OF TABLES

Table 2.1 An excerpt of the decision list for <i>Plant</i> (Adapted from Yarowsky, 1995, p. 191)..	60
Table 3.1 Thirty-one medical sub-disciplines covered by Hsu's MWL (2013).....	78
Table 3.2 Targeted categories of collocates in relation to their node's parts of speech	86
Table 3.3 Demographic profile of participants	92
Table 4.1 Development of the Medical Academic Word List and Medical Word List	104
Table 4.2 Wang and Nation's (2004, p. 302) scale of semantic relatedness for <i>Issue</i>	107
Table 4.3 Summarized interpretation of meanings for the word <i>Issue</i> using Wang and Nation's scale of semantic relatedness	108
Table 4.4 Evaluator details.....	111
Table 5.1 Details of the development of Hsu's (2013) Medical Word List	141
Table 5.2 Description of English Web 2020 and Medical Web Corpus.....	145
Table 5.3 Top 15 most frequent collocates and meanings of <i>Defect</i> (n) in English Web 2020	146
Table 5.4 Top 15 most frequent collocates and meanings of <i>Defect</i> (n) in Medical Web Corpus	146
Table 5.5 Top 15 most frequent collocates and meanings of <i>Defect</i> (v) in English Web 2020	147
Table 5.6 Descriptors of four technicality levels	149
Table 5.7 Forty sampled words used to develop SemiMed	152
Table 6.1 Details of English Web 2020 and Medical Web Corpus	171
Table 6.2 Technicality level description	175

LIST OF FIGURES

Figure 2.1 A semantic network of <i>Time</i> (Adapted from Evans, 2005, p. 52)	24
Figure 2.2 The two-phase process of compiling a corpus-based monolingual dictionary (Adapted from Atkins & Rundell, 2008, p. 98)	30
Figure 2.3 Three components in the entry for <i>Naked</i> . Definitions from <i>Collins English Dictionary</i> in the order in which they appear.....	32
Figure 2.4 Johnson’s (1755) entry for <i>Resound</i> as cited in Atkins and Rundell (2008, p. 271)	37
Figure 2.5 The entry for <i>Keen</i> with numbered senses (Adapted from Atkins & Rundell, 2008, p. 272)	38
Figure 2.6 The grammar-led (left) and meaning-led (right) entries for <i>Haunt</i> (Adapted from Atkins & Rundell, 2008, p. 247).....	38
Figure 2.7 Entries for <i>Necessary</i> in tiered (left) and flat (right) structures (Adapted from Atkins & Rundell, 2008, p. 250).....	40
Figure 2.8 The entry for <i>Icon</i> with frequency-ordered senses. Definitions from <i>Longman Dictionary of Contemporary English</i>	40
Figure 2.9 The entry for <i>Icon</i> with core meaning first. Definitions from <i>Oxford Dictionary of English</i>	41
Figure 2.10 Homograph entries for <i>Bear</i> . Definitions from <i>Collins English Dictionary</i>	43
Figure 2.11 The entry for <i>Green</i> . Definitions from the <i>Oxford English Dictionary</i>	45
Figure 2.12 A general model of WSD (Adapted from Kwong, 2013, p. 16).....	54
Figure 3.1 The three-stage research procedure	72
Figure 3.2 OED definitions of <i>Defect</i> (n)	83
Figure 3.3 OED definitions of <i>Defect</i> (v)	83
Figure 3.4 OED definitions of <i>Defect</i> (adj)	83
Figure 3.5 A generic microstructure in SemiMed.....	84
Figure 3.6 Sketch Engine-generated collocates for <i>Defect</i> (n) in English Web 2020.....	86
Figure 3.7 Concordance function in Word Sketch.....	88
Figure 3.8 Concordance view of collocate <i>Birth</i> and its node in English Web 2020	88
Figure 3.9 Participant recruitment procedure	91
Figure 4.1 The word entry <i>Fistula</i> in the OED.....	112
Figure 4.2 The semantic re-evaluation of <i>Fistula</i>	112
Figure 4.3. Research procedure	114

Figure 4.4 The evaluation of <i>Primary</i>	116
Figure 4.5 The evaluation of <i>Acute</i>	117
Figure 4.6 The evaluation of <i>Cataract</i>	118
Figure 4.7 The evaluation of <i>Plasma</i>	118
Figure 4.8 The evaluation of <i>Secrete</i>	119
Figure 4.9 The evaluation of <i>Resolve</i>	123
Figure 4.10 The evaluation of <i>Disorder</i> (n, v).....	125
Figure 4.11 An example of the core meaning of <i>Benign</i> for in-class teaching.....	127
Figure 4.12 The first seven examples of <i>Benign</i> in SKELL (2014-2021).....	127
Figure 5.1 Cantos et al.'s (2009) illustration of word meaning disambiguation based on collocations.....	137
Figure 5.2 Generic pattern of an LC (Adapted from Rizzo & Sanchez, 2010, p. 110).....	139
Figure 5.3 An LC of <i>Heart</i> (Adapted from Rizzo & Sanchez, 2010, p. 112).....	139
Figure 5.4 An excerpt from Hsu's (2013) Medical Word List.....	140
Figure 5.5 OED definitions of <i>Defect</i> used in Le and Miller (2023).....	142
Figure 5.6 Simplified OED definitions of <i>Defect</i>	142
Figure 5.7 Core and other related meanings of <i>Defect</i> used by Le and Miller (2023).....	143
Figure 5.8 Description of <i>Defect</i> resulting from the qualitative analysis.....	143
Figure 5.9 LCs of <i>Defect</i> resulting from the qualitative analysis.....	144
Figure 5.10 Description of <i>Defect</i> resulting from the quantitative analysis.....	150
Figure 5.11 LCs of <i>Defect</i> resulting from the quantitative analysis.....	150
Figure 5.12 A generic LC for single-meaning words.....	153
Figure 5.13 An example LC of a single-meaning word.....	154
Figure 5.14 A generic LC for multi-meaning words with a single core meaning.....	154
Figure 5.15 An example LC of a multi-meaning word with a single core meaning.....	155
Figure 5.16 A generic LC for multi-meaning words with more than one core meaning.....	155
Figure 5.17 An example LC of a multi-meaning word with more than one core meaning ...	156
Figure 5.18 An LC of <i>Primary</i>	157
Figure 5.19 <i>Primary</i> in <i>Cambridge Dictionary</i> (http://dictionary.cambridge.org).....	158
Figure 6.1 The linear structure of <i>Colon</i> and <i>Benign</i> . Definitions from <i>Cambridge Dictionary online</i> (https://dictionary.cambridge.org/) in the order in which they appear.....	168
Figure 6.2 The hierarchical structure of <i>Colon</i> and <i>Benign</i> . Definitions from <i>Merriam-Webster Dictionary online</i> (https://www.merriam-webster.com/) in the order in which they appear.....	169

Figure 6.3 Generic pattern of LCs of a polysemous word and a homograph (Adapted from Rizzo & Sanchez, 2010)	174
Figure 6.4 Procedural demonstration of <i>Diffuse</i>	177
Figure 6.5 Two homographs	179
Figure 6.6 A polysemous word	178
Figure 6.7 A polysemous word and a homograph.....	178
Figure 6.8 A Moodle interface of the LC of <i>Benign</i> (in Book 1: SemiMed A-C)	180
Figure 6.9 A pop-up box indicating the detailed technicality level	180
Figure 6.10 <i>Lobe</i> in the three dictionary formats.....	187

TABLE OF CONTENTS

A CORPUS-BASED SEMANTIC ANALYSIS OF SEMI-TECHNICAL VOCABULARY IN THE FIELD OF MEDICINE.....	i
LIST OF TABLES	viii
LIST OF FIGURES	ix
CHAPTER 1: INTRODUCTION	1
1.1 Background and rationale of the research	1
1.2 Statement of the problem	3
1.3 Scope of the research.....	4
1.4 Aims of the research.....	6
1.5 Research questions	7
1.6 Research significance.....	7
1.7 Thesis structure	8
CHAPTER 2: LITERATURE REVIEW	12
PART 1 – LEXICAL SEMANTICS	12
2.1 Key concepts	12
2.2 Lexical ambiguity.....	13
2.2.1 Homography and polysemy	13
2.2.2 Sources of homography and polysemy	13
2.2.3 Distinctions between homography and polysemy	16
2.3 Approaches to polysemy	18
2.3.1 The monosemy approach	18
2.3.2 The polysemy approach	19
PART 2 – LEXICOGRAPHY.....	30
2.4 Key concepts	30
2.5 Dictionary word senses	33
2.5.1 Dictionary word sense distinctions	34
2.5.2 Dictionary word sense presentations	36
2.6 Lexical semantics and lexicography	47
PART 3 – WORD SENSE DISAMBIGUATION AND CORPUS LINGUISTICS.....	53
2.7 Key concepts	53

2.8 WSD approaches	55
2.8.1 Methods of WSD	55
2.8.2 One-sense-per-collocation method	58
2.8.3 Problems in WSD	60
2.9 WSD and corpus-derived wordlists.....	66
CHAPTER 3: METHODOLOGY	72
3.1 Theoretical framework	72
3.1.1 Homography and polysemy distinctions.....	73
3.1.2 Word sense distinctions.....	74
3.1.3 Word sense presentations	75
3.1.4 Word sense disambiguation and word meaning frequency.....	76
3.2 The examination of Hsu's (2013) Medical Word List (MWL)	77
3.2.1 A brief description of Hsu's MWL.....	77
3.2.2 A source of semantic input for the examination of Hsu's MWL	78
3.2.3 A method for the examination of Hsu's MWL.....	79
3.3 The development of SemiMed	81
3.3.1 Semantic analysis.....	82
3.3.2 Corpus-based analysis.....	85
3.4 The piloting of SemiMed	89
3.4.1 Ethics approvals	90
3.4.2 Participant recruitment.....	90
3.4.3 Data collection	93
3.4.4 Data analysis	94
3.5 Methodological limitations	95
CHAPTER 4: THE EXAMINATION OF HSU'S (2013) MEDICAL WORD LIST	96
A core meaning-based analysis of English semi-technical vocabulary in the medical field	97
4.1 Introduction	97
4.2 Literature review	98
4.2.1 Nation's lexical categories	98
4.2.2 Semi-technical vocabulary.....	100

4.2.3 Wordlists of semi-technical vocabulary in the medical field	103
4.3 The study	105
4.3.1 Aims and research questions	105
4.3.2 Methodological framework.....	105
4.3.3 Research procedures	110
4.4 Results and discussion.....	115
4.4.1 Examining the boundary of semi-technical medical vocabulary	115
4.4.2 Identifying semi-technical medical vocabulary with multiple meanings	120
4.4.3 Disadvantages of the word form frequency-based list of semi-technical medical vocabulary	123
4.5 Potential implications of the research results for teaching semi-technical vocabulary	125
4.6 Limitations and directions for future research	127
4.7 Conclusion.....	128
CHAPTER 5: THE DEVELOPMENT OF SEMIMED	130
Developing a pilot version of semimed—A corpus-based resource of semi-technical medical words.....	131
5.1 Introduction	131
5.2 Literature review	132
5.2.1 Wordlists	132
5.2.2 Word sense disambiguation	134
5.2.3 Lexical constellations.....	137
5.3 The study	140
5.3.1 Qualitative analysis	141
5.3.2 Quantitative analysis	144
5.4 Findings and discussion	151
5.4.1 Selection of words from Hsu’s (2013) MWL to create a pilot version of SemiMed	151
5.4.2 SemiMed template	153
5.4.3 Pedagogical potential of SemiMed	156
5.5 Future work and conclusion	160
CHAPTER 6: THE SEMIMED PILOT	162

Report on A.S. Hornby Dictionary Research Award Project.....	163
6.1 Project summary.....	163
6.2 Background and objectives	164
6.2.1 Statement of research problem.....	164
6.2.2 Literature review	165
6.3 Description of research.....	169
6.3.1 Developing SemiMed	170
6.3.2 Piloting SemiMed	177
6.4 Results and evaluation.....	182
6.4.1 Conventional resources	182
6.4.2 SemiMed	185
6.4.3 Suggestions for future improvement.....	192
6.5 Overall reflections and future plans	193
CHAPTER 7: CONCLUSION	194
7.1 Summary	194
7.2 Implications.....	197
7.2.1 Lexical semantics.....	197
7.2.2 Lexicography	198
7.2.3 Corpus linguistics.....	199
7.2.4 Learning and teaching semi-technical medical vocabulary	200
7.3 Recommendations for future research.....	202
REFERENCES	204
APPENDICES	223
APPENDIX 1. Core meanings of 302 potentially confusing semi-technical medical words organised by frequency in Hsu's (2013) Medical Word List.....	223
APPENDIX 2. Qualitative and quantitative analysis results of 40 sampled words	240
APPENDIX 3. A pilot version of SemiMed (40 Lexical Constellations)	248
APPENDIX 4. Ethics approval (The University of Adelaide).....	288
APPENDIX 5. Ethics approval (A University of Medicine and Pharmacy in Vietnam – name deleted for anonymity).....	290

APPENDIX 6. Invitation email to rector (A University of Medicine and Pharmacy in Vietnam).....	291
APPENDIX 7. Invitation email to participants (A University of Medicine and Pharmacy in Vietnam).....	292
APPENDIX 8. Participant information sheet.....	293
APPENDIX 9. Consent form	298
APPENDIX 10. Medical scenarios	300
APPENDIX 11. Focus group questions	304

CHAPTER 1: INTRODUCTION

1.1 Background and rationale of the research

A word with multiple meanings is viewed as a common phenomenon in a language (Palmer, 1995). This phenomenon, often known as lexical ambiguity, is central to lexical semantics. Two types of lexical ambiguity—homography (e.g., *bank* as “financial institution” and “edge of river”) and polysemy (e.g., *bank* as “financial institution” and “bank of blood”)—have been perceived in lexical semantics as intricately intertwining with each other and typically manifesting the fuzzy nature of word meanings. They have thus posed long-standing challenges to lexical semanticists. Lexical studies have yet to agree upon a method that can draw sharp distinctions between homography and polysemy (Klepousniotou, 2002; Lehrer, 1974; Lyons, 1968, 1977; Panman, 1982), and there is much controversy around mental representations of polysemous words (Cruse, 1992; Janssen, 2003; Murphy, 2010).

Homography and polysemy have added a significant challenge to lexicography in terms of word sense distinctions and presentations. The fuzziness of polysemous meanings has challenged a lexicographic approach to word sense distinctions in which lexicographers usually rely on their intuition in clustering overlapped meanings into discrete senses, capturing polysemous senses inconsistently (Atkins et al., 2003; Ayto, 1983; Grefenstette & Hanks, 2023; Hanks, 1990; Kilgarriff, 1997, 2007; Van der Meer, 2004). This challenge has prevented the conventional presentation format, where word senses are neatly numbered and vertically listed under dictionary entries, from consistently representing the multidimensional structure of word meanings (Aitchison, 2003; Atkins & Rundell, 2008; Geeraerts, 2001; Hanks, 2000; Ostermann, 2015; Stock, 2008). Additionally, distinctions between homography and polysemy have remained in the grey area where lexicographers may decide not to explicitly signal relations (homography and/or polysemy) among meanings of a word in its dictionary entry (Atkins & Rundell, 2008; Moon, 1987).

Current computational, corpus-based word sense disambiguation (WSD) methods have yielded high-performance results in distinguishing meanings of homographs but not of polysemes (Edmonds, 2006). The constraints on WSD raised by homography and polysemy have therefore limited the effectiveness of corpus-derived, frequency-based wordlists. These lists often enumerate words which are frequently found in targeted contexts (corpora) and are usually multi-meaning (Dash, 2012; Todd, 2017). Because of its dependence on word frequency which is computed according to word forms rather than word meanings, the development of frequency wordlists has not adequately implemented the task of WSD and has attracted criticism for overlooking polysemy and homography (Gardner, 2007). The evaluation of these wordlists has mainly been conducted at the level of homographs alone, due to the lack of a robust WSD method that considers the fuzziness of word meanings and permits the disambiguation of polysemous senses (Kwong, 2013; Mihalcea, 2006).

These interdisciplinary challenges of homography and polysemy exist in learning and teaching all types of vocabulary, including that which is the subject of this thesis—semi-technical medical vocabulary. Located between general and technical vocabulary, semi-technical vocabulary is considered elusive because it carries both general and specialized meanings (Cowan, 1974; Flowerdew, 1993; Huizhong, 1986). In English for medical purposes (EMP), a word such as *defect*, for example, has polysemes (e.g., something that is not perfect, something wrong with part of the body) and homographs (e.g., to leave and join the other side) that can be general (e.g., something that is not perfect, to leave and join the other side) or medical (e.g., something wrong with part of the body), depending on context. This characteristic makes semi-technical medical vocabulary hard to learn and teach (Li & Pemberton, 1994; Shaw, 1991; Thurston & Candlin, 1998).

Because it is situated in the grey area between general and medical vocabulary, semi-technical medical vocabulary is not well treated in general and medical dictionaries, which tend

to focus on either general or medical meanings, respectively. Even if both general and medical meanings are present in a dictionary, the homographic and polysemous relations between them are not explicitly shown, since the dictionary's entry structure, as mentioned previously, has limited capacity to effectively capture both homography and polysemy (Aitchison, 2003; Ostermann, 2015).

The elusiveness of semi-technical medical vocabulary suggests the need to create its own wordlist. While wordlists of other types of vocabulary have thrived (e.g., West's (1953) General Service List, Brezina and Gablasova's (2017a) New General Service List, Browne's (2014) New General Service List, and Coxhead's (2000) Academic Word List), semi-technical medical wordlists remain relatively small in number. Hsu's (2013) Medical Word List (MWL) stands out as the most recent, well-designed list. However, the list suffers from a dearth of information on homography and polysemy as, like other corpus-derived, frequency-based wordlists, it gives frequency statistics of word forms but no explanations of word meanings. Also, although there has been a tendency to evaluate frequency wordlists with regard to lexical ambiguity (mostly homographs), studies evaluating semi-technical medical wordlists, particularly the MWL, are almost non-existent.

1.2 Statement of the problem

Challenges involving lexical ambiguity in lexical semantics, lexicography and corpus linguistics are rooted in (a) the fuzziness of word meanings and (b) the unclear distinction between homographs and polysemes. These challenges have been observed in semi-technical medical vocabulary, making this type of vocabulary problematic and thus under-researched. Two main resources for semi-technical medical vocabulary—dictionaries and wordlists—each present their own problems that limit their effectiveness in fully addressing homography and polysemy. While conventional dictionaries show word meanings in a format that removes

explicit indication of homography and polysemy, corpus-derived wordlists lack explanations of homography and polysemy and have not been comprehensively evaluated.

This study is thus driven to address these multi-meaning challenges from an interdisciplinary perspective, bringing lexical semantics, lexicography and corpus linguistics together to (a) adequately describe the fuzziness of word meanings and (b) reliably distinguish between homography and polysemy. Specifically, the study presents a corpus-based, semantic analysis of semi-technical medical vocabulary that addresses unresolved issues relating to homography and polysemy in wordlists and dictionaries. It revisits a wordlist of semi-technical medical vocabulary—Hsu’s (2013) MWL—to evaluate the list with the purpose of specifying issues resulting from the word form-based process through which it was developed. Then, a new resource that deals exclusively with semi-technical medical vocabulary is developed to resolve issues related to the MWL. The development of this resource also presents suggestions on how to overcome the limitations of distinguishing and presenting polysemes and homographs in conventional dictionaries.

1.3 Scope of the research

Within the scope of this study, semi-technical medical vocabulary is restricted to words with multiple (un)related meanings activated differently in general and medical contexts (e.g., *defect*). Although previous lexical studies have given various definitions and different names to this type of vocabulary, there is consensus on the salient characteristic of semi-technical vocabulary, i.e., having both general and technical meanings (Cowan, 1974; Flowerdew, 1993; Huizhong, 1986), that distinguishes it from other types of vocabulary and makes it problematic in learning and teaching.

Lexical ambiguity generally includes homonymy and polysemy (Cruse, 1986; Kempson, 1977; Lyons, 1977; Murphy, 2010). Homonymy refers to words with the same pronunciation (homophones: *see* and *sea*) and words with the same spelling (homographs: *bank*

as “financial institution” and “edge of river”). Polysemy refers to regular and irregular polysemy (Apresjan, 1974; Carston, 2021). Regular polysemy encompasses cases in which words like *bank* as “financial institution” can systematically refer to “building that houses the institution”. Irregular polysemy involves words whose meanings are metaphorically connected and not easily regularized, e.g., *bank* as “financial institution” and “bank of blood”. This study focuses on homography and irregular polysemy because (a) the study works with written corpora that essentially require disambiguation of words with identical spelling and (b) irregular polysemy is much more complex than regular polysemy and deserves further study.

Lexicography broadly covers theoretical and practical lexicography, which respectively refer to a body of theory that underpins dictionary structures and components, and the craft of compiling dictionaries (Atkins, 2008). Since this study aims to develop a lexicographic resource, practical lexicography is a central focus. Moreover, practical lexicography taps into not only general but also specialized knowledge of vocabulary, making the focus on practical lexicography, rather than its neighbouring discipline of terminology, relevant to developing a semi-technical medical resource that needs to consider both general and medical meanings. Terminology, by contrast, mainly focuses on delimited knowledge within a particular domain and is thus more suitable for studies on technical (e.g., medical terminology) rather than semi-technical (medical) vocabulary.

This study narrowly focuses on the use of corpus linguistics in vocabulary studies. More particularly, it reviews studies on how corpora can be used to investigate the frequency of English words and how to choose which words should be taught to learners (usually in the form of frequency-based wordlists). It also taps into the application of corpora in dictionary making and WSD. The task of WSD, which emerges from the field of natural language processing, is brought to the field of corpus linguistics because of its relevance to vocabulary-related research. This study does not delve into applications of WSD in natural language processing

such as machine translation, information retrieval, speech processing and text processing. Instead, it closely examines the application of WSD for disambiguating words in corpora used in creating wordlists and dictionaries.

1.4 Aims of the research

The aim of the research was to develop a lexicographic resource of semi-technical medical vocabulary (named SemiMed) that addresses issues of homography and polysemy in current dictionaries and wordlists. The specific aims are presented below.

- Semantically analyzing words in the MWL, i.e., identifying their polysemes and homographs, to see whether a substantial number of words in the list have polysemes and/or homographs. If they do, there is a need to create a new resource (SemiMed) that fully accounts for polysemes and homographs identified in the MWL.
- Specifying the MWL's words with polysemes and/or homographs and their relation to other types of vocabulary.
- Identifying problems that these words can bring about in learning and teaching as well as influences they have on the pedagogical effectiveness of the MWL.
- Creating SemiMed based on polysemes and/or homographs identified in the MWL which are potentially problematic for learning and teaching.
- Incorporating theories and practices from lexical semantics and corpus linguistics into the development of SemiMed, i.e., into computational and lexicographic tasks of word sense disambiguation, distinction and presentation.
- Examining the usefulness of SemiMed compared to corpus-derived wordlists and conventional dictionaries with the focus on features resulting from a combination of interdisciplinary theories and practices.

1.5 Research questions

Seven research questions were proposed as follows:

- Where does semi-technical medical vocabulary sit on the vocabulary continuum?
- What words in Hsu's (2013) Medical Word List can be identified as possessing multiple meanings?
- What are the main disadvantages of semi-technical medical wordlists based on word form frequency?
- How do lexical semantic and corpus-based word sense disambiguation principles inform the development of SemiMed?
- Does SemiMed have a pedagogical potential? If so, what features support or undermine SemiMed as a teaching and learning resource?
- How does SemiMed compare to conventional dictionaries in terms of facilitating the understanding of polysemy and homography in semi-technical medical vocabulary?
- How could SemiMed be improved for users?

1.6 Research significance

Since semi-technical vocabulary, particularly in the field of medicine, has remained an under-researched type of vocabulary, this study, which comprehensively examines semi-technical medical vocabulary, will significantly add to the current literature on semi-technical vocabulary. Unlike previous studies on the creation of wordlists, which only list written forms of frequently occurring semi-technical medical words, and studies on the evaluation of wordlists, which identify their homographs (and sometimes their polysemes), this study thoroughly investigates the root cause of problems in learning and teaching semi-technical

medical vocabulary through analyzing its characteristic of having polysemes and/or homographs, and proposes a solution to address the issue.

The solution (SemiMed) is significant as it is in the form of learning and teaching resource that presents an alternative to wordlists and dictionaries. SemiMed is considered an enhanced version of semi-technical medical wordlists because it aims to provide sufficient information about word meanings, i.e., polysemes and homographs, in addition to word forms. It also has additional features to present polysemes and homographs explicitly, which are not observed in general and medical dictionaries. It is therefore hoped that SemiMed will improve the learning and teaching of semi-technical medical vocabulary.

SemiMed, especially its development, contributes a new, replicable methodology that incorporates lexical semantics and WSD in corpus linguistics into current lexicographic practices. This methodology will pave the way for compiling lexicographic resources, particularly of multi-meaning words, that give due consideration to homography and polysemy. Equally importantly, it examines the issues of homography and polysemy in each discipline and initiates links between lexical semantics, lexicography and corpus linguistics that are significant in comprehensively addressing the multiplicity of meaning.

1.7 Thesis structure

Chapter 1 – Introduction briefly describes multi-meaning challenges in three disciplines (lexical semantics, lexicography and corpus linguistics) connecting to problems in semi-technical medical vocabulary learning and teaching which are the motivation behind this study. It also defines the scope of the research within the three disciplines and specifies the aims of the research, followed by research questions. It ends with the theoretical and practical significance of the research in the three disciplines and in semi-technical medical learning and teaching.

Chapter 2 – Literature review provides a lexical semantic, lexicographic and corpus linguistic perspective on multi-meaning issues.

Through the lens of lexical semantics, Part 1 discusses how multiple, context-derived interpretations of a word are distinguished from one another and stored in the minds of language users. It begins with an introduction to lexical ambiguity, emphasizing homography and polysemy, followed by a discussion on different sources of homography and polysemy, revealing the complexity of lexical ambiguity. Next, methods of homography and polysemy distinction are presented in detail, considering some of their practical limitations. Finally, approaches to mental representations of polysemous words are discussed from two contrasting perspectives, i.e., classical and cognitive, indicating the fuzziness of word meanings.

Part 2 takes a lexicographic perspective to extend Part 1's discussion by examining how multiple, context-derived interpretations of a word are represented via distinct word senses in dictionaries. This part shifts attention away from word and meaning in the mind to word and meaning in the dictionary. It discusses the lexicographic approach to distinguishing word senses from their instances in context and the default format of presenting word senses in conventional dictionaries. Challenges relating to homography and polysemy in dictionary word sense distinctions and presentation are highlighted in association with the homography and polysemy distinctions and fuzziness of word meanings discussed in lexical semantics.

Part 3 focuses on discussing the corpus-based task of WSD, i.e., how multiple, context-derived interpretations of a word are distinguished from one another in a corpus. It starts with the emergence of WSD in natural language processing and then reviews contemporary WSD approaches and methods. Advantages and disadvantages of individual methods are clarified, together with challenges related to homography and polysemy that limit the performance of these methods. The challenges are discussed in connection with lexical semantics and

lexicography. WSD is then reviewed within lexical studies in corpus linguistics relating to the creation and evaluation of corpus-derived wordlists.

Chapter 3 – Methodology presents a theoretical framework that addresses gaps identified in the literature review. The gaps are unresolved issues of lexical ambiguity in lexical semantics, lexicography and corpus linguistics, i.e., challenges in (a) distinguishing homography and polysemy and describing mental representations of polysemous words, (b) identifying distinct dictionary word senses and presenting them in a way that fully considers homography and polysemy, and (c) undertaking a corpus-based WSD task that satisfactorily performs the disambiguation of both homographic and polysemous senses. Overall descriptions of methods used in each paper are presented, detailing which theories and practices in the three disciplines are adopted.

Chapter 4 – Paper 1 answers the first three research questions by examining Hsu's MWL. The paper first reviews semi-technical medical vocabulary in relation to other types of vocabulary, its elusive nature, and the non-transparent characteristic relating to polysemy and homography that makes semi-technical medical vocabulary problematic to learn and teach. It then proposes a method that reconciles contradictory theories in lexical semantics to distinguish polysemes and homographs in the MWL, specifying the location of semi-technical medical vocabulary and revealing problems in learning and teaching this type of vocabulary.

Chapter 5 – Paper 2 goes on to answer the following two research questions through the development of SemiMed. It focuses on describing and demonstrating the methodology underpinning the development of SemiMed that incorporates theories and practices emerging from the review of lexical semantic and corpus-based WSD literature. The methodology involves semantic and corpus-based analyses of problematic words identified in Paper 1. Results of the analyses, i.e., SemiMed's components and their functions, are discussed. The

pedagogical potential of SemiMed is explored with an emphasis on features absent in wordlists and dictionaries.

Chapter 6 – Paper 3 presents a pilot study of SemiMed that aims to answer the last two research questions. The pilot study was designed to allow student participants to use SemiMed alongside general and medical dictionaries and then provide feedback on the usefulness of SemiMed compared to current conventional dictionaries. This paper discusses features of SemiMed that are considered an improvement on the conventional structure of current dictionaries and their beneficial impacts on polysemy and homography. It also explores some elements of SemiMed that need enhancement and offers suggestions for its improvement.

Chapter 7 – Conclusion opens with a summary of key findings from the three papers. Then, it discusses implications of these findings in (a) resolving issues of homography and polysemy in three disciplines, highlighting the importance of an interdisciplinary approach that connects lexical semantics with lexicography and WSD in corpus linguistics to the multiplicity of meaning and (b) suggesting a new direction to study, learn and teach semi-technical medical vocabulary. It concludes with recommendations for future research regarding how to address methodological limitations of this study to achieve a full, well-rounded version of SemiMed.

CHAPTER 2: LITERATURE REVIEW

PART 1 – LEXICAL SEMANTICS

2.1 Key concepts

Semantics is a branch of linguistics concerned with meaning in language. The *lexical* in *lexical semantics*, according to Murphy (2010), involves the lexicon, which indicates “the vocabulary of a language (also known as lexis)” and/or “a particular language user’s knowledge of her/his own vocabulary” (p. 4). This definition, with an emphasis on the second half, which is vocabulary in the mind of a language user, may more precisely refer to the mental lexicon. The mental lexicon is arranged into lexical entries, each of which “collects the appropriate information about a particular linguistic expression, called a lexeme” (Murphy, 2010, p. 5).

A lexeme is considered an abstract representation of a linguistic form, or, to put it another way, “a linguistic form (i.e., a bit of speech and/or writing) represents a lexeme if that form is conventionally associated with a non-compositional meaning” (Murphy, 2010, p. 6). In this view, a lexeme is characterized by non-compositionality and conventionality. Take *cat* as an example. The lexeme *cat* is *non-compositional* because constituent parts of its linguistic form, either phonological (e.g., the sounds /k/, /æ/, and /t/) or orthographical (e.g., the letters *c*, *a*, and *t*), do not make up (or cannot be used to predict) its meaning. Saussure (2011) uses the term *arbitrary* to describe a linguistic form and meaning relation such as *cat*, which is *conventional* in the sense that “form-meaning pairings are common knowledge among the speakers of the language, and we have had to learn these particular associations of form and meaning from other members of the language community” (Murphy, 2010, p. 6).

So, strictly speaking, from the analysis of the two components (*lexical* and *semantics*), lexical semantics is the study of lexeme meaning, although it is sometimes “loosely defined as the study of word meaning” (Murphy, 2010, p. 6). This loose definition of lexical semantics is acceptable, but it should be borne in mind that wherever in Part 1 the term *word* is mentioned,

it refers to a *lexeme* and distinguishes itself from a *word form*. Word forms can be described as “individuated by their form, whether phonological or graphic” (Cruise, 2000, p. 88). For example, the lexeme *run* is represented in multiple forms such as *run*, *runs*, *running* and *ran*. It can be observed from this example that “lexemes can be regarded as groupings of one or more word forms” (Cruise, 2000, p. 88).

2.2 Lexical ambiguity

A linguistic phenomenon in which a word form may have more than one interpretation is often known as lexical ambiguity (Cruse, 1986; Kempson, 1977; Lyons, 1977). Two types of lexical ambiguity—homography and polysemy—will be presented in detail below.

2.2.1 Homography and polysemy

Homonymy is defined as “a relation between different lexemes that are coincidentally similar in form” (Murphy, 2010, p. 90). More specifically, a pair of homonyms consists of two different lexemes that just happen to have the same spoken and/or written word form. If two lexemes have the same pronunciation (e.g., *sea* and *see*), they are homophones. If they have the same spelling (e.g., *bear* “an animal” and *bear* “to carry”), they are homographs. Since the main focus of this study is homography, from this point onward, only homography is discussed in relation to polysemy.

Polysemy indicates “a relation between [meanings] associated with a single lexeme” (Murphy, 2010, p. 90). Looking again at *bear* (*v*), this single lexeme has two distinguishable meanings: “to move while holding up and support” and “to hold in the mind” (Garner, 2007, p. 251). These meanings are related to each other and not different enough to split into two lexemes. Thus, “to move while holding up and support” and “to hold in the mind” are two meanings associated with only one lexeme. The lexeme *to bear* is called a polyseme; in other words, it is polysemous.

2.2.2 Sources of homography and polysemy

A review of current literature has highlighted several sources of homography and polysemy, particularly including language change, lexical borrowing and semantic change (Carston, 2021; Cowie, 1988; Béjoint, 1990; Bréal, 1900; Murphy, 2010; Ullmann, 1962; Vicente & Falkum, 2017).

The fact that the English language changes over time may accidentally cause unrelated words to come closer together in form. Homography is thus believed to mostly emerge through coincidence. By way of illustration, Murphy (2010, p. 87-88) shows how two lexemes, *sole* (n. “the bottom surface of a shoe”) and *sole* (adj. “only”), have evolved to become identical in spelling. Her etymological traces, which are aligned with etymological information about the two lexemes in the *Oxford English Dictionary*, reveal that *sole* as a noun has a Latin root *solea* which means “sandal” and *sole* as an adjective derives from a different Latin root *solus* which means “alone”. She reasons that despite being derived from different origins, these two gradually appeared to be form-related due to the language change over the centuries through which their final syllables were omitted and that possibly later led to the similarity in their spellings.

Another possible reason for homography is the fact that English vocabulary contains many loanwords from other languages (Jackson, 2013). This is called lexical borrowing. Murphy (2010, p. 94) takes *yen*, which originates from Japanese indicating “the currency of Japan” and is used in the English language as a loanword, as an example. She states that this word has no relation with *yen*, the existing English word meaning “a strong feeling of wanting or wishing for something”, in terms of their origins and usages. Each of them has evolved in its own way and only happens to share the same form. It is thus only a pure coincidence that the Japanese-originated *yen* has become a homograph with the English word *yen* since it was borrowed into English.

Unlike homography, polysemy originates neither from language change nor lexical borrowing. It is a diachronic phenomenon stemming from a mechanism that is termed semantic change (Bréal, 1900), semantic shift (Cowie, 1988) or shift of application (Ullmann, 1962), where old words are used in new ways. Instead of learning new words all the time, language users tend to extend existing meanings of a word in “predictable” (Murphy, 2010) or “conventionalized” (Vicente & Falkum, 2017) ways so that they can effortlessly understand newly created meaning(s) of an old word. Under this view, polysemy is an outcome of the semantic change in which old and new meanings coexist (Bréal, 1900; Vicente & Falkum, 2017). This is illustrated through Murphy’s (2010, p. 88) example of *coat*. *Coat* has the first historically recorded meaning of “an outer garment with sleeves for wearing outdoors”. This meaning later branches out into two meanings: “an animal’s covering of fur” and “a covering of paint or similar material”. The original and new extended meanings are still somewhat related to one another and exist in parallel, making *coat* itself a polyseme.

Nonetheless, semantic change only sometimes leads to polysemy. Carston (2021), Béjoint (1990) and Murphy (2010) state that through processes of change of meaning, a word may become polysemous if its original meaning is retained and its connection with other extended meanings (semantic link) is maintained (as in the case of *coat*). Otherwise, they expect that the word may either be no longer a polyseme or have homograph(s).

The first possibility occurs when the original meaning(s), for some reason, may be overshadowed by the extended ones and then die out so that they are no longer polysemous. This has been seen in the case of *undertaker* (Murphy, 2010, p. 95), which “originally meant anyone who undertakes some business for someone else”. This broad meaning was extended to particularly indicate “someone who undertakes funeral preparations for others”. The original meaning gradually fell out of use and the word is no longer considered polysemous.

The second possibility is that through a process of semantic change, meanings of a polyseme may drift so far apart from one another over time that their relatedness is no longer identifiable. Returning to *sole* as a noun, it has another meaning in addition to “the bottom surface of a shoe”, which is “a type of flatfish”. The “fish” meaning comes from the same root (*solea*) as the “sandal” meaning because of its resemblance to a flat shoe (Murphy, 2010; *Online Etymology Dictionary*). The meanings are thus etymologically related and should be deemed polysemous. However, language users today seem unable to see the link between “fish” and “sandal” and perceive them as homographs rather than polysemes.

2.2.3 Distinctions between homography and polysemy

It can be noted that homography and polysemy emerge from various sources, i.e., through accident and/or processes of semantic change, which makes the distinction between the two phenomena difficult to draw. Several attempts have been made to determine whether interpretations of a word form constitute a case of homography or polysemy. From reviewing the existing body of literature, the etymology of the word and the (un)relatedness of the word meanings stand out as two predominant approaches to homography and polysemy distinctions (Carston, 2021; Klepousniotou, 2002; Leech, 1974; Lehrer, 1974; Lyons, 1968, 1977; Panman, 1982).

The former tends to look back to the history of a word to identify its origin. If words are from the same lexical source, they are seen as polysemes. If they are from distinct lexical sources, they are regarded as homographs. This kind of distinction is observable in virtually all historical dictionaries, where polysemous meanings are listed (and usually numbered) under a single entry as different meanings of a single word. In contrast, homographs are treated as separate words and thus given separate entries. Relying on the etymological derivation of words may be helpful, because etymologically related rules, in theory, would be applicable to recognize the homography and/or polysemy derived from coincidence and semantic change.

However, in reality, decisions are not always straightforward (Klepousniotou, 2002; Lyons, 1968, 1977; Panman, 1982) because “there are many words about whose historical derivation we are uncertain” (Panman, 1982, p. 118) and “it is not always very clear how far back we should go in tracing the history of words” (Klepousniotou, 2002, p. 206). This can be seen again in the case of *sole* (fish and sandal). Even though the meanings are derived from the same lexical source, the source per se is still hard to trace back and not apparent to every present-day English L1 user.

The latter focuses on a native speaker’s judgement on the (un)relatedness of word meanings. According to Lyons (1977), relatedness in meaning indicates polysemy while unrelatedness in meaning indicates homography. This approach might be an alternative, especially, as mentioned earlier, when the word’s etymology is hardly traceable. Looking again at the example of *sole* (fish and sandal), present-day English L1 users unaware of the etymological connection between the “fish” and “sandal” meanings could possibly feel that “fish” and “sandal” are not related to each other and decide that the two meanings are in a relation of homography (Murphy, 2010). The possibility that some people may perceive “fish” and “sandal” as homographs, however, does not necessarily exclude the possibility that others may see a similarity in the shape of “fish” and “sandal” from which they conclude that these two meanings are related to each other and thus should be deemed polysemous. The native speaker-based distinction between relatedness and unrelatedness is therefore not as apparent as it may seem.

Hence, there is an issue for those who strive to establish a firm distinction between polysemy and homography relying on their subjective judgement of meaning (un)relatedness. It is relatively challenging for different native speakers to reach an agreed decision on whether a multi-meaning word should be assigned to either the polysemy or the homography category, as different people hold different views on how (un)related meanings of a word are enough for

polysemy and homography. In other words, they may face a dichotomous challenge which leads to a tendency to treat polysemy and homography as two opposite ends of a continuum, allowing native speakers to evaluate meaning(s) of a word against the continuum, i.e., closer to the polysemy or homography end, depending on the degree of (un)relatedness they come up with.

Possibilities that native speakers may diverge in their opinions about the (un)relatedness of meanings of a word could be attributed to the subjectivity that this approach relies on. Lehrer (1974, p. 10) uses the term “behaviorally valid” to describe the distinction drawn between polysemy and homography retrieved from the subjective judgement of a native speaker. She reports that disagreements escalate among different native speakers (and different responses are retrieved from the same native speakers at different times) when they get involved in distinguishing polysemy from homography based on their intuitive judgement, especially when meanings have some semantic similarity (e.g., “fish” and “sandal”). Lehrer adds that agreements, however, might be achievable for words whose meanings do not have semantic similarity like *sole* as a noun (fish or sandal) and *sole* as an adjective (alone).

2.3 Approaches to polysemy

There have been several approaches to explaining polysemy, among which *monosemy* and *polysemy* are two fundamental positions that diverge theories on polysemy into separate directions (Cruse, 1992; Janssen, 2003; Murphy, 2010).

2.3.1 The monosemy approach

The monosemy position considers polysemy a surface phenomenon, in which lexical entries are underspecified (Carston, 2021; Evans & Green, 2006; Frisson & Pickering, 2001). From this premise, lexical entries (which are described in 2.1 as containing information about particular lexemes) are generally abstract with minimum detail and then filled in by contexts (Ruhl, 1989) or generative rules (Pustejovsky, 1995). The monosemy approach emphasizes

that rather than representing multiple meanings of a word in the mind, only its general semantic representation is stored.

The emergence of monosemy has laid a foundation from which theories (notably, Pustejovsky's (1995) Generative Lexicon Theory and Ruhl's (1989) Monosemic Bias) on relations between meanings of a polysemous word have been developed. This approach sounds appealing in the sense that individual polysemous words are represented by a single abstract meaning, which advocates a simpler lexicon. Instead of having many meanings of polysemous words stored in the mental lexicon, different interpretations of polysemous words could be gained through contextual clues or applications of lexical generative devices.

Although monosemy-based explanations that represent a polysemous word via a single abstract meaning in the mental lexicon are viewed as elegant, such explanations tend to overlook the complex nature of polysemy. Theories derived from the monosemy approach may be subject to criticism due to "downplaying the amount and range of polysemy found in natural language" (Murphy, 2010, p. 101). The monosemy approach seems to pose a risk of oversimplification. It can be inferred from the monosemy explanations that specific interpretations of a mentally stored simplified representation of a word derived from either contexts or generative rules may follow regular patterns. Yet in reality, polysemous word meaning variations are part of irregularity, from which emerges an opposite approach to monosemy, known as the polysemy approach. This approach suggests representing each polysemous meaning of a word in the mind and treating a polysemous word as a complex network of mental representations.

2.3.2 The polysemy approach

2.3.2.1 Words as radial categories

The polysemy position, pioneered by Claudia Brugman whose work later inspired George Lakoff (Brugman & Lakoff, 1988; Lakoff, 1987), is dissimilar to the monosemy

position in terms of not viewing polysemy as a surface phenomenon but rather as a conceptual phenomenon. Their work brings a new perspective to word meanings, that is, a cognitive perspective, which departs from the monosemy approach. They claim that polysemy in language use is reserved in the mind via lexical organization where each (polysemous) word is stored as a category of distinct polysemous meanings rather than as a single abstract monosemous meaning. Their work has laid the foundation for the development of cognitive lexical semantics, where words are generally viewed as conceptual categories.

The term *conceptual category* was originally proposed by a cognitive psychologist, Eleanor Rosch (Rosch, 1973, 1975, 1977). In her study, she found that categorization is central to the human conceptual system. Her rationale behind this finding is that human beings tend to maximize their capacity to store as much information about the world as they can by grouping similar bits of information into categories rather than separating them. By this means, categorization is expected to give rise to concepts and account for the organization of concepts in the human mind. Rosch claims that a mechanism of forming and organizing categories is underpinned by the correlational structure existing in our world, for example, “wings co-occur with feathers more than with fur” (Rosch, 2004, p. 92). The existence of correlational structure sheds light on her *prototype theory*, which posits that a human being is in the habit of categorizing objects based on how closely they resemble the prototype of a category.

Following Rosch’s approach of categorization by family resemblance, a category is demonstrated through a prototype, the best example that exhibits the most representative features of the category (Rosch & Mervis, 1975). More prototypical members often exhibit a large number of features typical to the category, some of which are found in other less prototypical members grouped under the category. A classic example of typicality effects on categorization from Rosch’s study is to list members that constitute the category of *bird*. Rosch’s participants tended to more quickly and certainly decide that a *robin* was a bird

because it is more typically birdlike than other members, say, for example, *ostrich*. Participants made the decision based on their judgement that a *robin* has almost all the core features of a bird, including, but not limited to, having two legs and feathers and being able to fly and sing, while *ostrich* does not have the last two features. The example consolidates the prototype-based conception of categorization that categories such as *bird* are structured to include members (*robin* and *ostrich*) which resemble the prototype (of bird) to some extent.

Rosch's study provides insights into human categorization that is central to both cognitive psychology and cognitive lexical semantics because there has been a demand within the two disciplines for formulating theories that account for knowledge and linguistic representation in the human mind. Her findings were of fundamental importance for the work of Lakoff (1987), since he adopted the prototype theory to explain linguistic conceptual categories. At the heart of his work, words are equated with conceptual categories akin to Rosch's non-linguistic conceptual categories in terms of having a prototype structure. Lakoff (1987) states that different meanings of a word stored in a conceptual category exhibit the effects of prototypicality—these meanings are judged as more prototypical or less prototypical vis-à-vis prototype(s) of the category.

In support of this view, Lakoff (1987) uses *mother* as an example. *Mother* is a complex concept in which different aspects converge, i.e., a female who (a) gives birth to a child, (b) contributes to his/her genetic material, (c) nurtures and raises the child, (d) gets married to his/her father and (e) is the child's closest ancestor. These five aspects constituted Lakoff's *mother* prototype and some of them are taken as the basis for the extension of other (less prototypical) meanings. For instance, in mining, *mother*, which indicates the source of an ore (*mother lode*), is a less prototypical use of the word because it only focuses on (a) "giving birth". In syntax, *mother* is used to name one type of node in a syntactic tree diagram (*mother node*: the one under which some other node falls). *Mother* in this sense is less prototypical

because it only reflects one part of (e) “being the closest ancestor”. Another less prototypical use of *mother* is *to mother*. Anyone who commits an act of mothering is expected to take a nurturing role that only emphasizes (c).

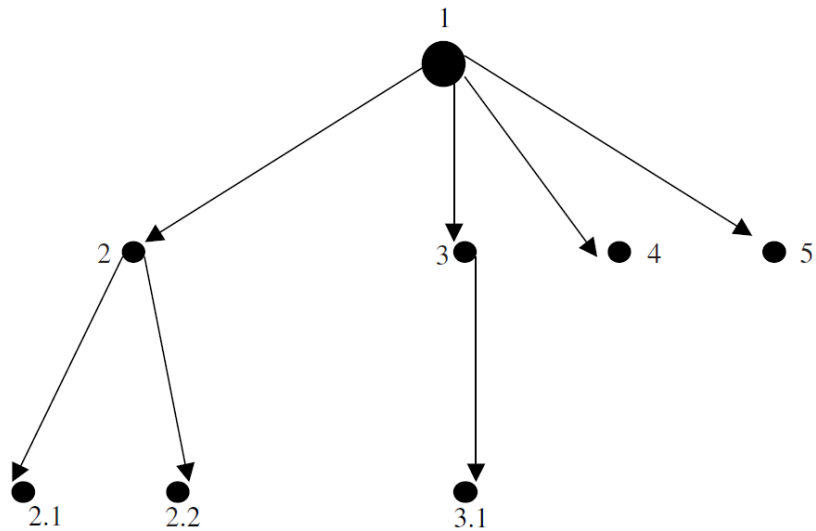
Lakoff suggests the term *radial* to describe the category of *mother*, as he reasons that different meanings of *mother* radiate from the key aspects of the concept. As such, a lexical conceptual category is a *radial category* where meanings are organized with respect to the prototype, i.e., closer to the prototype (more prototypical meanings) or further from the prototype (less prototypical meanings). He also asserts that meanings in a radial category are not generated from the prototype by predictable rules. Rather, they are related to the prototype by convention, i.e., most native speakers simply know the range of meanings associated with the prototype. Thus, radial categories are not meaning-generating devices. They instead model how word meanings are organized in the mental lexicon. In this regard, (polysemous) words are stored as “highly complex structured categories” (Evans & Green, 2006, p. 328). This view is opposed to monosemy as it rejects storing a single abstract meaning in the mind and applying predictable rules from contexts or generative devices to specify the single abstract meaning.

2.3.2.2 The Principled Polysemy approach

While Lakoff’s radial categories are influential in the field of cognitive lexical semantics, his approach to word meanings has attracted considerable criticism. As he emphasizes that word meanings in each conceptual category are conventionalized, critics (Tyler & Evans, 2003a; Dominiek, 1998) may question the ability of his theory to provide objective results. Evans and Green (2006, p. 342) point out that Lakoff’s model of radial categories is a result of “intuitions (and perhaps also the imagination)” of involved analysts rather than of lexical representations that language users actually store in their mind. Moreover, they argue that semantic analysts do not always agree about the central meaning (the prototype) from which other meanings in a category are derived.

Therefore, Vyvyan Evans and Andrea Tyler (Evans, 2004, 2005; Tyler & Evans, 2001, 2003a, 2003b, 2004) offer an approach named *Principled Polysemy* in response to the subjectivity existing in Lakoff's radial categories, which was initially raised by Dominiek (1998, p. 371) as a consequence of lacking "a set of scientifically valid decision principles". The Principled Polysemy approach comprises "decision principles" that ensure the analysis of polysemy is objective and testable. The aim of decision principles is twofold: (1) establish the central meaning and (2) identify distinct (more or less prototypical) meanings surrounding the central meaning stored in the mental lexicon.

Principled Polysemy was originally developed to model semantic networks of English prepositions and then extended to go beyond prepositions to account for an abstract noun, i.e., *time*. Evans (2005) takes a closer look at the noun *time* to explicate core tenets of the Principled Polysemy approach. A written form of *time*, according to Evans (2005, p. 38), embraces distinct meanings that are derived from and organized vis-à-vis a "historically earlier" meaning in a principled way. The historically earlier meaning, which is termed a "sanctioning sense", is believed to "typically (although not inevitably) [have] parallels with the diachronically earliest sense" and taken as central "prototypical" (Evans, 2005, p. 38-39). Meanings of *time* are separately stored in the mental lexicon via a *semantic network*, which is demonstrated in a "radiating-lattice structure".



- | | |
|-------------------------|---------------------------------|
| 1: The Duration Sense | 3: The Matrix Sense |
| 2: The Moment Sense | 3.1: The Agentive Sense |
| 2.1: The Instance Sense | 4: The Measurement-system Sense |
| 2.2: The Event Sense | 5: The Commodity Sense |

Figure 2.1 A semantic network of *Time* (Adapted from Evans, 2005, p. 52)

Evans (2005) formulates two sets of criteria to determine the central (prototypical) and other meanings of *time*. To establish the appropriate central prototypical meaning, or alternatively named, the sanctioning sense, for *time*, Evans (2005) proposes four criteria:

- (1) criterion of earliest attested meaning,
- (2) criterion of predominance,
- (3) criterion of predictability,
- (4) criterion of lived temporal experience.

To identify distinct meanings, he proposes three criteria:

- (1) meaning criterion,
- (2) concept elaboration criterion,
- (3) grammatical criterion.

Evans's (2005) criteria-based analysis results in eight meanings of *time*, three of which (1: The Duration Sense, 2: The Moment Sense and 2.1: The Instance Sense in Figure 2.1) are selected to illustrate the entire analytic procedure.

First, the sanctioning sense for *time* is determined by ensuring that the four criteria are met. The first criterion requires that the sanctioning sense should be most closely related to the earliest attested meaning of *time*. Evans (2005) refers to the *Oxford English Dictionary* to trace back the earliest attested meaning associated with *time*, which is “duration”. He asserts that even though the sanctioning sense is not necessarily the earliest attested meaning (origination sense), it may overlap with the origination sense because “duration” may “still play an active part in the synchronic network” (p. 40) of *time*. He thus nominates “duration” to become a candidate for the sanctioning sense. The sanctioning sense of *time* is specified as a bounded duration, more specifically, “an interval which is co-extensive with a particular state or process” (Evans, 2005, p. 48). The second and third criteria, according to Evans (2005), ensure that the sanctioning sense constitutes a meaning component which is “most predominant (frequent) in the semantic network” (p. 44) and from which “other distinct senses can be most plausibly predicted” (p. 50). Evans’s (2005) analysis of *time* reveals that “duration” is present in over half of the distinct senses and best meets the third criterion. These two criteria will be reiterated in greater detail in the discussion about distinguishing distinct senses of *time*. The fourth criterion links to human experience of *time* at the phenomenological level. Evans (2005) argues that our experience of time is related to “an awareness of temporal magnitude” (p. 45) which allows us to “distinguish past from present and . . . experience events as successive” (p. 50). According to him, “duration” most closely approximates this lived experience of *time* and satisfies this final criterion. Summing up, “duration” meets the four criteria and is thus acknowledged as the sanctioning sense of *time*.

Second, to determine whether a meaning is considered distinct, the meaning criterion and at least one other criterion need to be satisfied. The meaning criterion warrants that a distinct meaning “must contain additional meaning not apparent in any other [meanings] associated with *time*” (Evans, 2005, p. 41). Although the meaning criterion per se is sufficient

to justify a meaning as distinct, either concept elaboration or the grammatical criterion is, according to Evans (2005), still required to “safeguard judgements of meaning distinctiveness (on the part of the analyst) from the undue influence of context in identifying a particular usage as a particular [meaning]” (p. 42). The concept elaboration and grammatical criteria concern collocational dependences (Croft, 2001) and structural dependences (Evans, 2005) of a distinct meaning. They respectively suggest that a distinct meaning may be manifested through unique sets of lexical items co-occurring with that meaning and may appear in unique grammatical constructions. Put another way, the last two criteria provide syntagmatic and grammatical evidence in addition to the meaning criterion.

The examples of *time* in the following sentences exemplify how the three criteria are applied to identify distinct meanings:

- (1) a) The relationship lasted a long/short time. (Evans, 2005, p. 48)
- b) Looking back on the evening of their first date, it seemed to the couple that the time had flown by. (Evans, 2005, p. 42)
- c) Time seemed to stand still. (Evans, 2005, p. 39)
- d) Time seemed to have flown by. (Evans, 2005, p. 39)
- (2) Due to the volatile nature of the market, we left instructions to sell at an appropriate time. (Evans, 2005, p. 54)
- (2.1) a) Devine improved for the fourth time this winter when he reached 64.40 metres at a meeting in Melbourne. (Evans, 2005, p. 55)
- b) The horse managed to clear the jump 5 times in a row. (Evans, 2005, p. 56)

Time in (1a) expresses the sanctioning sense of “a bounded interval of duration” (Evans, 2005, p. 42). *Time* in (1b) is also linked to “a bounded interval of duration” but elaborated in a different way. While “duration” in (1a) is interpreted in terms of “physical length”, it is interpreted in terms of “motion” in (1b). The “physical length” and “motion” are

conceptualized by the use of modifiers (a long/short time) and verb phrases (have flown by), or in other words, have unique syntagmatic patterns and thus satisfy the concept elaboration criterion to become distinct meanings. However, both manifest “a bounded interval of duration” which means the meaning criterion is not satisfied in these cases. Therefore, (1b) is only seen as an elaboration (or a particular usage) of the sanctioning sense, not a distinct meaning. Likewise, (1c) and (1d) are two other elaborations of the sanctioning sense.

Unlike (1c) and (1d), which are elaborations of “duration”, (2) conveys “a discrete point” which according to Evans (2005, p. 53) brings additional meaning to “duration”. *Time* in (2) may appear to be a distinct meaning (termed “moment” sense) as it meets the meaning criterion. To confirm that “moment” sense is distinct from “duration” sense, the second and third criteria are considered. Evans (2005, p. 53-54) points out that (2) is elaborated in terms of “deictic motion”, which is a different kind of motion from the one in (1c) “protracted duration” and (1d) “temporal compression”. Moreover, *time* in (2) is a count noun while in (1c) and (1d) *time* is a mass noun. This means that “moment” sense meets not only the meaning criterion but also concept elaboration and grammatical criteria, making it a distinct meaning.

(2.1) features “a particular instance (i.e., occurrence) of an event or activity” (Evans, 2005, p. 56) rather than “a bounded interval of duration” and “a discrete point” in (1) and (2). Regarding the meaning criterion, *time* in (2.1) adds meaning not apparent in (1) and (2) and thus satisfies the first criterion. Regarding the concept elaboration criterion, the “instance” sense, according to Evans (2005), does not have unique patterns that can distinguish it from elaborations of “duration” and “moment” senses. So, “instance” sense fails to meet the second criterion. Regarding the grammatical criterion, even though *time* in (2.1) is a count noun, like in (2), it is modified by ordinal numbers (2.1a) and cardinal numbers (2.1b). This is a salient grammatical feature that is not observed in (2) and results in treating “instance” as a distinct sense (as it meets the meaning and grammatical criteria).

The application of the three criteria identifies two distinct meanings (moment and instance) apart from the sanctioning meaning (duration). For Evans (2005), “moment” and “instance” are two distinct meanings derived from the sanctioning sense “duration” (or more precisely, “moment” derives from “duration” and “instance” derives from “moment”) (Figure 2.1). He believes that a motivation for the derivation of “moment” and “instance” is highly plausible and mentions Flaherty’s (1999) phenomenon of time embeddedness as a driving force behind the derivation of “moment” (a discrete point) from “duration” (an interval). Time embeddedness is generally understood to mean that “events are embedded within other events” (Evans, 2005, p. 54), in other words, intervals are subsumed by greater intervals. This phenomenon underpins what Evans relies on to claim that the “moment” sense is plausibly predictable from the “duration” sense, that is, the embeddedness of “a discrete point” within “an interval”. He further clarifies the derivation of “instance” from “moment” by saying that “various intervals within larger intervals ... [are] enumerable” (p. 57). This means discrete points, which are embedded within intervals, “constitute particular instances which can be enumerated” (p. 57). This explains why the meaning extension of *time* from its sanctioning sense, i.e., the “instance” sense derives from the “moment” sense which derives from the “duration” sense, is principled, not conventionalized. “Duration” is thus a meaning component which is present in “moment” and “instance” and from which “moment” and “instance” are plausibly predictable. In other words, it meets the predominance and predictability criteria to be considered the sanctioning meaning.

By and large, Principled Polysemy offers a carefully articulated set of linguistic tools to analyze polysemy, i.e., justify (metaphoric) semantic extension. This is a “promising” model because it is among the first to propose “rigorous decision principles” (Gries, 2015, p. 29-30) to (a) identify the prototype of a polysemous category and (b) determine whether the usage of a polysemous word means that the user counts it as a distinct meaning stored in the mind. These

principles are considered important because they target the methodological problems in Brugman and Lakoff's (1988) and Lakoff's (1987) previous work. A primary advantage of Principled Polysemy is its replicable methodology, i.e., its criteria "help make decisions more replicable" (Gries, 2015, p. 30; Mahpeykar & Tyler, 2015). The application of Principled Polysemy has been widely observed in studies mainly on prepositions (Tyler & Evans, 2001; Van der Gucht et al., 2007), (abstract) nouns (Evans, 2005) and verbs (Dalpanagioti, 2018; Mahpeykar & Tyler, 2015). However, Evans (2005) suggests that the applications in these studies may be transferrable to other lexical classes, though there has been very little further research carried out so far.

PART 2 – LEXICOGRAPHY

2.4 Key concepts

The term *lexicography* has two interpretations—*theoretical lexicography* and *practical lexicography*—which, according to Atkins (2008, p. 31), can broadly be defined respectively as “a body of theory related to lexicography” and “the art and craft of dictionary-making”. In this section, lexicography is understood in terms of practical lexicography and is discussed with regards to the compiling of dictionaries.

The process of compiling a (monolingual) dictionary has two phases (Figure 2.2). An *analysis* is conducted first, during which lexicographers “[analyze] the word, trying to discover as many relevant linguistic facts as possible, record them, understand them, and order them” (Atkins, 2008, p. 33). The input to the first phase can be varied, but usually involves a corpus that can be manipulated by software tools. The corpus-based analysis involves a wide range of tasks, of which the most important is sense finding. The analysis output is usually recorded in the form of a database.

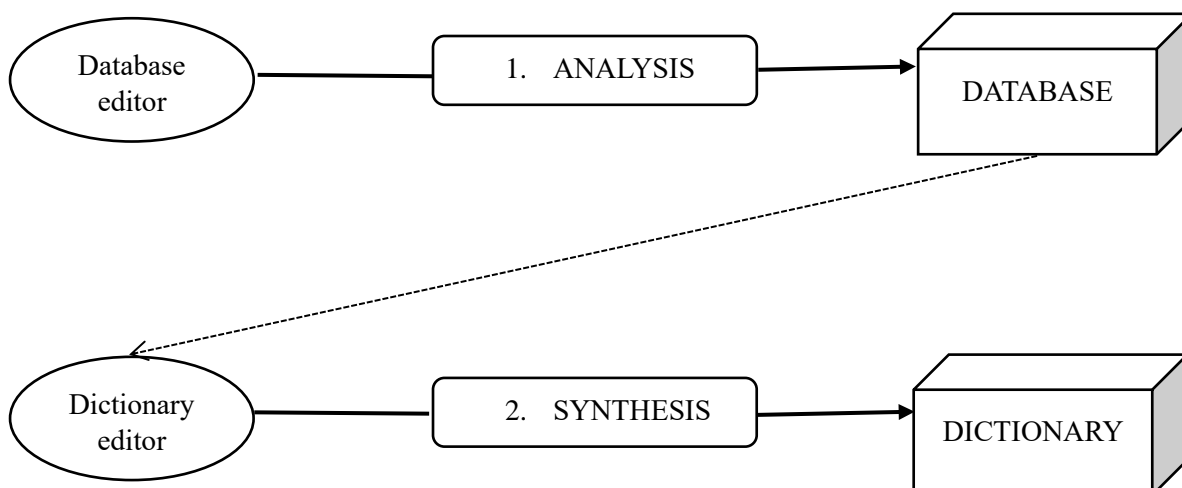


Figure 2.2 The two-phase process of compiling a corpus-based monolingual dictionary (Adapted from Atkins & Rundell, 2008, p. 98)

The database contains structured, computer accessible information that lexicographers rely on to execute the second phase, the *synthesis*. In this phase, lexicographers extract from the database relevant information and synthesize it to create a dictionary entry. The synthesis phase results in detailed guidelines that inform lexicographers when constructing and presenting dictionary entries. Among synthesis-involved tasks, determining *dictionary senses* is one of the “most problematic” as it requires skill and experience on the part of the lexicographer (Atkins & Rundell, 2008, p. 102).

Generally speaking, the first phase prepares resources on which the creation of a dictionary is based. The second phase involves decisions that determine the content of a dictionary, which is, more specifically, the *macrostructure* and *microstructure* of a dictionary. The former concerns which types of entry are presented and how headwords are organized, while the latter deals with which components are included and how these are structured in an entry (Atkins & Rundell, 2008; Hausmann & Wiegand, 1989).

A dictionary may contain different types of entry depending on the types of lexical item it features. There are two types of lexical item: single-word and multiword items (Atkins & Rundell, 2008; Pellicer-Sánchez, 2019). In single-word items, simple words create a category encompassing “the common words of the language”, subdivided into lexical and grammatical words (Atkins & Rundell, 2008, p. 164). Lexical words are open-class items, including nouns, verbs, adjectives, adverbs and interjections, while grammatical words are close-class items, including prepositions, conjunctions, pronouns, auxiliary verbs and determiners. Entries that feature lexical words are named *standard lexical entries* and this type of entry is the main focus of this study.

Within a standard entry, there are usually three components: a headword, lexical unit(s) and run-on (Figure 2.3). An example of headwords in a dictionary is the case of *play* (n, v) (Atkins & Rundell, 2008). *Play* as a noun has one inflected form (*plays*), while *play* as a verb

has three inflected forms (*plays*, *played* and *playing*). So, the two entries for *play* (n, v), representing all their inflected forms, are ideally treated as two headwords in a dictionary.

A *lexical unit* can be defined as “a headword in one of its senses” (Atkins & Rundell, 2008, p. 162) or simply a word sense (or a dictionary sense). Lexical units are listed and usually numbered (in bold in Figure 2.3, for example) under a headword. They are considered “core building blocks” (Atkins & Rundell, 2008, p. 163) of an entry because the ultimate purpose of writing dictionary entries is to provide word senses for which dictionary users look.

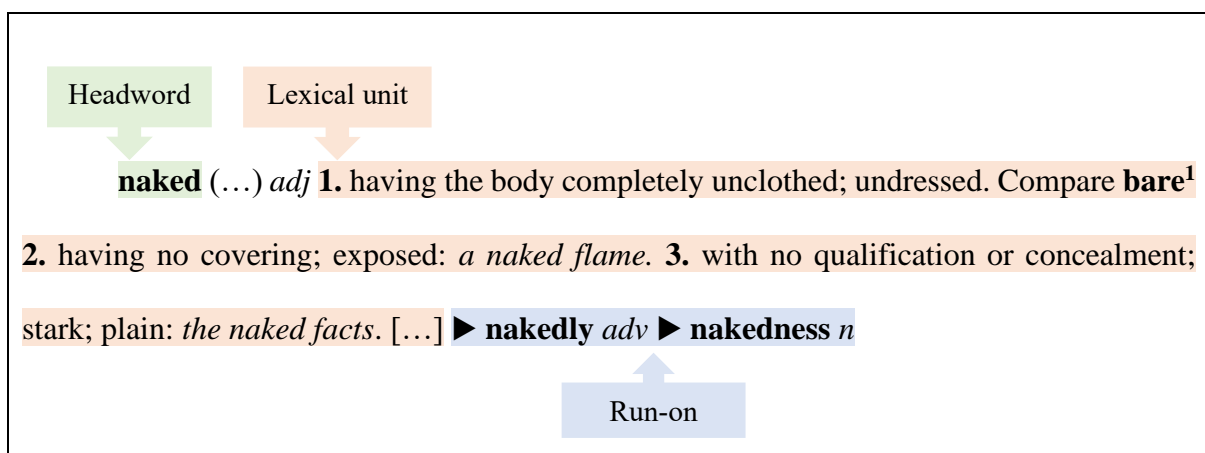


Figure 2.3 Three components in the entry for *Naked*. Definitions from *Collins English Dictionary* in the order in which they appear.

A *run-on* is part of an entry that is reserved for infrequent derived forms of a headword (e.g., *nakedly*, *nakedness*). Although this is an optional section in many dictionaries, the run-on is introduced here for the sake of later discussion. As can be seen in Figure 2.3, only the part of speech is shown in the run-on, so any derived form(s) appearing in the run-on should exhibit characteristics that do not confuse dictionary users: “its meaning is unambiguously deducible through the application of basic word-formation rules [,] its pronunciation can be predicted from the pronunciation of the headword it is attached to [,] its grammatical and collocational behaviour is simple and predictable” (Atkins & Rundell, 2008, p. 237).

2.5 Dictionary word senses

Since dictionaries are “designed for human users by humans” (Levin, 1991, p. 206) to provide “descriptions of our lexical knowledge” (Jorgensen, 1990, p. 168), lexicographers may relate their dictionary (entry) writing to the language user’s knowledge of vocabulary. There is thus an analogy between a lexical entry in lexical semantics (more precisely, cognitive lexical semantics) and a dictionary entry in lexicography. Both contain information about words, differently labelled within each discipline as lexemes and headwords. Although cognitive lexical semantics and lexicography appear to perceive and describe words from different perspectives, i.e., the former focuses on how meanings of a lexeme are stored in the mind and the latter focuses on how word senses under a headword are presented in the dictionary, they more or less tap into issues around distinct meaning/sense identification (Van der Eijk et al., 1995). Semantic linguists like Evans and Tyler, as mentioned in Part 1, have strived to establish parameters for differentiating distinct meanings from their usages in particular contexts to determine which meanings are stored in mental semantic networks. Lexicographers seem to work on the same ground, because they tend to generalize specific instances of a word in different contexts into a definite number of distinct word senses under a headword that provides language users with idealized descriptions of a word (Atkins & Rundell, 2008; Hanks, 2002; Kilgarriff, 2007; Mel’čuk, 1988). More particularly, according to Atkins and Rundell (2008, p. 273), “from the infinite number of individual situations in which a word appears, lexicographers derive a finite set of [lexical units] which collectively explain how that word contributes to the meaning of all of the individual events” and “which instantiate a one-to-many relationship (where one dictionary sense matches many language events)”. In this way, language users are expected to learn dictionary senses to prompt their interpretations of a word in various contexts where the word appears.

2.5.1 Dictionary word sense distinctions

2.5.1.1 A lexicographic approach to word sense distinctions

The identification of dictionary senses commences in the corpus-related analysis phase of compiling a dictionary and consists of several steps (Atkins & Rundell, 2008; Kilgarriff, 2013):

- (1) [analyzing] instances of usage, typically in concordances or lexical profiles, and
- (2) provisionally [identifying] different word senses (this is the subjective, intuitive part), then
- (3) [collecting] good, typical corpus examples for each of these provisional senses. As long as you have plenty of data, one-off oddities can usually be ignored, but ambiguous cases (the examples that you can't confidently assign to one or other of your provisional senses) should be stored for further analysis (step 5);
- (4) analysing each cluster of examples in turn, the lexicographer identifies the features that are typically associated with it (and that distinguish it from all the other clusters);
- (5) finally, our inventory of senses is refined if necessary (which may involve further splitting, or conversely, lumping of closely related clusters) so that all uses of the word that occur frequently in text are fully accounted for. (Atkins & Rundell, 2008, p. 312-313)

The fact that dictionaries mostly feature common words (as they are the core vocabulary of the language) which usually harbour more than one meaning (Jorgensen, 1990) implies that a major part of finding distinct word senses involves dealing with lexical ambiguity. The procedure of identifying dictionary senses is thus associated with word sense disambiguation, a term that has been commonly used in the field of natural language processing to refer to a machine-automated task of assigning senses to given word forms (discussed in more detail in Part 3).

The term “word sense disambiguation” (WSD) has been equally used in the field of lexicography, except that it does not embrace automation, as humans are involved in the task, i.e., lexicographers assign appropriate senses to word forms appearing in database-stored concordance lines (Step 2 in the above-described procedure for identifying dictionary senses). One method lexicographers use is to cluster corpus-derived sentences containing a word form that “exhibit similar patterning and meaning” together (Atkins et al., 2003; Kilgarriff, 1997, p.

92). They then assign a sense that represents corpus-based citations for a word form grouped in a cluster. The underlying idea behind this is that for an ambiguous word, the understanding of a sentence where the word appears is “built on the basis of just one of [its] meanings” (Kilgarriff, 1997, p. 91).

2.5.1.2 Challenges in word sense distinctions

Although sorting sentence citations sourced from corpora to distinguish word senses sounds straightforward, lexicographers in fact may not arrive at consistent sets of discrete, non-overlapping senses (Atkins et al., 2003; Grefenstette & Hanks, 2023; Kilgarriff, 1997). Two factors that are known to challenge this method are the *nature of word meanings* and the *subjectivity* it rests upon.

Word meaning is not a static but rather infinitely varied, context dependent entity which is not readily divided into distinct clusters (senses) (Kilgarriff, 2007). Kilgarriff (1997, 2007, p. 8) and Hanks (2000) agree that lexicographers often find corpus-derived citations contain “loose and overlapping word meanings, and standard or conventional meanings extended [...] and exploited in a bewildering variety of ways”. Here arises a challenge that is at the mercy of the complexity of lexical ambiguity. Kilgarriff (2007, p. 8) takes the classic example of *bank* to present this challenge. When it comes to a coarse-grained distinction between the word’s unrelated meanings (its homographs) “financial institution” and “edge of river”, *bank* has two clearly distinct senses. However, it becomes “bewildering” (Kilgarriff, 2007, p. 8) to make fine-grained distinctions among polysemous meanings of *bank* as “financial institution”, e.g., whether “a fund or reserve of money” and “a supply of something held in reserve (blood bank)”, to name just a few, should be treated as two distinct senses. A distinction between these two is harder to make than the one between “financial institution” and “river side” because they overlap, causing blurry boundaries around individual meanings. Not surprisingly, lexicographers, when distinguishing senses of polysemous words, have to “give a sharply

delineated presentation of something that is in fact fuzzy” (Hanks, 1990, p. 32) and that, according to Van der Meer (2004), tortures lexicographic practices.

Due to the fuzziness of word meaning, the underlying idea on which the method of clustering corpus-based citations to distinguish word senses relies is problematic. Although disambiguation is still achievable through the understanding of a corpus-derived sentence where a word appears (and where the understanding is based on one meaning), there is “no decisive way” of knowing when one meaning ends and another begins (Ayto, 1983; Kilgarriff, 2007, p. 29). Lexicographers thus must work toward less well-defined clusters (and consequently less well-defined senses) and tend to use their intuition in most judgements (Jorgensen, 1990; Stock, 2008). Since they “have strong intuitions about words having multiple meanings”, decisions they make on, for example, cluster/sense identifying (Step 2), splitting and/or lumping (Step 5) are therefore inevitably subjective (Atkins & Rundell, 2008; Kilgarriff, 1997, p. 92; Lew, 2013; Walter, 2010). Despite attempts to capture polysemous words systematically and consistently, a set of senses is “the product of the lexicographers’ intellectual labours” (Kilgarriff, 1997, p. 100), consequently varying according to each lexicographer’s subjective judgement. Kilgarriff (1997, p. 102) claims that “the identity test for a word sense in a particular dictionary is that two usages of the word belong to it if and only if the lexicographer would have put them in the same cluster”, given that lexicographers are unlikely to capture word senses in a systematic and unbiased way.

2.5.2 Dictionary word sense presentations

2.5.2.1 Sense enumeration

Dictionaries conventionally enumerate word senses in a vertical structure. In terms of macrostructure, word senses are often organized under alphabetically ordered headwords. In terms of microstructure, a dictionary entry is usually “a list of neatly separated, consecutively numbered lexical meanings” (Geeraerts, 1990, p. 198). The idea of numbering senses,

according to Atkins and Rundell (2008), may possibly originate from Johnson's (1755) work (see Figure 2.4).

To RESOU'ND. v.a.

1. To echo; to sound back; to celebrate by sound.
The sweet singer of Israel with his psaltery loudly *resounded*
the innumerable benefits of the Almighty Creator. *Peacham.*
The sound of hymns, wherewith they throne
Incompass'd shall *resound* thee ever blest. *Milton*
2. To sound; to tell so as to be heard far.
The man, for wisdom's various arts renown'd,
Long exercis'd in woes, oh muse! *Resound.* *Pope.*
3. To return sound; to sound with any noise
To answer and *resound* far other song. *Milton.*

Figure 2.4 Johnson's (1755) entry for *Resound* as cited in Atkins and Rundell (2008, p. 271)

This practice is supported by two “unarticulated” assumptions, which demonstrate what the five-step process of sense identification expects to achieve in terms of sets of discrete, non-overlapping senses.

first, that there is a sort of Platonic inventory of senses ‘out there’ (so if the dictionary says word W has N senses, it can't possibly have N – 1 or N + 2 senses)
second, that each sense is mutually exclusive and has clear boundaries (so if a specific occurrence of *keen* is assigned to sense 5, it cannot also belong to sense 6) (Atkins & Rundell, 2008, p. 271-272)

keen¹ *adj* **1.** Having a fine, sharp cutting edge or point. **2.** Having or marked by intellectual quickness and acuity. **3.** Acutely sensitive: *a keen ear* **4.** Sharp: vivid; strong: “*His entire body hungered for keen sensation, something exciting*” (Richard Wright). **5.** Intense: piercing: *a keen wind*. **6.** Pungent; acrid: *a keen smell of skunk was left behind*. **7.a.** Ardent;

enthusiastic: *a keen chess player*. **b.** Eagerly desirous: *keen on going to Europe in the spring* **8.** Slang Great; splendid; fine: *What a keen day!*

Figure 2.5 The entry for *Keen* with numbered senses (Adapted from Atkins & Rundell, 2008, p. 272)

Variations in implementing sense enumeration have been observed in different dictionaries. One option is organizing an entry based on grammar or meaning (Figure 2.6). The former tends to group meanings in regard to word classes, while the latter considers semantic proximity (distance) among meanings to group them.

haunt ► verb [with obj.] (of a ghost)	haunt ► (of a ghost) verb [with obj.]
manifest itself at (a place) regularly: <i>a grey lady who haunts the chapel.</i>	manifest itself at (a place) regularly: <i>a grey lady who haunts the chapel.</i>
■ (of a person of animal) frequent (a place): <i>he haunts street markets</i>	► (of a person of animal) verb [with obj.] frequent (a place): <i>he haunts street markets.</i>
■ be persistently and disturbingly present in (the mind): <i>the sight haunted me for years.</i>	■ noun a place frequented by a specified person: <i>the bar was a favourite haunt of artists of the time.</i>
■ (of something unpleasant) continue to affect or cause problems for: <i>cities haunted by the shadow of cholera.</i>	► be persistently and disturbingly present in (the mind) verb [with obj.]: <i>the sight haunted me for years.</i>
► noun a place frequented by a specified person: <i>the bar was a favourite haunt of artists of the time. [...]</i>	► (of something unpleasant) verb [with obj.]: continue to affect or cause problems for: <i>cities haunted by the shadow of cholera.</i>

Figure 2.6 The grammar-led (left) and meaning-led (right) entries for *Haunt* (Adapted from Atkins & Rundell, 2008, p. 247)

As shown in Figure 2.6, in a grammar-led entry, four meanings of *haunt* (v) are presented separately from the meaning of *haunt* (n), regardless of their semantic relations. In a

meaning-led entry, *haunt* (v, n) in the sense of “people frequently returning to a particular place” is put in one meaning group distinct from the other three.

Another option is presenting word senses in a flat or tiered structure, alternatively known as linear or hierarchical structure (Moerdijk, 2003, p. 285) (further discussed in Paper 3, section 6.2.2.2). In the flat structure, discrete senses are treated equally and numbered 1, 2, 3 and so on. In the tiered structure, meanings are categorized into main senses (numbered 1, 2, 3, ...) and sub-senses (numbered 3a, 3b, 3c, ...), indicating which sub-senses are nested under the same main sense (Figure 2.7).

<p>necessary (...) <i>adj</i> 1 needed to achieve a certain desired effect or result; required. 2 resulting from necessity; inevitable; <i>the necessary consequences of your action.</i> 3 Logic. 3a (of a statement, formula, etc.) true under all interpretations or in all possible circumstances 3b (of a proposition) determined to be true by its meaning, so that its denial would be self-contradictory. 3c (of a property) essential, so that without it its subject would not be the entity it is. 3d (of an inference) always yielding a true conclusion when its premises are true. 3e (of a condition) entailed by the truth of some statement or the obtaining of some state of affairs.</p>	<p>necessary (...) <i>adj</i> 1 needed to achieve a certain desired effect or result; required. 2 resulting from necessity; inevitable; <i>the necessary consequences of your action.</i> 3 Logic. (of a statement, formula, etc.) true under all interpretations or in all possible circumstances 4 Logic. (of a proposition) determined to be true by its meaning, so that its denial would be self-contradictory. 5 Logic. (of a property) essential, so that without it its subject would not be the entity it is. 6 Logic. (of an inference) always yielding a true conclusion when its premises are true. 7 Logic. (of a condition) entailed by the truth of some statement or the obtaining of some state of affairs.</p>
--	--

◆ Compare sufficient (sense 2). 4	◆ Compare sufficient (sense 2). 8
Philosophy. (in a nonlogical sense)	Philosophy. (in a nonlogical sense)
expressing a law of nature [...]	expressing a law of nature [...]

Figure 2.7 Entries for *Necessary* in tiered (left) and flat (right) structures (Adapted from Atkins & Rundell, 2008, p. 250)

Word senses within an entry are also variously ordered. The three common orderings are historical, frequency and semantic (Atkins & Rundell, 2008). The first method puts historically earlier senses before later ones, depicting the evolution of a word over time. The second method relies on corpus-based calculations of meaning frequency to order senses. A plausible assumption behind the frequency order is that senses with a higher frequency of occurrence in a corpus are more likely to be encountered by users and, thus, should be listed first.

<p>icon /.../ n [C] a small sign or picture on a computer screen that is used to start a particular operation: <i>To open a new file, click on the icon at the top of the screen.</i> 2 someone famous who is admired by many people and is thought to represent an important idea: <i>a 60s cultural icon.</i> 3 also ikon a picture or figure of a holy person that is used in worship in the Greek or Russian Orthodox Church.</p>

Figure 2.8 The entry for *Icon* with frequency-ordered senses. Definitions from *Longman Dictionary of Contemporary English*

The third method orders a word’s core meaning first, followed by its semantically closest ones. Other senses marginally relevant to the core meaning appear later, toward the end of an entry. The core meaning, the “psychologically salient” one, according to Atkins and Rundell (2008, p. 251), tends to be the meaning users learn first as a child, though it is not necessarily the one they encounter most frequently. Hence, the semantic order (with the core

meaning first) appears to “give the user the most satisfying account of meaning” (Atkins & Rundell, 2008, p. 251).

<p>icon /.../ (also ikon) noun a devotional painting of Christ or another holy figure, typically executed on wood and used ceremonially in the Byzantine and other Eastern Churches</p> <p>■ a person or thing regarded as a representative symbol or as worthy of veneration: <i>this iron-jawed icon of American manhood</i>. ■ Computing a symbol or graphic representation on a VDU screen of a program, option, or window. [...]</p>
--

Figure 2.9 The entry for *Icon* with core meaning first. Definitions from *Oxford Dictionary of English*

2.5.2.2 Challenges in word sense presentations

Although the vertical sense enumeration has long served as a default presentation of dictionary word senses, “the numbered lists of definitions found in dictionaries have helped to create a false picture of what really happens when language is used” (Hanks, 2000, p. 205). Atkins and Rundell (2008) advocate Hanks’s (2000) contention by pointing out a weakness of Johnson’s numbered sense idea, especially of the second assumption, via the entry of *keen*. They argue that the use of *keen* in the example of sense 6 (*a keen smell of skunk*) can potentially fit in senses 4 or 5. A reason for Atkins and Rundell’s (2008) argument lies in the difference between the structure of the mental lexicon and the vertical layout of dictionary word entries. Aitchison (2003, p. 13) claims that the “fluidity and flexibility of the mental lexicon [...] contrasts strongly with the fixed vocabulary of any book [dictionaries]”, revealing a problem in dictionary word sense presentation. This problem is articulated through “the fact that lexicographers [...] have to project a multidimensional clustered semantic structure onto the linear order of the dictionary” (Geeraerts, 2001, p. 14; Stock, 2008). With such “a desire to be neat and tidy”, dictionaries follow an alphabetical fashion (macrostructure) with countable,

fixed content (microstructure) so strictly that semantic considerations are consequently outweighed (by written words) (Aitchison, 2003, p. 11; Ostermann, 2015).

From the examples above, it can be clearly noticed that semantic similarities/differences are often overlooked in some methods of organizing word senses. Grammar-based organization—the commonest in lexicographic practices (Atkins & Rundell, 2008)—pays attention only to an exhaustive list of senses from the same word classes, regardless of varying semantic distance. This is a concerning disadvantage of the grammar-based entry, when closely related meanings may be placed far apart in the entry because they belong to different word classes, e.g., *haunt* (v): (of a person or animal) frequent (a place) and *haunt* (n): a place frequented by a specific person. The flat-structured organization, although not likely to widen the distance between related senses, causes difficulties in distinguishing major and minor senses. Jorgensen (1990, p. 185) warns that treating “all senses as equally important and equiprobable” may be misleading for dictionary users who do not know words in the first place. The grammar-based and flat-structured entries are, however, still prevalent in dictionaries because they can be applied objectively and systematically, even if they are semantically implausible. Other methods of structuring dictionary entries that are meaning-centred, e.g., meaning-based organization and semantic order, are less common, even though they are more considerate of semantic relations. This may be because word meaning is less clear-cut than, for example, word class. The application of meaning-based organization thus entails more subjectivity than grammar-based organization. Similarly, despite being semantically plausible, the semantic order is “the least scientific” (and “a compromise solution”) among the three orderings because lexicographers must intuitively identify a core meaning (Atkins & Rundell, 2008, p. 251).

Taking a closer look into homography and polysemy, concerns about the minimal treatment for semantic relations that some sense enumeration methods offer are heightened. In

theory, homographs represent discrete headwords that accidentally share identical orthographic forms; therefore, they should be treated in separate entries (and are usually given superscript numbers as in Figure 2.10).

bear ¹ (beə) <i>vb.</i> bears, bearing, bore, borne. (mainly <i>tr.</i>) 1 to support or hold up; sustain. 2 to bring or convey <i>to bear gifts.</i> 3 to take, accept or assume the responsibility of: <i>to bear an expense</i> 4 (<i>past participle</i> born in passive use) [...]	bear ² (beə) <i>n. pl.</i> bears or bear 1 any plantigrade mammal of the family <i>Ursidae</i> : order <i>Carnivora</i> (carnivores). Bears are typically massive omnivorous animals with a large head, a long shaggy coat, and strong claws [...]
---	---

Figure 2.10 Homograph entries for *Bear*. Definitions from *Collins English Dictionary*

Nevertheless, separate entries are only sometimes reserved for homograph headwords in lexicographic practices. Atkins and Rundell (2008, p. 192-193) state that since the classical perception of homographs as words with identical spellings is somewhat vague, lexicographers must make case-by-case decisions on whether there should be separate entries for homograph headwords:

Case 1 – Same spelling, different meaning and etymology

*bear*¹ “animal” and *bear*² “carry, tolerate, support”

Case 2 – Same spelling, different meaning and pronunciation

*tear*¹ /tɪə/ (from weeping) and *tear*² /teə/ (in paper, cloth)

Case 3 – Same spelling and pronunciation, different meaning and capitalization

*may*¹ (modal verb) and *May*¹ (month)

Case 4 – Same spelling and pronunciation, different meaning

*bank*¹ “edge of river” and *bank*² “financial institution”

For cases (2) and (3), lexicographers tend to give a separate entry for each homograph.

The reason for generating two homograph headwords for each pair of words is to appropriately

present their different pronunciations (*tear*¹ and *tear*²) and capitalizations (*may*¹ and *May*²). For case (1), when pronunciation and capitalization are not applicable, etymology is a common rule for lexicographers. The consideration of etymology is, however, useful in historical dictionaries because the central function of these types of dictionaries is to describe the word's development (Atkins & Rundell, 2008; Van der Meer, 1997). Homographs thus always appear as separate entries with complete descriptions of their origins in historical dictionaries. Most current dictionaries, however, especially learners' dictionaries, rarely pay attention to the etymology of homographs or even to homographs themselves (e.g., they do not divide separate entries for homographs). This may be due to doubts over the value of homography to a synchronic account of meaning. Atkins and Rundell (2008) argue that few language learners know the origin of words, so the division of entries for homographs seems "pointless", i.e., "the connections – or lack of them – among the various uses of a word form will not necessarily be obvious" (p. 282). Moon (1987, p. 88) agrees with Atkins and Rundell (2008) and further reasons that

... because access to an item is through its orthographic form, and because etymological [homography] depends on knowledge that is not available to the dictionary user before he or she locates the word in the dictionary, it was decided to ignore [homography] completely.

For (4), Atkins and Rundell (2008, p. 193) state that the difference in meanings of *bank*¹ and *bank*² (and possibly of their polysemes) is "a grey area, and there are no clear criteria for lexicographers to apply (and of course the user looking up a word often does not know its meaning)". This may explain why homographs of *bank*, and of other words belonging to this type, are all put in a single entry (with polysemous senses) under the same headword, possibly in a flat structure.

Regarding polysemy, sense numbering causes difficulties in putting overlapped meanings into discrete senses (as discussed in the case of *keen*). The two commonest organizations of meanings—grammar-based and flat-structured entries—largely disregard

semantic relationships among polysemous meanings. The former tends to widen the gap between related senses of different word classes and the latter does not show semantic hierarchy. The tiered structure, which is expected to address semantic drawbacks of the grammar-based and flat structures when presenting polysemy, is still not considered a satisfactory method, especially in dealing with domain-specific meanings. L'Homme (2020) exemplifies how the hierarchical alphanumeric systems in general dictionaries expose limitations in showing semantic links between domain-specific senses with the presentation of *green* in the *Oxford English Dictionary*. Her example is taken from the field of environment.

<p>green</p> <p>I. With reference to colour.</p> <p>[...]</p> <p>2. Of a colour intermediate between blue and yellow in the spectrum; of the colour of grass, foliage, an emerald, etc.</p> <p>2a. Covered with or abundant in foliage or vegetation; verdant; (of a tree) in leaf. Also in extended use.</p> <p>[...]</p> <p>III. In extended uses</p> <p>[...]</p> <p>13b. Of a product, service, etc.: designed, produced, or operating in a way that minimizes harm to the natural environment.</p>
--

Figure 2.11 The entry for *Green*. Definitions from the *Oxford English Dictionary*

As can be seen in Figure 2.11, although senses are categorized neatly into main and sub-senses, the numbering of senses accidentally pushes the two environmental senses of *green* (2a and 13b) away from each other. L’Homme argues that even though 2a and 13b are remotely (i.e., metaphorically) linked, it is necessary to pull them closer rather than push them further apart with several general meanings in between. The issue L’Homme (2020) raises here is relevant to arguments against the application of numbering senses in presenting metaphoric extensions (irregular polysemy). The fact that literal and figurative senses are treated as discrete entities (Moon, 2004) (as observed in senses 2a and 13b of *green*) weakens the link between them. The numbered senses hinder the association between the figurative use of *green* (e.g., 13b) with its actual meaning (e.g., 2a) (L’Homme, 2020; Van der Meer, 1997), making it more difficult for users to gain a subtle understanding of relations between figurative and literal uses.

Furthermore, the numbered senses in frequency order, where figurative uses sometimes happen to be placed first (or before literal meanings), also lead to problems of understanding. Although the assumption that putting the most frequently occurring meanings first (because they are most likely to be looked up by users) sounds plausible, in such general dictionaries as *Oxford Advanced Learner’s Dictionary*, *Cambridge International Dictionary of English*, *Collins COBUILD English Dictionary* and *Longman Dictionary of Contemporary*, first-placed meanings are usually figurative (Van der Meer, 1997, 1999). This also happens to a few cases in specialized dictionaries, for example, *benign* in *Merriam-Webster Medical Dictionary* (see in 6.2.2.2). As Van der Meer (1997, 1999) claims that “the figurative uses of a specific word cannot be fully understood except by reference to its literal meaning” (p. 196), the literal (basic) meaning should be placed first, or at least its relation to other figurative uses should be explicit (further discussed in 6.2.2.2), to facilitate users in “the realisation that meanings may be related to other, more basic meanings” (p. 559). In other words, vocabulary development, i.e., (metaphoric) meaning extension, should be adequately considered and explicitly presented in

both general and specialized dictionaries, as suggested by Scholfield (1999), because it is a vital aspect of vocabulary learning.

A run-on may additionally confuse users, especially when it is embedded in an entry of a polysemous word with numbered senses (e.g., *naked*, in Figure 2.3). Although Atkins and Rundell (2008, p. 237) describe three characteristics (in 2.4) that run-on sections need to possess to avoid confusing users, these may only be existent in entries of words with a single meaning. For words with polysemous meanings, the first characteristic, meaning(s) of words in a run-on “unambiguously” retrieved from word-formation rules, seems hardly to be retained. In the case of *naked*, for example, there is no further information about meanings of *nakedly* and *nakedness* except their parts of speech. This lack of semantic indication may lead to users knowing little about which senses of *naked* the words *nakedly* and *nakedness* semantically link with.

2.6 Lexical semantics and lexicography

Since lexicographers work on the same ground as semantic linguists (mentioned in 2.5), root causes of challenges in sense distinctions and presentations may be associated with lexical semantic issues, particularly the nature of word meanings and their representations in the mind.

Dictionary word sense distinctions have faced a deep-rooted problem in lexical semantics—the fuzziness involved in the disambiguation of words. Meanings of a word, especially of a polysemous one, are perceived, in lexical semantics, as being overlapping. Word senses in lexicography, idealized manifestations of word meanings in lexical semantics, should therefore be constructed with respect to the fuzzy characteristic of word meanings. Nonetheless, the lexicographic approach to distinguishing word senses seems to disregard (and consequently, only superficially address) the fuzziness of word meanings (Copestake & Briscoe, 1995; Kilgarriff, 1992; Kwong, 2013) when establishing clear-cut boundaries around individual senses, with little consideration of lexical semantic theories. Kilgarriff (1992, 1997),

therefore, casts doubt upon the conception of word sense because of the lack of theoretical foundations on which this concept, and the entire procedure of identifying a distinct word sense, is based. He states that lexicographers may need to rethink “what sorts of distinctions the dictionary [makes]” and “what rationale underlines them” (Kilgarriff, 1992, p. 365).

Dictionary word sense presentations structurally relate to mental representations of word meanings. The previously discussed analogy between a dictionary and a lexical entry in lexicography and lexical semantics uncovers the connection between a dictionary and the mental lexicon. Although at first glance there appear to be different perspectives from which these two describe words, word meanings presented in dictionaries are ultimately subject to human perception. That means dictionaries are expected to offer “access points” to a bigger picture of the language in one’s mind (Ostermann, 2015, p. 65), given that word sense presentations in dictionaries should follow semantic principles in accordance with those that govern how word meanings are represented mentally. However, alphanumeric systems, in which dictionaries present word senses vertically, actually contrast with the multidimensional structure of word meanings in the mental lexicon (Aitchison, 2003; Geeraerts, 2001; Miller, 1986; Stock, 2008). A consequence of the contrast between presentations of dictionary word senses and mental representations of word meanings is that semantic relations, i.e., homography and polysemy, are buried under a “spuriously neat” view that dictionaries favour (Aitchison, 2003, p. 14; Ostermann, 2015).

In short, lexicographic challenges are rooted in (a) the lack of lexical semantic theories that underpin the concept of word senses and (b) the discrepancy between dictionaries and the mental lexicon. These unresolved issues entail (a) sets of word senses that are inconsistently captured and (b) lexicographic formats that are unlikely to facilitate the processing of semantic information. To overcome the challenges, it is suggested to combine lexical semantics, more precisely, cognitive lexical semantics, and lexicography (Aitchison, 2003; Csábi, 2002;

Geeraerts, 2001, 2007; Ostermann, 2015). A rationale behind this suggestion is that cognitive lexical semantics “adequately describes how language users process language” (Ostermann, 2015, p. 48). It potentially theorizes the fuzziness and conceptualizes the “multidimensional, clustered semantic structure” of words (Geeraerts, 2001, p. 14, 2007), consequently proposing “a framework for analysis and description that will do least distortion to evidence” (Hanks, 2008, p. 221). The application of cognitive lexical semantic research findings in the making of dictionaries, therefore, would improve and enrich traditional elements of dictionary content and structure. More particularly, this would offer possible solutions to dictionary word sense distinctions and presentations, making dictionary content and structure more realistic and efficient (Csábi, 2002).

Theories in cognitive linguistics have exerted an impact on different aspects of lexicographic practices. Regarding the lexical ambiguity-related challenges in the division of word senses and microstructure of dictionaries, Rosch’s (1973) Prototype theory, Lakoff’s (1987) linguistic categories and Tyler and Evans’s (2004) Principled Polysemy are highly relevant theories that have great value for addressing the unresolved issues (Atkins, 2008; Atkins & Rundell, 2008; Béjoint, 1990, 2000; Geeraerts, 1990; Ostermann, 2015; Rundell, 2012). Prototype effects in language have had significant implications for lexicography because of their capacity to deal with fuzzy categories. The fundamental principle of prototype effects—the family resemblance, which demonstrates “networks of overlapping attributes” (Rosch & Mervis, 1975, p. 575)—underlies explanations of “membership and the position of members in fuzzy category structures” (Ostermann, 2015, p. 53). This principle helps prototype theory “accurately [model] the kind of semantic phenomena that lexicographers have to face up to”, thereby making prototype theory “well suited as a theoretical basis” (Geeraerts, 1990, p. 210) for the task of identifying distinct word senses. Atkins and Rundell (2008) specify that a major part of the task to which the prototype theory has made contributions is WSD because:

It reflects the way people create meanings when they communicate, and thus it goes with the grain of the language, and accommodates creativity and fuzziness. It makes the lexicographer's task more manageable, because it allows us to focus on the prototype and its common exploitations, rather than requiring us to predict and account for every possible instantiation of a meaning. (p. 280)

Rundell (1988, p. 134) adds that “learners will be better served by accounts of word-meaning based on a prototype approach, which deals in core meanings that admit of minor variation and degrees of category membership”. The influence of prototype theory in lexicographic practices is also substantially extended via Lakoff's (1987) linguistic categories and Tyler and Evans's (2004) *Principled Polysemy*. The knowledge of motivations that Lakoff (1987) generates to account for meanings of polysemous words has additional benefits for the structure of dictionary entries (Atkins, 2008; Csábi, 2002). He elucidates the conceptual mechanism, i.e., the conventionalization that motivates how a prototypical meaning is extended to less prototypical meanings via a radial model (as discussed previously in the case of *mother*). Since less prototypical meanings, which are usually figurative ones, “appear to be used in a variety of unrelated senses”, they are “easily confusable” and pose problems for lexicographers (e.g., environmental meanings of *green*) (Csábi, 2002, p. 250). Lakoff's mechanism (and model) of polysemous extension, which *Principled Polysemy* parameterizes, can thus assist lexicographers in describing polysemous words, especially their figurative meanings, in a systematic and effective way. More particularly, they may implement the mechanism (and model) when structuring dictionary entries of polysemous words (Dalpanagioti, 2018) and ensure that “conceptual links between words and their meanings should be made explicit whenever possible” (Csábi, 2002, p. 250). In this way, dictionaries will help raise dictionary users' (e.g., language learners') awareness of connections among polysemous meanings, thereby enabling users to master word meanings, especially figurative ones, more easily.

Despite the positive benefits of cognitive lexical semantics, there has been little trace of cognitive lexical semantic elements in current dictionaries. The prototypical principle of

categorization can only be detected in, for example, the entry of *climb* in the *Oxford Dictionary of English* (Atkins & Rundell, 2008, p. 279; Hanks, 1994; Rundell, 2012). The dictionary structures the entry of *climb* with the verb's prototypical meaning of "ascend" (core meaning) appearing first, followed by its sub-senses that "approximate in varying degrees to the prototype" (Rundell, 2012, p. 279):

A car may climb up a steep hill.

A plane may climb into the air after take-off.

A column of smoke may also climb into the air.

A plant may climb up a wall.

A road or path may climb up the side of a hill.

The conceptual mechanism which is systematically described in Principled Polysemy is seemingly absent from current dictionaries, though it has inspired investigations (Dalpanagioti, 2018; Ostermann, 2015) into pairing cognitive lexical semantic principles with aspects of lexicographic practices to improve or replace some traditional elements in dictionaries. Investigations into implementing Principled Polysemy in dictionaries have, however, been restricted to prepositions, abstract nouns and verbs (as mentioned in 2.3.2.2). The radial model has yet to be depicted in current dictionaries.

The lack of a cognitive lexical semantic approach in dictionaries may result from the view of some lexicographers that contributions of linguistics (in general) and cognitive lexical semantics (in particular) to practical lexicography are for "consciousness-raising discussion rather than immediate applicability" (Atkins, 2008, p. 42; Rundell, 2012, p. 70). A big part of the lexicographic task is eventually to make a good decision by "trying to find the underlying regularity" (Zgusta, 1992, p. 92). Such theories as Lakoff's radial categories, according to Atkins (2008, p. 48), have "a great bearing on practical lexicography" because they guide lexicographers to discover the 'underlying regularity'.

However, these theories are “not a prerequisite for being a proficient lexicographer – still less a guarantee of success in the field” (Atkins & Rundell, 2008, p. 130). An awareness of these theories can only help lexicographers be better placed to “perceive order and system in the apparent randomness of language” (Rundell, 2012, p. 48). Lexicographers may be well prepared to analyze word senses and produce dictionary entries if they understand the mechanism by which different word meanings develop. This may not, however, necessarily make practical tasks of identifying word senses, for example, any easier, but understanding the underlying system will only help lexicographers tackle this job with greater confidence. If they feel confident, they will “make good judgments in the more marginal, less clear-cut cases” and keep subjective judgments to a minimum (Atkins & Rundell, 2008, p. 294; Rundell, 2012).

Even considering the long-standing relationship between lexical semantics and lexicography, cognitive lexical semantic theories, according to some influential lexicographers, have been explicitly acknowledged but have yet to be applied. Regarding immediate applicability, these theories “have not been found convincing by the [lexicographic] community” (Béjoint, 2010, p. 381). Rundell (2012, p. 71) states that since “lexicographers and linguists have different agendas”, it may require time for “a process where linguistic theories need to be adapted in order to be of use in the specific environment of a dictionary”.

PART 3 – WORD SENSE DISAMBIGUATION AND CORPUS LINGUISTICS

2.7 Key concepts

Word sense disambiguation (WSD) is one of the oldest problems in the field of natural language processing (Cohn, 2003; Ide & Véronis, 1998; Mihalcea, 2006). WSD has been considered ever since computers were used to resolve issues in human language, i.e., lexical ambiguity. It was first introduced in some of the earliest work in the 1940s on machine translation, e.g., Weaver's (1955) memorandum on translation (Hutchins, 1999). WSD is considered a distinct computational task, which is relevant when a word has multiple meanings, "to determine which sense of a word is activated by the use of the word in a particular context" (Edmonds, 2006, p. 2).

The task usually involves two phases: (1) the determination of all possible senses for a word in a text and (2) the assignment of each occurrence of a word to its appropriate sense (Ide & Véronis, 1998; Schuemie et al., 2005). Kwong (2013) offers a general model for the current practice of WSD (Figure 2.12). In this model, WSD is often grounded on a sense inventory and viewed as a task which "compares the Triggering Context (TC), that is, the actual context embedding a new occurrence of a word, with the Conventionalised Context (CC) characterising individual senses of the word, to find the sense with the closest resemblance between TC and CC" (Kwong, 2013, p. 16).

To start with, a *sense inventory* "listing a finite set of predetermined senses for different words" is often required (Kwong, 2013, p. 11). This list of senses is linked with *lexical resource(s)* that "provide the lexico-semantic information of the senses, by means of definitions, semantic relations, or other forms of knowledge representation" (Kwong, 2013, p. 11). Such lexical resources are varied, and are broadly categorized as structured or unstructured (Vidhu Bhala & Abirami, 2014). Dictionaries, thesauri and computational lexicons are structured resources. Dictionaries have been used since the inception of WSD. As computing

power increased in the late 20th century, dictionaries and thesauri were scaled up and supplemented with richer semantic interrelationships. Machine-readable dictionaries and Roget’s *International Thesaurus* became widely used, giving rise to dictionary-based WSD (Agirre & Edmonds, 2006, p. 6). WordNet later emerged as a computational lexicon, hierarchically organizing word senses into synsets (groups of synonymous words). It is computationally accessible and has been one of the most-used resources in WSD research. During the same time, corpora became available, constituting large-scale, unstructured resources. Corpora are digital collections of texts having “related or similar topics that may be raw, sense-annotated (manually or automatically) or may represent word collocations” (Vidhu Bhala & Abirami, 2014, p. 164). The advent of corpora preceded the emergence of corpus-based WSD.

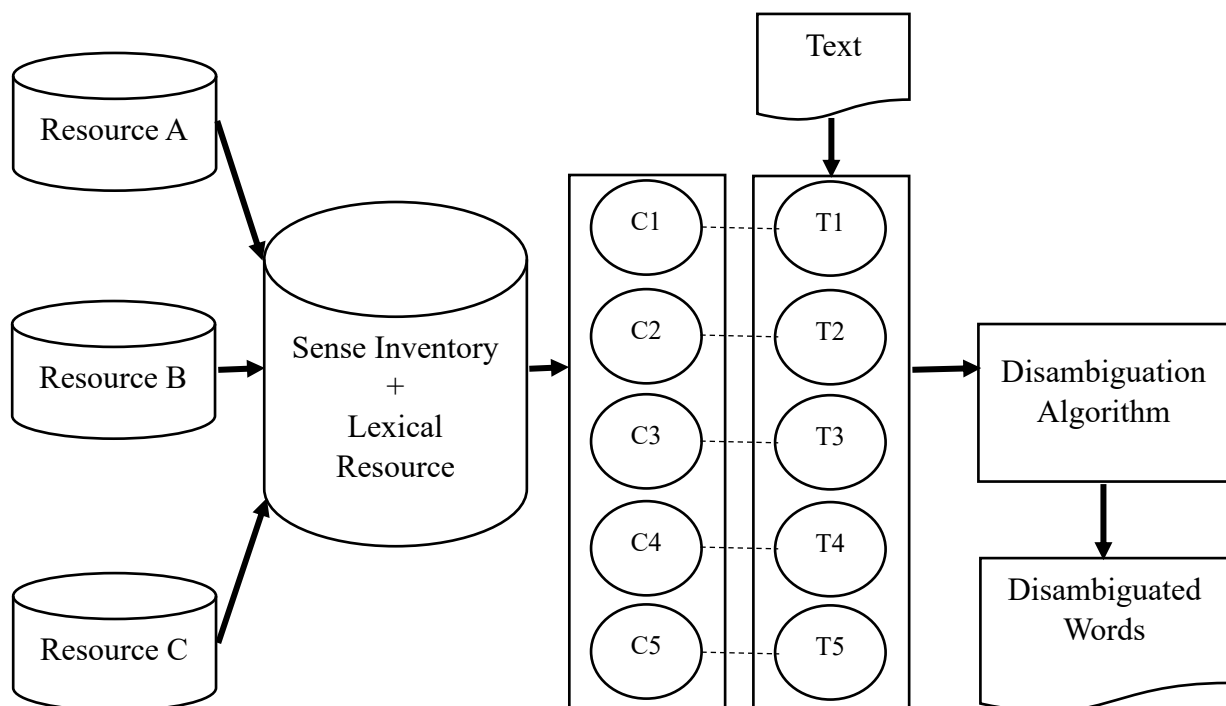


Figure 2.12 A general model of WSD (Adapted from Kwong, 2013, p. 16)

Knowledge from lexical resources constructs the CC of word senses (indicated by C1, C2, ..., C5) to be compared with the TC (indicated by T1, T2, ..., T5)—the actual context

embedding the occurrence of the word. An example of this process is illustrated in Covington et al.'s (1994) disambiguation of the word *pen* in the sentence “The pen is full of pigs”. A lexical resource was consulted to provide predetermined senses of *pen*: “a writing instrument” and “an enclosure for livestock”, which were respectively labelled as *pen_for_animals* and *pen_for_writing*. In Covington et al.'s resource, word senses were characterised by identifying different groups of cue words. Accordingly, cue words for *pen_for_animals* were *farm* and *pig* and cue words for *pen_for_writing* were *ink* and *paper*. The CC, in this case, is thus in the form of cue words and the WSD task was to match the cue words in the CC with those in the TC. Consequently, the match between *pig* in both CC and TC assigned the sense *pen_for_animals* to *pen* in the sentence “The pen is full of pigs”.

2.8 WSD approaches

2.8.1 Methods of WSD

Kwong (2013) states that lexical resources and disambiguation algorithms respectively determine *what* knowledge (i.e., what is available as CC) and *how* it is compared with the TC (Figure 2.12). Depending on kinds of lexical resources and algorithmic approaches adopted, WSD methods are classified into knowledge-based, supervised, unsupervised and semi-supervised methods (Agirre & Edmonds, 2006; Ide & Véronis, 1998; Kwong, 2013; Màrquez et al., 2006; Pedersen, 2006).

A *knowledge-based* (also known as *dictionary-based*) method takes advantage of established knowledge available in lexical resources such as dictionaries and thesauri. This method is developed from a hypothesis that knowledge encoded in dictionaries and thesauri, including definitions of words and relations between words, can contribute to selecting a correct sense for an ambiguous word (Chen et al., 2009; Schuemie et al., 2005). Lesk (1986), who proposed to use dictionary definitions for WSD, is considered the forerunner of the dictionary-based method. Lesk disambiguated two words by “finding the pair of senses with

the greatest word overlap in their dictionary definitions” (Chen et al., 2009, p. 28). Following this lead, others started to use and evaluate the dictionary-based method, e.g., Wilks et al. (1990) conducted experiments on the classic word *bank* and achieved an accuracy of 45% on the identification of polysemous senses and 90% on homographic senses. This result indicates that the dictionary-based method is useful for homographs but not robust for polysemous cases of word sense ambiguity, as “dictionaries lack complete coverage of information on sense distinctions” (Agirre & Edmonds, 2006, p. 6).

To surmount the problem of limited coverage of word sense information in the dictionary-based method, several researchers decided to construct knowledge about word senses and uses through “learned information derived from statistics over large corpora” (Levow, 1997, p. 2), rather than utilizing precoded knowledge in dictionaries and thesauri. Since then, WSD has progressed to a *supervised corpus-based* method: “a classifier is trained for each distinct word over a corpus of manually-annotated examples of each word in context” (Edmonds, 2006, p. 3). The supervised learning method requires a large dataset of sentences with the occurrence of an ambiguous word labeled by hand with the correct sense, where machine learning techniques are applied to achieve the sense classification. Even though systems (e.g., Bayesian learning and support vector machines) that employ the supervised learning method have become mainstream, with the best performance in evaluation exercises, they have been constrained by a significant problem—the so-called knowledge acquisition bottleneck (Abed et al., 2016; Agirre & Edmonds, 2006; Başkaya & Jurgens, 2016; Gale et al., 1992a; Kwong, 2013). Since the classifier can only distinguish between senses present in the training set, the supervised learning method ideally needs a significant amount of sense-annotated data per ambiguous word to be accurately executed (Chen et al., 2005). This, however, is likely to be impractical, because it requires tremendous effort and time to acquire

sufficient contextual information for every sense of a large number of words existing in the natural language (Abed et al., 2016; Chen et al., 2009; Schuemie et al., 2005).

Due to the challenge of creating adequately sense-annotated corpora in the supervised method and the dearth of word sense information in the dictionary-based method, using unannotated corpora without a reference to fixed sense inventories seems to be the most promising solution for a WSD task (Gale et al., 1992a). An *unsupervised method* therefore emerged, with a different aim as compared to knowledge-based and supervised methods, i.e., “[detecting] sense clusters instead of allocating sense labels” (Abed et al., 2016, p. 227). This method neither requires a labelled dataset nor takes advantage of any machine-readable resources such as dictionaries or thesauri. Rather, it relies on an assumption that the same senses of a word tend to have similar neighbouring words and employs a technique of word sense induction. In this regard, word senses are induced through the input of unannotated corpora via clustering of word occurrences. The unsupervised method carries out word sense disambiguation by first examining contexts (corpora) where a word is used to form prompted clusters (senses) and then distributing occurrences of the word that contain similar corpus-based evidence into appropriate clusters (senses) (Schütze, 1998). In this way, two goals are achieved: (1) knowing senses of a word induced from text and (2) disambiguating a new usage of the word as an instance of one of the induced senses. However, although the unsupervised method overcomes the knowledge acquisition bottleneck problem, it is yet to outperform the supervised method.

Although the supervised method addresses the dearth of word sense information in the dictionary-based method, current sense-annotated corpora do not contain a sufficient number of instances per word to train supervised systems for all words. While the unsupervised method has been proposed to overcome this data sparsity problem, it has not surpassed the accuracy of the supervised method, nor does it take advantage of what sense-annotated data is available

(Kilgarriff & Rosenzweig, 2000; Mihalcea et al., 2004). A *semi-supervised* method has been introduced as an alternative to supervised and unsupervised WSD. This method typically uses “corpora of unlabeled data and a small amount of labeled data to build a more accurate classification model than would be built using only labeled data” (Zhou & Meng, 2019, p. 143). The semi-supervised method has recently received significant attention because it provides advantages over the supervised and unsupervised methods. The semi-supervised method can potentially remove the knowledge acquisition bottleneck of requiring significant amount of sense-annotated data in the supervised method, as it makes use of unannotated corpora. Although using unannotated data, the accuracy of classifiers in the semi-supervised method is generally more improved than in the unsupervised method as the classifiers are built up by bootstrapping from a small amount of annotated data.

2.8.2 One-sense-per-collocation method

One semi-supervised method that ensures high accuracy of WSD in large, unannotated corpora, eliminating the need for costly hand-annotated training data, is the one-sense-per-collocation method (Yarowsky, 1993). This method implements two heuristics that are powerful properties of human language and widely accepted by natural language processing community: (1) one sense per discourse and (2) one sense per collocation.

The former was introduced by Gale et al. (1992b), suggesting that a word tends to preserve its meaning in a given discourse. The latter, which was introduced by Yarowsky (1993), is “similar in spirit to the one-sense-per-discourse [heuristic but] has a different scope” (Mihalcea, 2006, p. 124). It postulates that a word tends to preserve its meaning when used in the same collocation. In other words, nearby words provide useful and consistent clues to disambiguate senses of a target word (a node).

By saying “collocation”, Yarowsky (1995, p. 189) means “a juxtaposition of words [in which no] idiomatic or non-compositional interpretation is implied”. He examines the co-

occurrence of two words in several defined relationships, from which he categorizes different types of collocation and measures their effects on WSD. More particularly, he considers (1) direct adjacency (first word to the left and/or right of a node), (2) syntactic relations (verb/object, subject/verb and adjective/noun) and (3) word classes of collocations (content and function words).

The effects of these types of collocation on WSD are evaluated by observing the “tendency for [target] words to exhibit only one sense in a given collocation” (Yarowsky, 1995, p. 190). He found out that adjacent collocations had a stronger tendency to indicate a particular sense of their node than non-adjacent collocations, and the tendency weakened with distance. It was stronger for collocations in a predicate-argument relationship and collocations with content words (nouns, verbs, adjectives, and adverbs) than those in “arbitrary associations at equivalent distance” and with function words (Yarowsky, 1995, p. 190). Content words from different word classes tended to behave differently from each other in relation to their nodes.

Verbs, for example, derive more disambiguating information from their objects (.95) than from their subjects (.90). Adjectives derive almost all of their disambiguating information from the nouns they modify (.98). Nouns are best disambiguated by directly adjacent adjectives or nouns, with the content word to the left indicating a single sense with 99% precision. Verbs appear to be less useful for noun sense disambiguation, although they are relatively better indicators when the noun is their object rather than their subject. (Yarowsky, 1993, p. 269)

In sum, Yarowsky’s (1993, 1995) findings provide reliable indicators that are useful for sense disambiguation. From these findings, the Yarowsky bootstrapping algorithm, which is generally an “iterative and incremental” algorithm (Mihalcea, 2006, p. 181), was developed. It initializes by sense-tagging a small number of examples to build a simple classifier based on a decision list, which is a set of seed collocations. These seeds first include some collocations that are strongly indicative of each sense of a target word, for example, including *life* and *manufacturing* as seed collocations for two senses “living” and “factory” (labeled as sense A and B, respectively) of the target word *plant*. The classifier is then used to tag a few more new

contexts through which more seed collocations are added, expanding the decision list (see Table 2.1). This whole process is repeated until a large amount of data is sense-tagged.

Table 2.1 An excerpt of the decision list for *Plant* (Adapted from Yarowsky, 1995, p. 191)

Collocation	Sense
life (within ± 2 -10 words)	A
manufacturing (within ± 2 -10 words)	B
animal (within ± 2 -10 words)	A
equipment (within ± 2 -10 words)	B
employee (within ± 2 -10 words)	B

Yarowsky’s experiment on disambiguating two senses of *plant* achieved an overall precision of above 90% across a large set of hand-annotated examples. However, although the percentage is high, this achievement is only for coarse-grained sense distinctions, in other words, distinctions of homographs, words with 2-way ambiguity like *life plant* versus *manufacturing plant*. Martínez and Agirre (2000) found that the precision of the one-sense-per-collocation method drops significantly to about 70% or even less when words with higher degrees of ambiguity (e.g., polysemy) are considered.

2.8.3 Problems in WSD

Problems in the WSD task are attributable to homography and polysemy (Abed et al., 2016). WSD problems with homography, which are related to coarse-grained distinctions, are satisfactorily addressed with homographic disambiguation achieving above 90% accuracy (Edmonds, 2006; Wilks et al., 1990; Yarowsky, 1993, 1995). However, WSD problems with polysemy, which are related to fine-grained distinctions, remain unsolved, as polysemous disambiguation has not yet surpassed 70% accuracy (Martínez & Agirre, 2000). Since the

number of polysemous words in a natural language is significant, more issues relating to WSD errors lie in polysemous than homographic senses (Abed et al., 2016; Dash, 2012).

Issues have emerged from “the traditional conception of WSD via an explicit sense inventory” (Ide & Wilks, 2006; Kwong, 2013, p. 2; McCarthy, 2006; Resnik, 2006). As for the dictionary-based method that refers to sense inventories derived from structured lexical resources such as dictionaries, problems in polysemous (and homographic) disambiguation are linked with the lexicographic challenges (discussed in Part 2). Ide and Wilks (2006, p. 64) state that “the WSD community has grappled for years with the issue of sense distinctions because of its reliance on pre-defined sense inventories provided in monolingual dictionaries”. Since the organization of senses in these inventories usually adheres to lexicographic principles (e.g., grammar-based organization or frequency order), the indication of the degree of distinguishability among polysemous and homographic senses is not often explicit.

Moreover, the intuitive guidance lexicographers use to split and/or lump senses causes the number of senses to vary considerably in different dictionaries. Consequently, senses in different dictionaries (or thesauri) are rarely compatible, challenging WSD tasks that only utilize a single source as a sense inventory (Cohn, 2003). Not only the number but also the granularity of senses exacerbates problems for WSD. The granularity of senses has been one of the most often-cited obstacles to WSD research, since several authors have questioned the suitability of sense granularity from dictionaries and other lexical resources alike for WSD (Edmonds, 2006; Ide & Wilks, 2006; Ide & Véronis, 1998; Kwong, 2013; McCarthy, 2006). Ide and Wilks (2006) and Kwong (2013, p. 38) remark that “most WSD researchers have relied on the sense distinctions in existing lexical resources, typically machine-readable dictionaries or WordNet” that are often too fine-grained for the purposes of WSD. Ide and Véronis (1998, p. 22) clarify that “overly fine sense distinctions create practical difficulties” for WSD systems that rely on, for example, the *Longman Dictionary of Contemporary English* and WordNet. A

machine-readable version of the *Longman Dictionary of Contemporary English* appears to be unrealistic for automatic WSD, as even human analysts still cannot reliably distinguish fine-grained senses in this dictionary (Ide & Véronis, 1998; Ide & Wilks, 2006; Kilgarriff, 1992). WordNet senses are criticised for “being [so] unrealistically fine-grained, and sometimes overlapping” that it is not easy for even human analysts to reach high agreement on distinguishing such senses (Kwong, 2013, p. 45), creating “the WordNet problem” for automatic WSD (Edmonds, 2006; Ide & Véronis, 1998; Ide & Wilks, 2006, p. 52).

Dictionary-based inventories have several disadvantages because dictionaries (and thesauri) are not designed for WSD researchers and are “subject to standard market pressures, which dictate the size of the dictionary, the coverage and depth, and crucially the granularity and interpretation of sense distinctions” (Edmonds, 2006, p. 13). WordNet initially emerged from “a psycholinguistic project on network models for the mental lexicon” that had no connection with WSD but later became one of the most popular semantic lexicons used by WSD researchers “despite its relative weakness in capturing syntagmatic relations” (Kwong, 2013, p. 58). All of these issues bring about the problem of limited coverage of word sense information discussed earlier in the dictionary-based method. Dictionaries have been criticised for “leaving out some common sense information that would [be] very useful in WSD” (Edmonds, 2006, p. 15). Kwong (2013, p. 12) and Véronis (2001) add that dictionaries and other similar reference materials (e.g., WordNet) “often lack distributional criteria like syntactic and collocational information which are usually required to match a given sense with a new occurrence”. Also, dictionaries usually contain static sets of senses that are not often updated, thereby rendering them unable to cover new usages of words (Kwong, 2013).

Since senses from structured lexical resources do not appear to match those that are required by WSD systems, recent efforts have resorted to deriving sense inventories (or obtaining sense distinction information for some methods that do not require a sense inventory)

from unstructured lexical resources like corpora (Dash, 2012; Edmonds, 2006). Those who are in favor of using corpora instead of structured lexical resources hold the view that “polysemy is an intrinsic quality of words, [and] ambiguity is an attribute of text”, so “context works to remove ambiguity” (Edmonds, 2006, p. 8). They therefore suggest that ambiguity would be resolved by considering the evidence derived from the context of a word’s use, i.e., the corpus evidence. Corpora bring authentic contexts to WSD research through texts of actual language that are more reliable (and possibly more up to date, as they are constantly renewed) than dictionary data, which mainly consist of an “exhaustive list of citations of sense variations of words” resulting from intuitive assumptions (Dash, 2012, p. 2). Corpora are also richer in linguistic knowledge than dictionaries and other structured resources because they show patterns of usage for a given word (e.g., its syntactic structure and collocational behaviour) that are considered useful information for signaling its use in different senses. For these reasons, corpora appear to be more relevant to WSD tasks than structured lexical resources. Word sense distinction information derived from corpora may be more enriched, and corpus-based sense inventories may be more appropriate to a WSD-dependent level of sense granularity, than those derived from structured lexical resources. The advantage of the corpus-based WSD is, however, also a disadvantage, reflected in the challenges that corpus-based methods such as supervised, unsupervised and semi-supervised methods have confronted. Despite containing rich linguistic knowledge information, corpora are still under-exploited due to the problem of the knowledge acquisition bottleneck in the supervised method. Induced senses from the unsupervised method may create a suitable level of sense granularity for WSD, but the word sense induction entails the challenge (raised in 2.5.1.2) relating to clustering corpus-derived citations that contain overlapping usages of a word into discrete groups, thus limiting its performance. Although the semi-supervised method overcomes the challenges of the

supervised and unsupervised methods, such semi-supervised methods as one-sense-per-collocation have not performed polysemous disambiguation successfully.

The limit posed by sense inventories derived from both structured and unstructured lexical resources on WSD performance has, according to Kwong (2013, p. 33), “led to a rethinking of what level of sense granularity is optimal for WSD and more importantly, what is really needed”. Several WSD researchers have long argued that “the standard fine-grained division of senses by a lexicographer for use by a human reader may not be an appropriate goal for the computational WSD task” (Agirre & Edmonds, 2006, p. 20) and agreed on more coarsely grained senses (Dagan & Itai, 1994; Ide & Wilks, 2006). Given the state of play, Dagan and Itai (1994) and Ide and Wilks (2006) suggest that the realistic level of sense granularity that WSD systems require corresponds roughly to that achieved by homograph distinctions. Homograph distinctions do not always necessarily require lexicographers to locate them. Homographs with different parts of speech, e.g., *play* as a noun and verb, can be easily disambiguated with reliance on current, reliable part-of-speech taggers. Homographs belonging to the same word classes, e.g., *stool* as “solid waste from the body” and “a wooden seat”, can be effortlessly found in parallel texts in different languages (parallel corpora). Moreover, considering results obtained by different WSD methods, the performance in resolving homographs is much better than in resolving polysemes, and therefore Ide and Wilks (2006) propose to redirect WSD to what it can perform with a high level of accuracy.

A root of the WSD problems causing some researchers to advocate the redirection toward coarse-grained distinctions lies in the difference between the characterization of WSD and the nature of word meanings. WSD is considered a task of classification in which words are assumed to have a fixed set of discrete senses (Cohn, 2003). Word meanings are, by nature, contextually varied and overlapping. Consequently, WSD attempts to classify a particular

occurrence of a word into an appropriate cluster (e.g., the unsupervised method) or sense in a sense inventory (e.g., the dictionary-based method) raise two questions:

Will senses ever be the same so that a sense previously seen can be used to name a later one? Are sense boundaries definite enough so that we can say a new occurrence of a word falls under one sense but not the other? (Kwong, 2013, p. 9)

These questions uncover a connection between WSD and lexical semantics, as the main endeavour of lexical semantics, more precisely, cognitive lexical semantics, is to (a) identify meanings (stored in the mind) that account for all of their instances in various contexts and (b) deal with the fuzziness of word meanings. However, despite this connection, WSD has established a closer relationship with lexicography than with cognitive lexical semantics because WSD and lexicography share the same assumption that “word uses can be grouped into coherent semantic units” (Edmonds, 2006, p. 8). Since differences still exist between lexicography and cognitive lexical semantics, e.g., the fuzziness of word meanings in the mental lexicon is inadequately represented in dictionaries, the relationship between WSD and lexicography pushes WSD further away from cognitive lexical semantics. This may explain why dictionary-based sense inventories aggravate rather than ease WSD problems.

Kwong (2013) therefore supports WSD to reconnect with cognitive lexical semantics. A union of these two paradigms, in which WSD strategies used by machines and humans are examined in parallel, may open an opportunity to remove the root of the WSD problems. This may also create a synergy between WSD and cognitive lexical semantics, in which both are closely supported and mutually advanced. Nevertheless, even though the potential of cognitive lexical semantics in resolving WSD problems is acknowledged, the union of these two paradigms has yet to come. This is because from the perspective of some WSD researchers, cognitive lexical semantics “has always been more concerned with representational issues [...] and models of word meaning and polysemy so far too complex for WSD” (Agirre & Edmonds, 2006, p. 2). This may lead to hesitation or even resistance from the WSD side in bringing

together the two paradigms. Agirre and Edmonds (2006, p. 2) note that “WSD has never really found a home in lexical semantics”. Also, WSD researchers are happy with what they are doing and achieving in WSD, so “notwithstanding the theoretical concerns to the logical or psychological reality of word senses, the field of WSD has successfully established itself by largely ignoring lexical semantics” (Edmonds, 2006, p. 8), “much as lexicographers do in order to produce dictionaries” (Agirre & Edmonds, 2006, p. 9).

2.9 WSD and corpus-derived wordlists

Wordlists are much-used resources in many disciplines. They are often known as unigram lists, “a compact representation of a corpus, lacking much of the information (being decontextualised), but small and easily tractable”, which are essential for applications in natural language processing like machine translation (Kilgarriff et al., 2014, p. 123). In psychology, psychologists investigating language acquisition and understanding are interested in word frequency because it is correlated with the speed with which a word is acquired and understood. Educationalists are also interested in word frequency in English language education because it can guide curricula for learning and teaching English vocabulary in particular contexts. Psychologists and educationalists thus tend to work towards wordlists that contain words frequently found in target contexts. Wordlists in these two disciplines are viewed as frequency wordlists.

Frequency wordlists are variously designed for general English (e.g., West’s (1953) General Service List), academic English (e.g., Coxhead’s (2000) Academic Word List) and English for specific purposes (e.g., Hsu’s (2013) Medical Word List). A standard method to develop a frequency wordlist is from a corpus. Words make their way to a corpus-derived wordlist only if their frequency of occurrence in a target corpus passes a pre-determined threshold. Word frequency is usually calculated using automatic computer programs (e.g., the RANGE program). In computer-automated calculation, lemmas and word families are two

word constructs often used for word frequency counts. The frequency of a word calculated using lemmas and word families is the aggregate frequency of its inflections and derivations. For example, the calculation of the lemma *work* or the word family *diagnosis* targets inflected forms of *work* (*works*, *working* and *worked*) or inflected and derived forms of *diagnosis* (*diagnosable*, *diagnose*, *diagnoses*, *diagnosing* and *diagnosed*). Base word forms of lemmas (e.g., *work*) or word families (e.g., *diagnosis*), often known as headwords, are presented in wordlists with their frequency statistics.

An underlying principle on which the constructs of a word (and the making of corpus-derived wordlists) are based is that there is “a core or basic meaning that inherently exists in all of the derivations and inflections of a certain root word or base form” (Anderson & Nagy, 1991; Bauer & Nation, 1993; Graham, 2008, p. 23; Nerlich et al., 2003; Sinclair, 2004; Stubbs, 2002). As corpus-derived wordlists are created for deliberate decontextualized learning, i.e., base word forms are learned and taught out of context, the principle offers a “logical approach” to acquiring and understanding base word forms in wordlists (Bauer & Nation, 1993). Accordingly, wordlist users, who are often English language learners, are expected to acquire a high-frequency word form, e.g., *work*, based on what they deem as the core meaning of the word. Then, they can use contextual clues to elaborate the word’s core meaning to understand a particular use of the word in a specific context, e.g., “The old woman slowly worked her way across the street” (Graham, 2008, p. 24).

Since corpus-derived wordlists rely on word frequency and words have the nature of being multi-meaning, several studies have been conducted to examine the relationship between word frequency and the number of word meanings (Graham, 2008; Ravin & Leacock, 2000; Skoufaki & Petric, 2021). Results indicate that the relationship is positive: “low frequency words tend to have only one [meaning] but once a frequency threshold is passed, the number of word [meanings] increases as word frequency increases” (Skoufaki & Petric, 2021, p. 9).

Because of appearing in many contexts, high-frequency words tend to have more meanings than low-frequency words, thus having the potential for more homography and polysemy. Graham (2008, p. vi) and Ravin and Leacock (2000, p. 1) confirm that “the presence of homography tends to be extensive in many high-frequency word forms” and “the most commonly used words tend to be the most polysemous”.

As homography and polysemy are pervasive among high-frequency words, they have become the main concerns for making corpus-derived wordlists, i.e., calculating word frequency. These two phenomena have complicated “the process of defining the construct of word and consequently how words are counted and what words are included in word lists” (Gardner, 2007; Graham, 2008, p. 19; Hyland & Tse, 2007; Knowles & Mohd Don, 2004). Lemmas and word families are, albeit “two primary ways in which words have been defined, grouped, and counted in wordlists of the last 20 to 30 years”, not ideal for defining the construct of word that accounts for homography and polysemy (Gardner, 2007; Graham, 2008, p. 27). This is because these two constructs of word focus on word forms rather than word meanings, posing a significant obstacle to the implementation of WSD. The computer-automated process of counting word frequency using lemmas demonstrates limited capacity for disambiguating polysemous and homographic senses. This can be seen in Gardner’s (2007) analysis of the word forms *bear* and *bore*. These two forms, which are sometimes in a relation of polysemy (e.g., bear/bore trays of drinks, bear/bore a burden) and sometimes in a relation of homography (e.g., to bore a hole, a black bear), should be separately counted where possible. Nevertheless, the computer-automated frequency calculation of lemmas cannot disambiguate polysemous and homographic forms, consequently linking the forms *bear* and *bore* together. The use of word families in the computer-automated frequency calculation is even worse because “they include so many forms under the guise of one meaning, consequently bringing out all of the problems listed with lemmas, but to an even more exaggerated level” (Graham, 2008, p. 30).

Hyland and Tse (2007) warn that, especially for word forms not used in the same way in different disciplines, lumping the variety of their uses together under a word family may misrepresent them.

The constructs of word disconnect themselves with WSD, causing corpus-derived wordlists to have shortcomings (Kilgarriff et al., 2014; Nation et al., 2016). According to Nation and Parent (2016), ideally, in corpus-derived wordlists, homographs should be presented as separate lemmas or word families, as they are two different words. Polysemes should not be presented separately, as they are meanings of the same word, unless the wordlist is aimed at low-proficiency learners. Nevertheless, in reality, neither homographic forms nor polysemous senses (for low-proficiency learners) are separately presented, because the computer-automated frequency calculation of both lemmas and word families cannot reliably distinguish homography and polysemy, as in the case of *bear* and *bore*. This leads to a lack of supplementary information about word meanings, which may exacerbate the problem of polysemy and homography presentation in wordlists. Polysemous senses are usually represented under a base word form whose meaning is absent. If the word is presumably unknown to learners, its base word form alone may give no clue for them to figure out its (core) meaning nor to understand its usage in a particular context. Failure to provide word meanings may complicate the learning of words with homographs because homographs, especially ones found in different disciplines, which are not usually presented separately in wordlists, can cause misunderstanding for learners.

These shortcomings reveal that corpus-derived, frequency-based wordlists do not provide core meanings to facilitate the understanding of polysemous and homographic senses grouped under the same base word form. This may undermine the logical approach on which the deliberate decontextualized learning of word forms is based, making the quality of wordlists questionable (Thompson & Alzeer, 2019). There has thus been a widespread tendency to

evaluate wordlists, mainly in general and academic English, with a focus on their shortcomings, to enhance their utility. However, the evaluation of wordlists in English for specific purposes, such as Hsu's (2013) Medical Word List (MWL), is still rare.

Several studies have been conducted to examine homography in wordlists used for learning general and academic English. Among these, Parent's (2012) and Wang and Nation's (2004) studies are notable because they extensively evaluate two well-known wordlists—West's (1953) General Service List (GSL) and Coxhead's (2000) Academic Word List (AWL). The studies aim to semantically analyze individual word forms in the two lists to identify which word forms have homographs. Both found that the number of words with homographs in the GSL and AWL is relatively modest. Nevertheless, it does not mean homography is ignored when creating wordlists. According to Parent (2012, p. 79), “although homography [occurs] with a reasonable number of word forms, typically the different meanings occur with very different frequencies”. This is exemplified through Wang and Nation's (2004) identification of three AWL word families whose homographs did not satisfy a frequency threshold for inclusion in the list. Hence, it is still worth separating homographs when doing word frequency counts, as “the more accurate a count, the more valid it is” (Parent, 2012, p. 78).

Polysemy in wordlists is also examined, but not as much as homography, probably because of Nation and Parent's (2016) suggestion of only distinguishing polysemes for low-proficiency learners. Skoufaki and Petric's (2021) evaluation of Gardner and Davies's (2014) Academic Vocabulary List (AVL) is one of the few studies that takes polysemy into account. Skoufaki and Petric (2021) consulted two resources—*Collins COBUILD Advanced Learners' Dictionary* and WordNet—to identify polysemous lemmas in the AVL. The findings revealed that over half the AVL lemmas had more than one definition in the two resources and were thus deemed polysemous. Since the AVL consists of words frequently found in academic discourse across disciplines, Skoufaki and Petric (2021) also highlighted the need to separately

present their polysemous senses. They further reasoned that learners tend to resist inferring word meanings from context and insist on applying a meaning they already know to all contexts. So, if the known meaning is a general one that does not work in discipline-specific contexts, the separate presentation of general and discipline-specific meanings is necessary for assisting learners in learning different meanings accurately. (Bensoussan & Laufer, 1984; Frantzen, 2003; Skoufaki & Petric, 2021)

The studies above suggest that the utility of wordlists could be enhanced if word meanings rather than word forms are counted and presented. However, this suggestion appears to be impractical because of the constraint on WSD raised by homography and polysemy. Kilgarriff et al. (2014, p. 131) state that counting word senses, especially polysemous senses, is almost impossible because “50 years of research in automatic WSD has not delivered programs which can automatically say, with a reasonable level of accuracy, which sense a word is being used in”. Henceforth, studies on evaluating wordlists that go beyond identifying word forms with polysemes and/or homographs to supplement frequencies of word meanings, especially polysemous ones, may suffer from an insurmountable burden. This is because an enormous amount of laborious sense coding would be required. Moreover, corpora widely used for creating wordlists, such as the Corpus of Contemporary American English (COCA) and British National Corpus (BNC), lack semantic tagging.

CHAPTER 3: METHODOLOGY

3.1 Theoretical framework

The study aimed to develop a lexicographic resource of semi-technical medical vocabulary that would comprehensively address unresolved issues of polysemy and homography in corpus-derived, word form frequency-based wordlists and conventional dictionaries. The procedure consisted of three main stages in which theories from lexical semantics, lexicography and corpus-based WSD were applied.

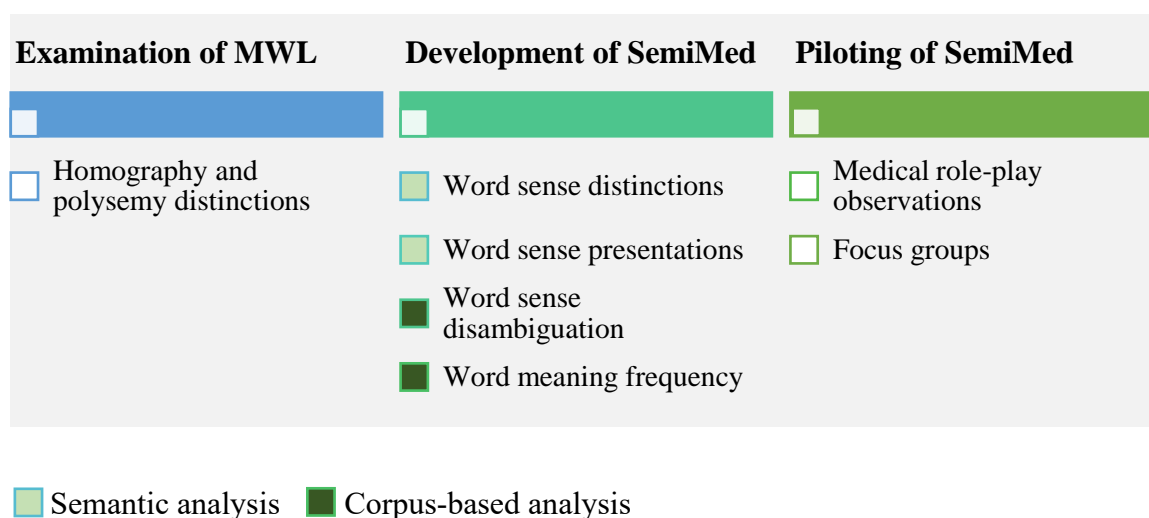


Figure 3.1 The three-stage research procedure

First, the study started with a corpus-derived list of semi-technical medical vocabulary—Hsu’s (2013) MWL, whose reliance on word forms regardless of word meanings means that its treatment of polysemy and homography may be lacking. Lexical semantic theories regarding the distinction between polysemy and homography were considered when conducting the examination of the MWL. Secondly, SemiMed was developed based on findings from the examination of the MWL. The development of SemiMed involved semantic and corpus-based analyses. The semantic analysis worked towards two key aspects of lexicographic practice—word sense distinctions and presentations. Relevant lexical semantic theories that accommodate the fuzziness of word meanings were directly applied to set out a fundamental theoretical background to satisfactorily perform the lexicographic tasks of

distinguishing and presenting polysemous meanings of a word and its homograph's meaning(s). The corpus-based analysis focused on overcoming word form-related drawbacks of corpus-derived wordlists by considering word meanings. The task of WSD was introduced in the context of corpus linguistics to facilitate the calculation of word meaning frequency—a task that has rarely been undertaken during the creation of corpus-derived wordlists. Thirdly, SemiMed was piloted in medicine-oriented role-plays with EFL medical student participants. Observations and focus groups were carried out to examine the practicality of SemiMed, especially functions that were the result of incorporating theories in lexical semantics and corpus-based WSD into lexicographic practices.

3.1.1 Homography and polysemy distinctions

After considering the limitations of etymology- and native speaker's judgement-based approaches, the method that the study adopted for drawing distinctions between homography and polysemy was to combine two methods predominant in the field of lexical semantics—word etymology and native speaker's judgement on (un)relatedness of word meanings. The combination of these two methods, not previously found in the lexical semantic literature, aimed to complement each method fully. As mentioned in 2.2.3, although the etymology-based approach offers objective evidence that is usually accessible in virtually all historical dictionaries, etymological evidence is sometimes untraceable, making this method alone unable to deal with words whose historical derivation took place too far in the past to be precisely captured. The native speaker's judgement-based approach, which rarely relies on etymological evidence, potentially offers an alternative to cases that the etymology-based approach has limited capacity to cope with.

Despite the potential to supplement the etymology-based approach, the native speaker's judgement-based approach only produces consistent results for homographs (Lehrer, 1974). Regarding polysemous words, subjectivity usually reaches an undesirable level, i.e., subjective

judgement results scatter along the continuum of polysemy and homography. To minimize an undesirable level of subjectivity, the study referred to core meaning theories to set a benchmark for subjective judgement-based decisions on meanings of a word having semantic similarity (e.g., meanings “fish” and “sandal” for *sole*). Core meaning theories unite the monosemic and polysemic approaches to polysemy in lexical semantics. In essence, the two approaches agree on the concept of core meaning, which has also been acknowledged in fields other than lexical semantics (e.g., lexicography and corpus linguistics) (Bauer & Nation, 1993; Béjoint, 1990; Carston, 2021; Hyland & Tse, 2007; Mihalcea, 2006; Rundell, 2012). A core meaning, variously named core/central/prototypical meaning in monosemic and polysemic approaches, is perceived to be a psychologically plausible agent from which polysemous meanings of a word extend over time. It is considered a feature of polysemy only and thus valuable in differentiating polysemy from homography, especially when it is hard to determine the etymological connection between words which have semantic similarity.

3.1.2 Word sense distinctions

The task of distinguishing word senses, specifically making fine-grained distinctions of polysemous word meanings, poses lexicographic challenges. Although lexical semantics, particularly cognitive lexical semantics, has been acknowledged as beneficial in resolving challenges related to word sense distinctions, theories in cognitive lexical semantics have not been directly applied in lexicographic practices. This study therefore adopted Principled Polysemy, a cognitive lexical semantic approach to polysemy that fully accounts for the fuzzy nature of word meanings, to perform word sense distinctions. Principled Polysemy offers explicit criteria that parameterize the process of identifying distinct senses. Elements that the criteria include, such as collocational and structural patterns, are measurable through corpus-based analyses, allowing subjective, criteria-based judgement to be cross-checked with objective, corpus-derived evidence, minimizing subjectivity to an acceptable level.

Moreover, Principled Polysemy was among the first theories to propose decision criteria for determining a core meaning. Although there is a growing consensus among semanticists about the existence of core meaning, semantic analysts sometimes have diverging opinions on which meaning should be considered core, due to their sole reliance on intuition. The use of Principled Polysemy in previous related studies is testimony to the capability of this approach to validate intuitive decisions to arrive at a mutually agreed core meaning (Dalpanagioti, 2018; Evans, 2005; Mahpeykar & Tyler, 2015; Tyler & Evans, 2001; Van der Gucht et al., 2007). Although Principled Polysemy has been so far applied to identify core meanings of prepositions, (abstract) nouns and verbs, Evans (2005) suggests Principled Polysemy offers a duplicable and transferrable methodology to lexical studies on different parts of speech. Following this suggestion, my study therefore extends the application of the Principled Polysemy approach to the determination of core meaning of not only nouns and verbs, but also adjectives and adverbs, because these are common word classes found in semi-technical medical vocabulary.

3.1.3 Word sense presentations

The meaning-based presentation of word senses, i.e., semantic order in respect of core meaning, has been perceived as one of the few satisfying methods that takes the semantic aspects of words into consideration. However, despite being semantically plausible, semantic order lacks objectivity, causing this method to be less frequently used than other methods such as grammar-based organization. Primary concerns about semantic order stem from the fact that lexicographers must rely on their intuition to identify a core meaning; this can be alleviated by adopting Principled Polysemy (e.g., criteria to determine a core meaning). This meaning-based presentation of word senses was therefore employed in this study. Rather than vertically listing core and related meanings, Lakoff's (1987) radial format was chosen to underpin the meaning-based presentation, as it fully reflects the multidimensional structure of word meanings.

Moreover, as Lakoff's radial categories align with the Principled Polysemy approach, Lakoff's radial format-inspired presentation is the best fit for presenting distinct senses resulting from Principled Polysemy.

Cantos and Sanchez's (2001) Lexical Constellation (LC) model was employed for the realization of Lakoff's (1987) radial categories. Although Evans (2005) illustrates the radial category of *time* in the radiating-lattice structure (Figure 2.1), this model appears not to be lexicographically appropriate because it does not provide optimal space for full explanations of word meanings. Cantos and Sanchez's (2001) model, on the other hand, allows texts to be inserted into bubble-shaped meaning clusters, creating more detailed views of word meanings. In addition, this model is successfully duplicated in Perez's (2013) study on semi-technical words and showcases "semantic hierarchies existing amongst the general and specialised semantic features of these terms and their dependencies in a very clear and visual manner" (p. 165). The default format of LCs in which word meanings radiate circularly (not downward as in the radiating-lattice structure) from the core meaning seems to articulate the idea of "radial" better. This format is also more likely to clear up the first (figurative) meaning-related confusion in the conventional linear format, as meanings are not numbered, allowing users to fully concentrate on relations between literal and figurative meanings vis-à-vis core meaning.

3.1.4 Word sense disambiguation and word meaning frequency

The lexical resource that the study used to disambiguate word senses was mainly corpus-based rather than purely dictionary-based. Since dictionary-based WSD faces limited information about different word senses in dictionaries, WSD that relies on corpora, which store richer data, was expected to provide broader lexical knowledge. The primary linguistic knowledge that the corpus-based WSD in this study referred to was collocations. The one-sense-per-collocation heuristic (Yarowsky, 1993) shed light on the entire WSD process. This process was semi-automatic in the sense that collocates were first automatically exported using

corpus analysis software and then manually assigned to appropriate meanings by human analysts. The semi-automatic method of corpus-based WSD with reference to collocational data is practical and feasible because collocational data can be reliably retrieved from corpus analysis software. This method is also open to a wide range of corpora because it does not necessarily require sophisticatedly sense-tagged corpora.

Because it characterizes a type of vocabulary that usually carries multiple meanings across different contexts, SemiMed places more focus on word meaning frequency than word form frequency. Within the scope of this study, WSD was viewed as an intermediate task that contributed to the calculation of word meaning frequency, which has often been overlooked in the development of corpus-derived lists of the most frequently occurring semi-technical words. The corpus-based WSD with reference to collocational data offered a means of determining word meaning frequency, i.e., collocations. While calculating the frequency of word meaning itself is impossible due to the scarcity of sense-tagged corpora, calculating the frequency of word meaning via the frequency of collocates is more attainable. This is because, as mentioned above, collocates are automatically exported, and their frequencies can be reliably calculated by corpus analysis software.

3.2 The examination of Hsu's (2013) Medical Word List (MWL)

3.2.1 A brief description of Hsu's MWL

The MWL enumerates 595 words situated between non-technical and technical vocabulary, ranging from the BNC (British National Corpus) 4th 1,000 to 14th 1,000 words (and beyond). This is a corpus-derived, word form frequency-based list, i.e., containing the most frequently occurring word forms in a corpus that satisfy a pre-determined set of criteria. The MWL was created from a custom-made corpus, using the RANGE program as an analysis tool and adopting three selection criteria. The corpus was compiled from 155 online medical

textbooks, containing 15 million words, covering 31 subject areas in medicine. This corpus was named the Medical Textbook Corpus.

Table 3.1 Thirty-one medical sub-disciplines covered by Hsu's MWL (2013)

1	Anaesthesiology	17	Neurology
2	Allergology/Immunology	18	Nephrology
3	Alternative/Complementary Medicine	19	Obstetrics/Gynaecology
4	Cardiology	20	Oncology
5	Dermatology	21	Ophthalmology
6	Dentistry	22	Orthopaedics/Rehabilitation
7	Endocrinology/Metabolism	23	Otorhinolaryngology
8	Emergency Medicine	24	Perinatology/Paediatrics
9	Forensic Medicine	25	Psychiatry
10	Gastroenterology	26	Pathology
11	Hematology	27	Pulmonary/Respiratory Medicine
12	Hepatology	28	Public Health
13	Health Informatics	29	Radiology
14	Urology	30	Surgery
15	Infectious Diseases	31	Transplantation
16	Intensive Care Medicine		

Lexical items that became final candidates for the list had to satisfy all three criteria: (1) specialized occurrence, (2) range and (3) frequency (more details in 4.2.3 and Table 4.1). The RANGE program (Nation & Heatley, 2005) was used to calculate the range and frequency. The unit of counting was the word family. The RANGE program automatically read all inflections and derivatives of a base (word) form and counted their range and frequency as one word family. For example, the range and frequency of *diagnosis* were the sum total of its inflected and derived forms (*diagnosable, diagnose, diagnoses, diagnosing* and *diagnosed*).

3.2.2 A source of semantic input for the examination of Hsu's MWL

The *Oxford English Dictionary* (OED) was selected as the source of semantic input for the examination of the MWL.

Purpose: Since the MWL only presents base forms (headwords) of 595 word families, their range and frequency statistics, the OED was used to find the meanings of the MWL's headwords.

Description: The study used an online version of the OED (full text available at <https://www.oed.com/>). The OED online database offers a guide to the meaning, history and pronunciation of 500,000 words, together with 3.5 million annotations from a wide range of texts and genres.

Rationale: There were two reasons for selecting the OED to look up the MWL's headwords. First, since the MWL contains semi-technical medical words—a type of vocabulary with general and medical meanings—it was necessary to choose a dictionary that presents both meanings in great detail. The OED was a suitable option because, unlike medical dictionaries (which usually focus on presenting only medical meanings), it can be used as a general dictionary that presents word meanings in general contexts. The OED is more suitable than other general dictionaries because it is a historical dictionary comprising a wide range of meanings in different disciplines, including medical ones. This means it does not miss out medical meanings, which are sometimes absent in general dictionaries. Secondly, as a historical dictionary, the OED includes etymological information rarely seen in learners' (general and medical) dictionaries and offers a complete account of each word's history, i.e., past and present word meanings. These two features of the OED are beneficial in providing rich data for a core meaning-based analysis.

3.2.3 A method for the examination of Hsu's MWL

The examination of Hsu's MWL was undertaken to re-evaluate the list with a focus on the semantic aspects of its 595 headwords. The examination began by looking up the 595 headwords in the OED and then identifying polysemes and/or homographs among the OED definitions.

A core meaning-based method was proposed to analyze polysemy and homography. This method combined both etymology and subjective judgement to distinguish polysemy and homography vis-à-vis the core meaning. More specifically, the OED definitions of the

examined words were evaluated against a set of criteria, involving etymological and subjective judgement, to identify whether there was a shared core meaning. If there was, they were deemed polysemes. Otherwise, they were classified as homographs.

The set of criteria the study used was a modified version of Evans's (2005) criteria to determine a word's central meaning. Since the criteria were originally applied to analyze the abstract noun *time*, the fourth criterion of lived temporal experience seemed to be irrelevant when analyzing other word classes and was therefore excluded from this analysis. Additionally, as a core meaning is not always the earliest attested meaning (Atkins & Rundell, 2008; Evans, 2005), it is not mandatory for a meaning to meet the first criterion to be recognized as a core meaning. However, since Evans (2005) states there are some overlaps between the earliest attested meaning and the core meaning, it was critical to know the earliest attested meaning of an examined word so that the word's (provisional) core meaning could be determined. Therefore, a core meaning in the current study needed to satisfy at least one of the following three criteria: (1) be the earliest attested meaning, (2) be predominant in the semantic network and/or (3) be predictable in regard to other meanings.

Two native and one non-native English speaking evaluators (details in 4.3.3 and Table 4.4) were invited to analyze polysemy and homography of the MWL's headwords with reference to the modified set of criteria. The first criterion was associated with etymology, i.e., a diachronic analysis of words. The evaluators consulted the OED, which played a role as a source of etymological information to gain evidence for identifying the earliest attested meaning. The remaining criteria required their subjective judgement. More particularly, they then evaluated the earliest attested meaning to see whether it was a core meaning of an examined word by judging whether it was predominant and/or predictable among the word's OED definitions. The core meaning was confirmed when either or both the second and third criteria were satisfied. Polysemes and/or homographs were identified by judging whether the

OED definitions of the word linked to the confirmed core meaning in terms of the second and third criteria. If they did, they were polysemous meanings of the word. If not, they were meanings of the word's homographs.

3.3 The development of SemiMed

SemiMed was developed as a lexicographic resource that exclusively deals with semi-technical medical vocabulary and has specific properties, referring to Atkins and Rundell's (2008, p. 24-25) properties to classify dictionaries, as follows:

- Language: SemiMed is a monolingual resource.
- Coverage: It aims to cover both general and medical meanings of words.
- Medium: SemiMed is ideally compatible with an e-format. However, it can also be adapted to a printed format.
- Organization: Users are expected to use SemiMed to find meanings of a word. Hence, its organization is what Atkins and Rundell (2008, p. 25) term "word to meaning".
- Intended users: SemiMed targets EFL/ESL students whose majors are medicine-related.
- Functions: Users will use SemiMed to understand meanings (relations between meanings, e.g., polysemy and homography) of a word and to interpret a word appropriately in particular (general and medical) contexts. It is potentially a resource for EMP teaching.

The development of SemiMed was based on a sample of Hsu's (2013) MWL, i.e., core meanings, polysemes and/or homographs of 40 MWL words (see sampling procedure in section 6.3.2.1, Paper 3 in this thesis). It comprised semantic and corpus-based analyses, which respectively aimed to generate the microstructure and calculate the frequency of meanings of MWL's words. The semantic analysis determined other related meanings of a word in addition

to its core meaning identified in the examination of Hsu's MWL, modified their OED definitions, and presented relationships (e.g., polysemy and homography) between these meanings in a semantic network. The corpus-based analysis checked whether the other related meanings identified in the semantic analysis were distinct meanings and examined their frequency of occurrence in medical and general contexts.

3.3.1 Semantic analysis

3.3.1.1 Determine other related meanings

Other meanings of a word relating to its confirmed core meaning were identified among the word's OED definitions, referring to Evans's (2005) three criteria to determine distinct meanings. A distinct meaning was expected to (1) feature an additional meaning (meaning criterion), (2) have collocational dependencies (concept elaboration criterion) and/or (3) have unique structural patterns (grammatical criterion).

The identification of other related meanings served two purposes: (a) preparing a sense inventory as a source of reference for disambiguating the word in the corpus-based analysis and (b) suggesting an appropriate degree of sense granularity for the final version of SemiMed. Evans's (2005) criteria played a critical role in re-adjusting the sense granularity in the OED, which usually appeared too fine-grained to be realistically distinguished through corpus-based tasks of WSD. The re-adjusted granularity, after being double-checked in the corpus-based analysis, resulted in parameterized distinct meanings in SemiMed.

The procedure started with the removal of obsolete senses. For instance, meanings 4 and 5 of *defect* (n); meanings 1, 2, and 4 of *defect* (v); and meanings 1 and 2 of *defect* (adj) were indicated in the OED as being rarely used in present-day language (see Figures 3.2, 3.3 and 3.4), so they were taken out before Evans's (2005) criteria were applied. In the semantic analysis, the criteria were used to parameterize mostly sense lumping, i.e., lumping OED definitions that did not appear to be distinct senses, e.g., merging meanings 2 and 3 of *defect*

(n). The last two criteria were rechecked in the follow-up corpus-based analysis, where further sense splitting might occur if necessary based on corpus-based evidence.

Defect, n.

1. Lack or absence of something necessary or desirable; a deficiency, a want. Also: the state or fact of being deficient or falling short.
2. An imperfection in a person or thing; a shortcoming, a failing; a fault, flaw, or abnormality.
3. The quality, state, or fact of being imperfect; defectiveness, faultiness.
4. (*Obsolete*) An act of abandoning or renouncing something; a defection.
5. (*Obsolete*) The failure of the moon, sun, or another celestial object to (fully) appear; an eclipse, an occultation.

Figure 3.2 OED definitions of *Defect* (n)

Defect, v.

1. (*Obsolete*) To hurt, to damage; to cause to have defects.
2. (*Obsolete*) To fail, to fall short; to become deficient or wanting.
3. To abandon or desert a person, party, organization, or cause, esp. in favour of an opposing one.
4. (*Obsolete*) To desert or abandon (something).

Figure 3.3 OED definitions of *Defect* (v)

Defect, adj.

1. (*Obsolete*) Disfigured
2. (*Obsolete*) Defective, deficient; wanting.

Figure 3.4 OED definitions of *Defect* (adj)

3.3.1.2 Modify OED definitions of other related meanings:

After being lumped (and split), OED definitions of other related meanings appeared to be lengthy, making them harder for users with lower levels of English (e.g., beginner, pre-intermediate, intermediate, or even upper-intermediate) to understand. Therefore, they were simplified as far as possible. The modification of OED definitions was underpinned by two principles (Atkins & Rundell, 2008, p. 433; Mel'čuk, 1988, p. 171-175):

- The decomposition principle: Defining language must be semantically simpler than the content it defines.
- The univocity principle: Each defining term used in defining a word must not be ambiguous.

The two principles were also applied to write definitions of other related meanings resulting from the splitting in the corpus-based analysis.

3.3.1.3 Present core and other related meanings in a semantic network

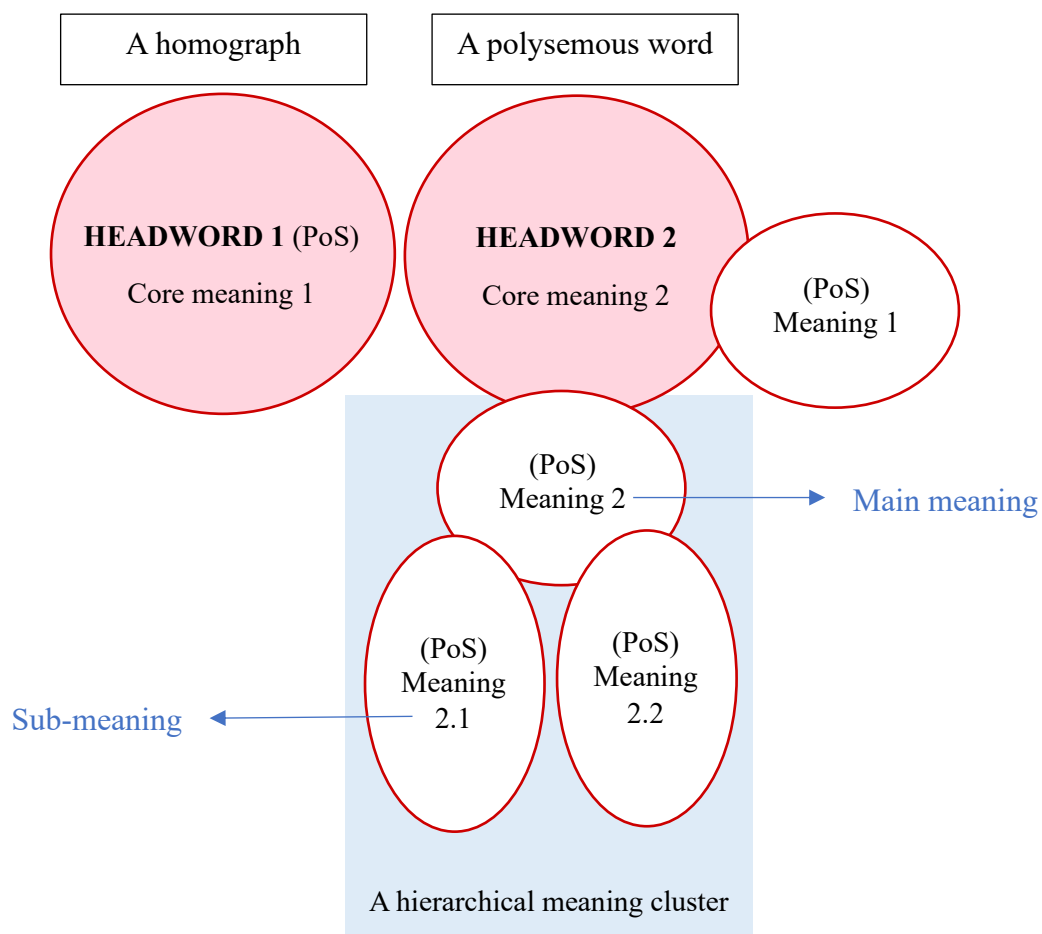


Figure 3.5 A generic microstructure in SemiMed

The microstructure of SemiMed is meaning-based and has a radial format. Other related meanings (polysemous meanings) of a word are grouped together regardless of their parts of speech. Meaning-based groups (meaning clusters) are structured in a hierarchy of meanings. Specifically, meanings in a cluster were categorized into main meanings and sub-meanings. A polysemous word was then visualized in a semantic network (an LC) with its core meaning and a base form of its headword placed at the centre. Meaning clusters were placed around the core meaning, to which main meanings of individual clusters were attached. Within each cluster, sub-meanings were directly attached to their main meaning. If the word had homograph(s), its homograph(s) would be presented in a separate LC, next to the LC of the polysemous word.

3.3.2 Corpus-based analysis

The meanings of SemiMed's semantically analyzed words then underwent a corpus-based analysis to check their granularity and examine their frequency of occurrence in medical and general corpora.

3.3.2.1 Undertake word sense disambiguation

As SemiMed's headwords (their base word forms) have multiple meanings, the corpus-based analysis commenced with a WSD task. The method of WSD used in this study was one-sense-per-collocation. The analysis tool selected to undertake the WSD task was Sketch Engine, a corpus query system. It was used because

- it offers full access to a wide range of general and specialized corpora from which this study could select target contexts to examine SemiMed's words, and
- it features Word Sketch, a unique function that allows a one-page display of automatically corpus-generated summary of a word's collocational and grammatical behaviour; therefore, it is significantly advantageous for the disambiguation of word senses based on their collocations and the checking of Evans's (2005) concept elaboration and grammatical criteria.

Two corpora, English Web 2020 and Medical Web Corpus, were chosen to represent general and medical contexts respectively (see a detailed description of the corpora in Table 5.2 in 5.3.1). A word was disambiguated by examining its collocates in these two corpora using Word Sketch. As can be seen in Figure 3.6, Word Sketch exported adjacent collocates and categorized them into groups based on grammatical relationships with their node word, e.g., modifiers of *defect*, nouns modified by *defect*, verbs with *defect* as object, verbs with *defect* as subject, etc.

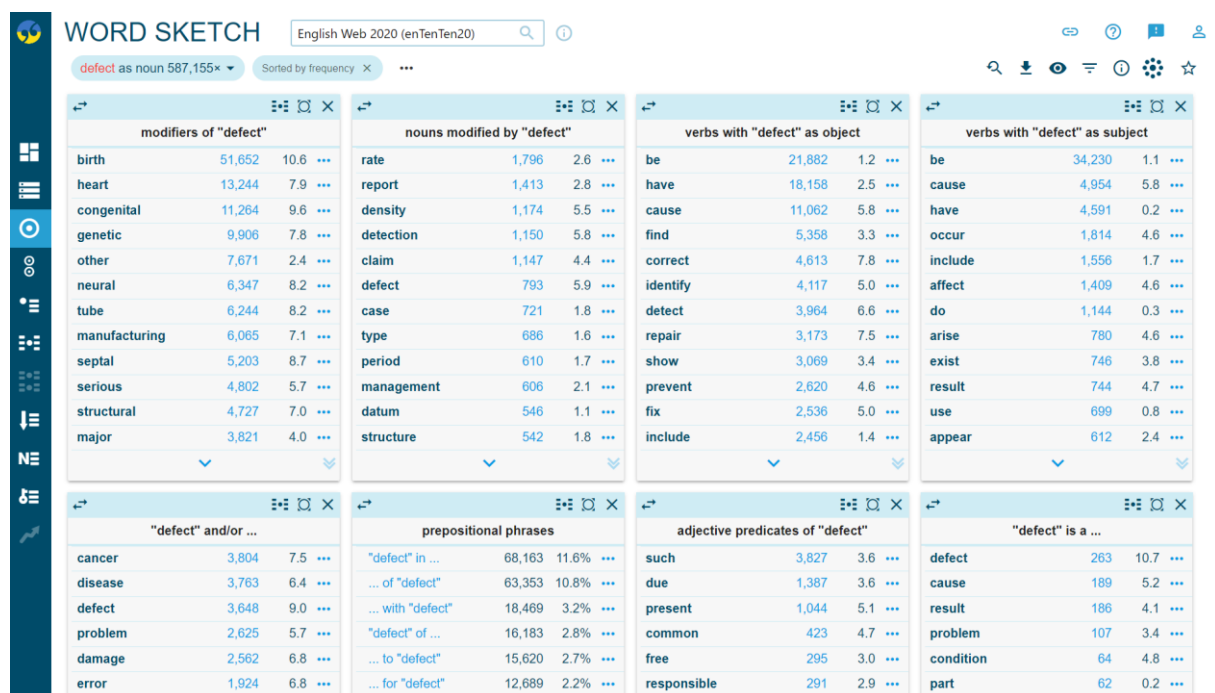


Figure 3.6 Sketch Engine-generated collocates for *Defect* (n) in English Web 2020

This default setting effectively facilitated the filter of certain types of collocates (Table 3.2), which, according to Yarowsky (1993, p. 269), could help provide more disambiguating information. In this regard, specific categories were targeted, depending on a word's parts of speech, to narrow down the number of examined collocates.

Table 3.2 Targeted categories of collocates in relation to their node's parts of speech

Parts of speech of the <i>node word</i>	Categories of the <i>node word's</i> collocates
Noun	Adjectives modifying the <i>node word</i>

	(e.g., spastic <i>colon</i>)
	Nouns modified by the <i>node word</i>
	(e.g., <i>colon</i> cancer)
	Verbs with the <i>node word</i> as an object
	(e.g., cleanse the <i>colon</i>)
Verb	Objects of the <i>node word</i>
	(e.g., <i>circulate</i> air)
Adjective	Nouns the <i>node word</i> modifies
	(e.g., <i>benign</i> lesions)
Adverb	Verbs modified by the <i>node word</i>
	(e.g., run <i>parallel</i>)

Collocates from targeted categories were further ranked according to their frequency scores (discussed in 3.3.2.3). Only the most frequent collocates were assigned meanings.

3.3.2.2 Check the granularity obtained by semantic analysis

The assignment of meanings to selective collocates was based on a sense inventory, resulting from the semantic analysis, and corpus-based evidence, to ensure that the granularity was cross-checked with contextual data.

The sense inventory initially provided a set of a word's possible meanings from which the most appropriate meaning would be assigned to a collocate. If none of the meanings listed in the sense inventory matched a particular collocate, a new meaning would be adduced from examples containing the collocate and its node word. The examples could be easily retrieved via the Concordance directly linked to all functions of Sketch Engine, including Word Sketch (Figures 3.7 and 3.8).

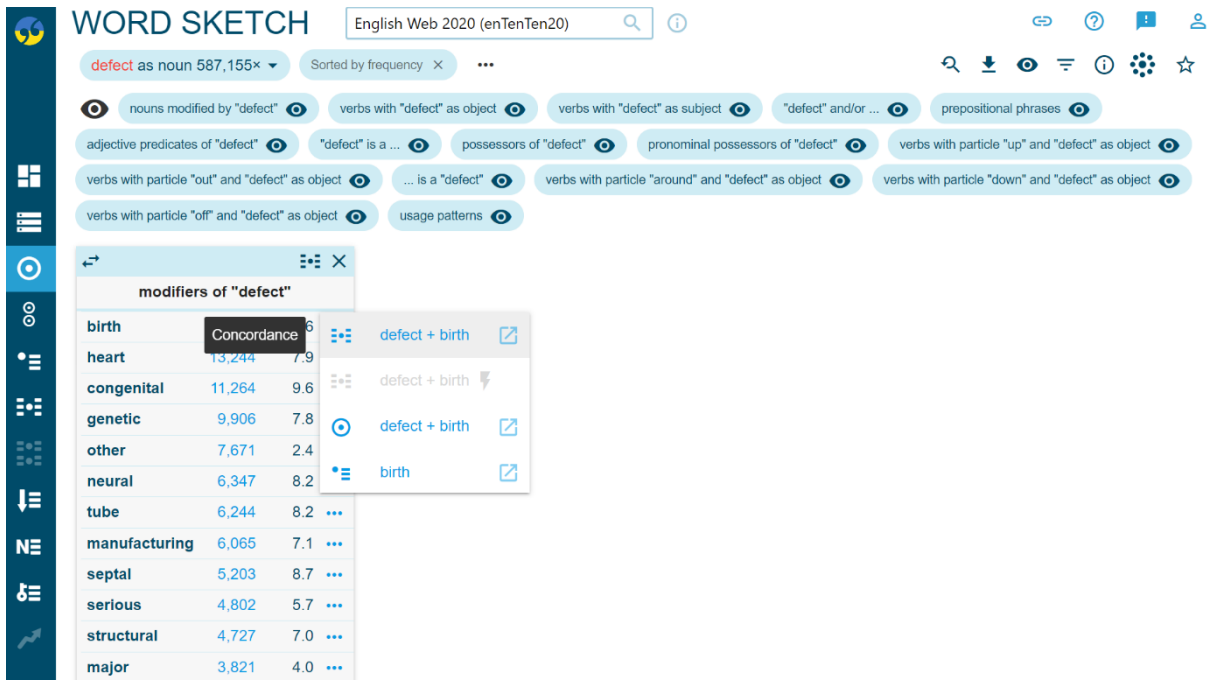


Figure 3.7 Concordance function in Word Sketch

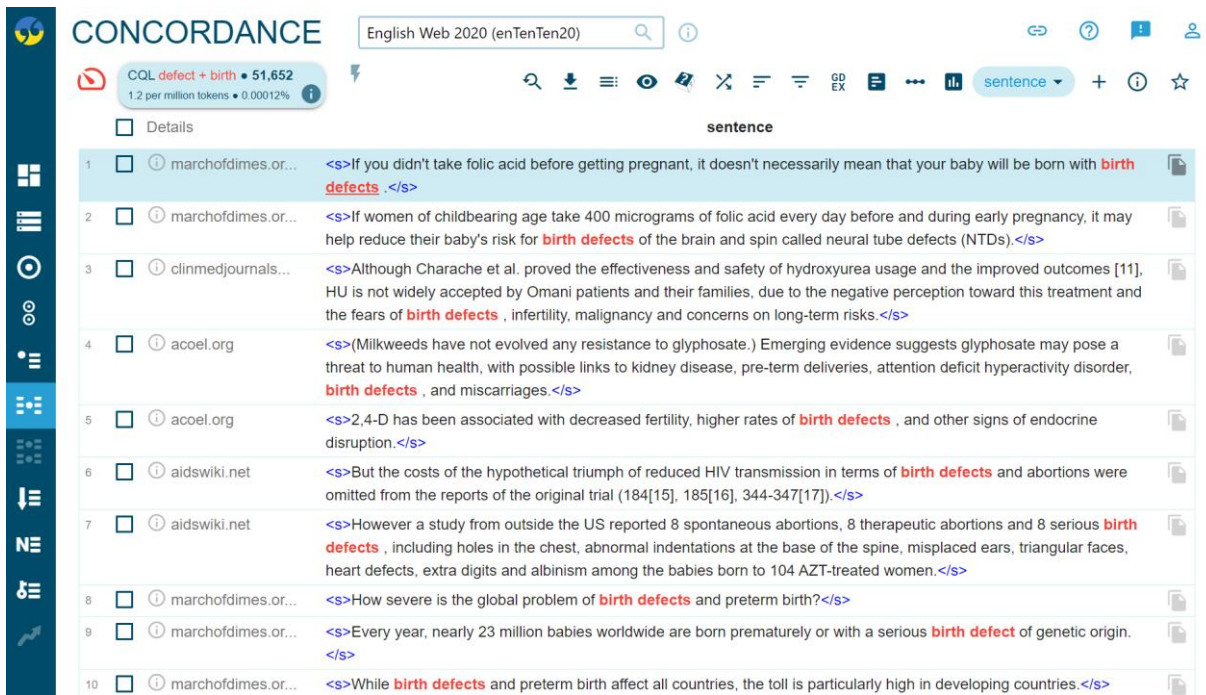


Figure 3.8 Concordance view of collocate *Birth* and its node in English Web 2020

There would also be a possibility of new meanings emerging from the splitting of word meaning(s) in the sense inventory. For example, the meaning of *defect* as “an imperfection in a person or thing”, which appeared to contain overlapping meanings—“something wrong with

part of the body” and “something that is not perfect”—could be further split up into two distinct meanings. The splitting decision would be made by checking collocational and grammatical patterns exported by Word Sketch to see whether “something wrong with part of the body” and “something that is not perfect” satisfied Evans’s (2005) second and/or third criteria to determine distinct meanings. If so, they would be split into two rather than merged into one.

3.3.2.3 Examine the word meaning frequency

The frequency of a meaning of a word was examined via the frequency of the word’s collocates. Word Sketch selects collocates for inclusion in the result page based on frequency and typicality. The two scores are automatically calculated and by default presented along with collocates in parallel columns, e.g., *generic* (Frequency: 9,906 and Typicality: 7.8) in Figure 3.7. Frequency indicates how frequently a collocate co-occurs with its node. For example, *generic* (Frequency: 9,906) is found to co-occur more frequently with *defect* than *structural* (Frequency: 4,727) (see in Figure 3.7). Typicality indicates which collocate is more likely to co-occur with a particular node. For example, *structural* (Typicality: 7.0) is a more typical collocate of *defect* than *other* (Typicality: 2.4).

Collocates in Word Sketch were sorted by their frequency scores to help create a shortlist of the most frequent ones. Typicality was also referred to because some frequent collocates with a lower typicality score appeared to be less useful in disambiguating their node word than ones with a higher typicality score. This is obvious in the cases of *structural* (Typicality: 7.0) and *other* (Typicality: 2.4). It is easier to assign a meaning to *defect* when it co-occurs with *structural* than with *other*. Collocates found their way to the shortlist by having high frequency and acceptable typicality scores (see a discussion of the cut-off line in 5.3.2). Meanings were then assigned to collocates in the shortlist, eventually revealing the most frequent meanings of a word found in medical and general contexts.

3.4 The piloting of SemiMed

3.4.1 Ethics approvals

The study was covered by ethics approvals from the University of Adelaide's Office of Research Ethics, Compliance and Integrity (No: H-2022-004) and the Institutional Ethics Committee of a University of Medicine and Pharmacy in Vietnam (UMP) (No: H2022/015) (see Appendices 4 and 5). In accordance with ethical guidelines for low-risk research, this study undertook to obtain voluntary participation, to inform consent and to maintain the confidentiality of participants throughout the conduct of the pilot phase and in the publication of relevant research findings.

3.4.2 Participant recruitment

3.4.2.1 Recruitment setting and procedure

SemiMed targets student users whose majors are medicine-related and who learn English as a foreign/second language (EFL/ESL). The SemiMed pilot study therefore recruited participants from a university of medicine and pharmacy in Vietnam where the medical students are EFL learners. The recruitment procedure, aligned with ethical guidelines, is presented in Figure 3.9.

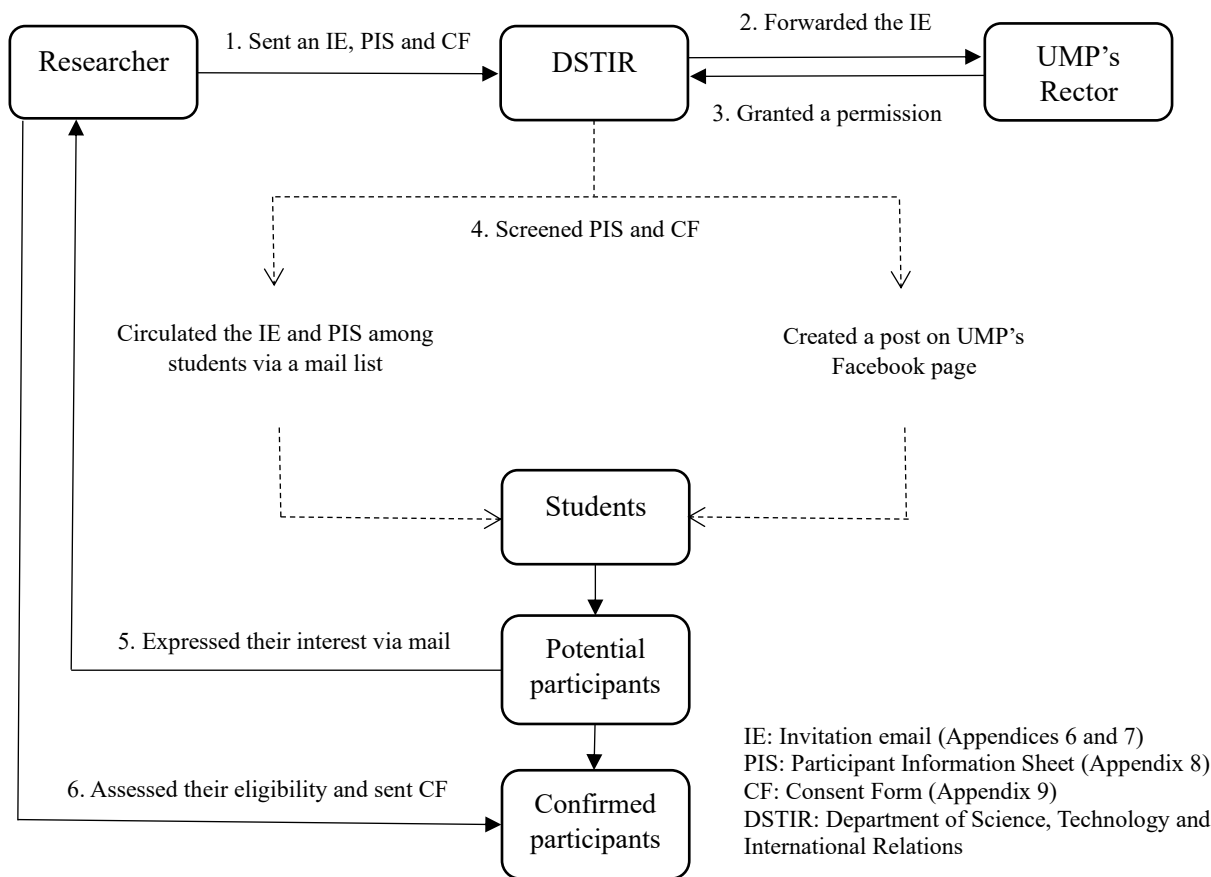


Figure 3.9 Participant recruitment procedure

3.4.2.2 Recruitment criteria

To identify participants for the pilot study, the following inclusion criteria were used.

Be 18 to 24 years of age (Optional): The study sought undergraduates who were enrolled in/had completed English for medical courses, to ensure they had sufficient background to perform the required tasks. Undergraduate medical students usually fall into the 18-24 age range (undergraduate medicine degree programs take between four and six years of study). Students above this age range but satisfying the two criteria below were also included in the study.

Majoring in a medical field of study (Mandatory): Participants had to be undertaking medicine-related programs, including but not limited to General Medicine, Odontostomatology, Preventive Medicine, Traditional Medicine, Nursing, Medical Laboratory

Techniques, Medical Imaging Techniques, Public Health and Midwifery. Participants who majored in pharmacy were also included.

Meet English language requirements (Mandatory): Since SemiMed is a monolingual resource and the pilot study was conducted in English, participants had to have attained a standard of English at upper-intermediate level and above (equivalent to IELTS 5.5/ TOEFL 46-59/ CEFR B2 and above) so that they could perform assigned tasks satisfactorily. They were required to provide valid test results by the time the pilot study commenced.

3.4.2.3 Participant demographics

Eighteen students who satisfied the selection criteria were recruited for this study. Of the 18 participants, eight were male and ten were female. All participants were between 19 and 24 years of age and had valid IELTS results, ranging from 6.0 to 8.0. Most (13) of the participants majored in General Medicine, some (3) majored in Odontostomatology, and the other two majored in Traditional Medicine (1) and Pharmacy (1).

Table 3.3 Demographic profile of participants

No	Name	Gender	Age	Major	IELTS
1	Participant A	Male	20	General Medicine	6.5
2	Participant B	Male	23	General Medicine	7.5
3	Participant C	Male	24	General Medicine	7.5
4	Participant D	Male	20	General Medicine	7.5
5	Participant E	Female	19	Odontostomatology	7.0
6	Participant F	Female	20	General Medicine	7.0
7	Participant G	Female	20	General Medicine	6.0
8	Participant H	Female	20	Odontostomatology	7.5
9	Participant I	Male	24	General Medicine	6.5

10	Participant J	Female	24	General Medicine	6.0
11	Participant K	Female	19	Pharmacy	7.0
12	Participant L	Male	22	Traditional Medicine	6.0
13	Participant M	Male	21	General Medicine	6.5
14	Participant N	Female	19	General Medicine	7.5
15	Participant O	Female	24	General Medicine	8.0
16	Participant P	Male	21	General Medicine	7.0
17	Participant Q	Female	19	Odontostomatology	7.5
18	Participant R	Female	23	General Medicine	6.0

3.4.3 Data collection

The pilot study consisted of two parts. Participants were requested to: (a) use SemiMed and designated conventional dictionaries to perform medical role-plays; and (b) provide feedback on the use of SemiMed compared with the use of conventional dictionaries (see 6.3.2 for more details). Due to COVID limitations, the study was conducted virtually over Zoom. The 18 participants were divided into six groups. A 60-minute Zoom meeting was scheduled for each individual group. The six meetings took place over the course of a month. All participants gave the researcher consent to record Zoom meetings (video and audio). Data were collected using observations and focus groups.

3.4.3.1 Observations

Overt observations (Cohen et al., 2011) were conducted, where participants were informed that they were being observed while role playing. Each group of three participants was guided on how to prepare for a randomly assigned medical scenario. There were five scenarios, each relating to a medical topic: bowel, eye, heart, liver and pregnancy. Each scenario had three characters (a patient, specialist and nurse) and was approximately 100-200

words in length (see Appendix 10). Participants looked up the essential vocabulary required for their particular scenario using SemiMed and designated conventional dictionaries to ensure understanding before acting out their parts. The primary purpose of the observations was to capture participants' interactions with and attitudes toward SemiMed and the designated conventional dictionaries through role-plays. The participants were observed for the entire time, from commencing preparations for their performances until their role-plays ended. Notes were taken during observations to record relevant data, which later informed and was cross-checked with focus group data.

3.4.3.2 Focus groups

Follow-up focus groups took place in the last 30 minutes of the Zoom meetings. Each focus group was structured based on a pre-determined set of nine questions (see Appendix 11) that prompted in-depth discussions around the experience of using SemiMed and the designated conventional dictionaries. The focus group discussion began with a generic question asking about academic majors to encourage participants to introduce themselves to each other and engage in discussions. Two following questions were about their experience in using conventional dictionaries prior to and during role playing. The next five questions focused on SemiMed, more particularly, their experience in using SemiMed while role playing, the potential of SemiMed beyond medical role-plays, and the strengths and weaknesses of SemiMed in comparison with conventional dictionaries. The last question sought suggestions from participants to improve SemiMed. The discussions were moderated to ensure that everyone had an equal chance to voice their opinion. In addition to the pre-set questions, other questions, informed by observation notes, were raised where necessary to gain a complete view of SemiMed's practicality and usefulness.

3.4.4 Data analysis

Focus group recordings were manually transcribed, replacing each participant's name with a pseudonym (see Table 3.3) so as not to disclose personal information. The transcriptions were then analysed using thematic analysis. Two main themes emerged: experience in using (a) SemiMed and (b) conventional dictionaries. Sub-themes (including both those anticipated and unexpected) were classified under the main themes. Theme-based data were recorded and catalogued in NVivo to assist in finding and making sense of connections between themes.

3.5 Methodological limitations

First, since the core meaning-based analysis was proposed as a solution to the highly undesirable level of subjectivity in native speaker judgement, three evaluators (two native and one non-native English speaker) are still considered the minimum number that warrants the inter-evaluator reliability of outcomes. Although this was satisfactory within the scope of this study, a larger number of evaluators would be more desirable.

Second, owing to time constraints, the development of SemiMed only targeted a small-sized sample of the MWL. This study therefore was unable to address all the words in the list and their core meanings.

Third, although SemiMed's format was able to be electronically transferred, it was not ideally suitable for printed resources, because meaning clusters in individual LCs may take up considerable space on paper.

Fourth, the foci of this study were necessarily restricted to one specific type of vocabulary (semi-technical medical vocabulary) and its semantic aspects (polysemy and homography); consequently, SemiMed comprises mainly content words and is meaning-focused.

CHAPTER 4: THE EXAMINATION OF HSU'S (2013) MEDICAL WORD LIST

Statement of Authorship

Title of paper	A core meaning-based analysis of English semi-technical vocabulary in the medical field
Publication status	Published – revised for this thesis for stylistic consistency
Publication details	Le, C. N. N., & Miller, J. (2023). A core meaning-based analysis of English semi-technical vocabulary in the medical field. <i>English for Specific Purposes</i> , 70, 252-266.
Conference presentation	Le, C. N. N. (2021, September 1-2). <i>Are word form frequency-based lists reliable? A close examination of Hsu's (2013) Medical Word List</i> [Paper presentation]. AustralaLex Conference, Auckland, New Zealand.

Principal Author

Name of principal author (Candidate)	Chinh Ngan Nguyen Le		
Contribution to the paper	Researched and developed conceptual framework, performed all data collection and analysis, interpreted data, developed first draft, wrote and revised manuscript, and acted as corresponding author.		
Overall percentage (%)	80%		
Certification	This paper reports on original research I conducted during the period of my Higher Degree Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis.		
Signature		Date	21/11/2023

Co-author contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of co-author	Julia Miller		
Contribution to the paper	Supervised development of work, helped in data analysis and interpretation and manuscript revisions.		
Signature		Date	21/11/2023

A core meaning-based analysis of English semi-technical vocabulary in the medical field

Abstract Semi-technical vocabulary, a type of vocabulary with both a technical and non-technical meaning (e.g., *colon*: part of the large intestine; punctuation mark), is an area of controversy owing to disagreement over its definition and characteristics. While it is widely held that learning technical vocabulary is critical for learners of English for Specific Purposes (ESP), several studies have also focused on semi-technical vocabulary because these words often have multiple meanings, depending on the context, and may therefore be harder to learn and understand than purely technical words. This study aims to revisit semi-technical vocabulary in medicine to address these controversial issues by re-evaluating a 595 semi-technical medical word list developed by Hsu (2013). A core meaning-based analysis identified 302 potentially confusing semi-technical medical words. These are mostly mid-frequency words; some are academic and low-frequency words. The findings also revealed pedagogic challenges associated with word form frequency-based lists deserving of further research.

Keywords: Semi-technical medical, core meaning, English for Specific Purposes, wordlist, vocabulary

4.1 Introduction

For decades, vocabulary acquisition has been viewed as a key component of English language learning. In the realm of English for Specific Purposes (ESP), technical vocabulary is of primary importance, as this type of vocabulary is central to specialized materials. Nevertheless, what learners find problematic regarding acquisition and understanding are not technical words themselves but vocabulary that is semi-technical in nature. Such words (e.g., *stool*) lie between technical and non-technical vocabulary, and their meanings vary according to context, making them challenging to learn or teach (Durrant, 2009; Hyland & Tse, 2009; Gardner, 2007; Li & Pemberton, 1994; Shaw, 1991; Thurston & Candlin, 1998).

The focus in this paper is on one area of ESP, namely English for Medical Purposes (EMP). In learning and teaching EMP, semi-technical vocabulary has caused more problems than fully-technical medical terminology in terms of meaning interpretation (Li & Pemberton, 1994; Shaw, 1991; Thurston & Candlin, 1998). Medical terminology usually has static meanings across different contexts. Semi-technical vocabulary, however, has not only non-technical meanings but tends to carry additional meanings when appearing in the medical context. These additional meanings sometimes differ from generally used meanings. The word *stool*, for example, is frequently understood as the “solid waste from the body” in the medical context. This meaning is only distantly related to its general meaning of “a wooden seat” and consequently may challenge EMP learners.

Until now, the problems associated with semi-technical language in medical practice have been addressed through the development of wordlists; however, such wordlists do not indicate word meanings that vary according to context. In order to address this situation, we conducted an extensive investigation of semi-technical vocabulary in medicine, to characterize semi-technical medical words and highlight the shortcomings of the existing semi-technical medical wordlists without semantic explanations. This paper argues that semantic analysis of wordlists is critical to initiate an improvement in current material resources.

4.2 Literature review

4.2.1 Nation’s lexical categories

Nation (2018) developed a 25,000-word list compiled chiefly from the British National Corpus (BNC) and Corpus of Contemporary American English (COCA). On the basis of frequency, words in the BNC/COCA list are sorted in descending order and divided into 1,000-word bands covering three lexical categories (Nation, 2013): high-frequency words (the first 2,000 words), mid-frequency words (from the third to the ninth 1,000 words), and low-frequency words (from the tenth 1,000 words onwards).

Schmitt and Schmitt (2014) adhere to Nation's (2013) proposed categorization, except for an amendment through which high-frequency vocabulary expands to 3,000 words and low-frequency vocabulary is lowered to the 9,000 level. They base their change on the re-evaluation of previous studies on vocabulary size and word frequency. Although Schmitt and Schmitt's (2014) category slightly differs from Nation's (2013), they agree with Nation that "8,000-9,000 word families are sufficient to provide the lexical resources necessary to be able to read a wide range of authentic texts" (Schmitt & Schmitt, 2014, p. 484). We believe that Nation's (2013) lexical categories are comprehensive and present them in detail below.

High-frequency vocabulary is perceived as basic English words most frequently encountered in spoken and written discourse (Chen & Ge, 2007; Hsu, 2013). The most classical work that delimits the size of high-frequency vocabulary is West's (1953) General Service List (GSL), listing around 2,000 headwords, accounting for 80% lexical coverage across academic texts (Nation, 2013). Updates to West's (1953) GSL are still being provided—for example, Brezina and Gablasova's (2017a) 2,494-headword New General Service List and Browne's (2014) 2,800-headword New General Service List—suggesting that high-frequency vocabulary might scale up to 3,000 words, which is in line with Schmitt and Schmitt's (2014) suggestion.

Mid-frequency vocabulary comes after high-frequency vocabulary and ranges from the third to the ninth 1,000 words (Nation, 2013). This type of vocabulary varies between 6,000-7,000 words, depending on whether high-frequency vocabulary is perceived to be within 2,000 or 3,000 words. Mid-frequency words are worth learning because "together with the high-frequency words, they represent the amount of vocabulary needed to deal with English without the need for outside support" (Nation, 2013, p. 18)

Low-frequency vocabulary, as its name suggests, is understood to mean "words that we rarely meet in our use of the language" (Nation, 2013, p. 19), including technical terms.

Although the low-frequency vocabulary size is relatively large, i.e., from the tenth 1,000 words onwards (Nation, 2013), this type of vocabulary provides only modest coverage in academic discourse. Nation (2013, p. 19) also further described two categories that are listed under “specialized vocabulary” as follows:

Academic vocabulary is considered as lexical items “that are common in different kinds of academic texts” (Nation, 2013, p. 19). The Academic Word List (AWL) was developed by Coxhead (2000) to itemize the words most frequently found in academic reading materials. According to Nation (2013, p. 30), the AWL “is drawn from words from the third 1,000 to the seventh 1,000” and thus fits in the area of mid-frequency vocabulary. The resulting 570-word list claims around 10% extra coverage of academic texts in addition to the 80% coverage given by West’s (1953) GSL’s top 2,000 words.

Technical vocabulary refers to “words that are very closely related to the topic and subject area of the [texts], . . . reasonably common in this topic area but . . . not so common elsewhere, [and] . . . [different] from subject area to subject area” (Nation, 2013, p. 19). According to the text, technical vocabulary may be considered high-, mid-, low-frequency and academic (Nation, 2013). Lists of technical words are compiled by examining the frequency of these words in discipline-specific corpora. Identifying sets of technical words frequently found in particular disciplines aims to increase the aggregate coverage of West’s (1953) GSL, Coxhead’s (2000) AWL and a list of technical words to 95-98%, a text coverage that Laufer and Ravenhorst-Kalovski (2010) propose is desirable to gain adequate comprehension of reading materials.

4.2.2 Semi-technical vocabulary

Among Nation’s (2013) lexical categories, technical vocabulary is pivotal in ESP teaching and learning. The “technical” concept is perplexing because it is not dichotomous; in other words, it is not always possible to say whether a word is either technical or non-technical.

Moreover, studies on technical vocabulary acknowledge the existence of semi-technical vocabulary, which locates between non-technical and technical vocabulary.

4.2.2.1 Previous studies on semi-technical vocabulary

Researchers have long disagreed over the nature of semi-technical vocabulary and how it should be named. Higgins (1966) uses the term “frame words”. Cohen et al. (1988) prefer the term “specialized non-technical lexis”. Li and Pemberton (1994, p. 184) call it “subtechnical” or “semi-technical” vocabulary and describe it as being “context-independent”. Cowan (1974), Flowerdew (1993) and Huizhong (1986) focus on the nature of semi-technical vocabulary and state that it occupies the space between generally used and highly specialized vocabulary. In this regard, semi-technical vocabulary has not only a general but also a technical meaning. A semi-technical word meaning is interpreted depending on a particular context where the word is found. For example, the word *orbit* (as a noun) indicates “one complete circuit made by an object around the orbited body” and carries an additional meaning when it appears in the medical context: “the eye socket” (*Oxford English Dictionary*, 2021).

Although Cowan (1974), Flowerdew (1993), and Huizhong (1986) point out that semi-technical vocabulary lies in an area between non-technical and technical vocabulary, they do not articulate whether it overlaps with other types of vocabulary. Other linguists have made greater efforts to specify where semi-technical vocabulary lies in relation with other types of vocabulary. Fraser (2007, 2009, 2012) labels semi-technical words “cryptotechnical” words, by which he means polysemous words with a meaning that may be obscure to the non-specialist. He asserts that more than 12% of cryptotechnical words in his Pharmacology Word List also appear in Coxhead’s (2000) AWL. Watson-Todd (2017) investigated semi-technical words (referred to as “opaque words” in his study) and suggested that high-frequency vocabulary sometimes takes on technical meanings in discipline-specific contexts, thus revealing itself as semi-technical vocabulary. This finding is consistent with Quero and

Coxhead's (2018) observations in their study on high-frequency vocabulary in medical contexts.

4.2.2.2 A taxonomy of semi-technical vocabulary in the field of medicine

To comprehensively perceive semi-technical vocabulary, drawing on the work of Baker (1988) and Fraser (2009), Hsu (2013, p. 467-468) devised a taxonomy to classify semi-technical vocabulary in the field of medicine into distinct groups:

(1) Words, themselves and/or their family members, [that] express some academic notions, approaches or procedures, and can be found across a wide range of disciplines . . . (2) Words of general use whose technical meaning may be hidden and only emerge from the context . . . (3) Words [that] are used equally with general and specialized meanings . . . (4) Words with a medical dress [that] may undergo a semantic transfer when used in general language . . . (5) Words [that] reveal a technical sense . . . , mainly used in the medical register . . . (6) Words [that] are used almost exclusively in the medical contexts . . .

4.2.2.3 Difficulties in learning and teaching semi-technical vocabulary

Hsu's (2013) six-level taxonomy showcases the complex nature of semi-technical vocabulary. Moreover, previous studies (Fraser, 2007, 2009, 2012; Quero & Coxhead, 2018; Watson-Todd, 2017) have uncovered a non-clear-cut boundary between semi-technical and other types of vocabulary, implying that this type of vocabulary has non-transparent characteristics, highlighting the necessity to take semi-technical vocabulary into account in ESP courses. Some who hold a view that learners should concentrate on technical vocabulary may remain skeptical about the significance of revisiting semi-technical vocabulary. It is thus worth reiterating that the mastery of technical vocabulary alone is, according to Cohen et al. (1988), insufficient to achieve successful reading of specialized materials. Additionally, multiple studies have claimed that semi-technical vocabulary is more problematic than technical vocabulary in terms of learning and teaching (Li & Pemberton, 1994; Shaw, 1991; Thurston & Candlin, 1998). This is because meaning variation increases from technical to semi-technical vocabulary (Gardner, 2007). Technical words are mostly single-meaning and

consistently used in a particular discipline, so they may be easier to acquire. Conversely, due to their hybrid nature, semi-technical words usually carry more than one meaning, which can cause confusion for learners (Durrant, 2009; Hyland & Tse, 2009). It is generally agreed that while learners are familiar with general meanings, they may not be aware of additional meanings activated in technical contexts; as a result, they are unable to interpret semi-technical vocabulary accurately when reading specialized materials (Cohen et al., 1988; Hyland & Tse, 2009). Moreover, due to the unclear boundary between semi-technical and other types of vocabulary, neither ESP nor content teachers deliver direct instruction to equip learners with semi-technical vocabulary learning strategies (Durrant, 2009; Hyland & Tse, 2009; Peters & Fernández, 2013).

4.2.3 Wordlists of semi-technical vocabulary in the medical field

The lack of explicit instruction on semi-technical vocabulary in the ESP classroom gives learners no other choice but to rely on specialized dictionaries. However, even specialized dictionaries do not include semi-technical words (Peters & Fernández, 2013). Wordlists have therefore been created as an alternative custom-made teaching (or learning) resource for ESP.

In medical disciplines, Hsu's (2013) Medical Word List and Wang et al.'s (2008) Medical Academic Word List are two well-known semi-technical lists whose words were extracted from two corpora: Hsu's (2013) corpus of medical textbooks and Wang et al.'s (2008) corpus of medical research articles. To be included in the lists, a word had to satisfy all three of the following criteria: *specialized occurrence* (occurrence in medicine-related texts), *range* (the number of texts in which a word is repeated), and *frequency* (the number of occurrences of a word across different texts). The lists are based on relevant corpora representing target material sources learners usually encounter, and sound criteria were adopted to rank words in order of frequency of occurrence (see Table 4.1). There is a general consensus that the 623-word Medical Academic Word List (Wang et al., 2008) and 595-word Medical Word List (Hsu,

2013) are beneficial in respect of indicating which words are worth an investment of learners' time and effort; i.e., the higher a word is ranked, the more frequently it occurs in medical materials, and the more time and effort should be devoted to mastering the word.

Table 4.1 Development of the Medical Academic Word List and Medical Word List

Wang et al.'s (2008) Medical Academic Word List (623 words)	Hsu's (2013) Medical Word List (595 words)
Medical Article Corpus: A one million-word corpus of 288 medical research articles across 32 subject areas	Medical Textbook Corpus: A 15 million-word corpus of 155 medical textbooks across 31 medical subject areas.
Selection criteria:	Selection criteria:
Specialized occurrence: Outside West's (1953) 2,000-word GSL	Specialized occurrence: Outside the first BNC 3,000 words
Range: Occur in more than half of 32 medical subject areas	Range: Occur in more than half of 31 medical subject areas
Frequency: Occur at least 30 times in the corpus of medical articles	Frequency: Occur at least 863 times in the Medical Textbook Corpus

Although the lists constitute useful reference resources for teaching EMP, some downsides remain. The corpus-based automatic calculation seems to treat word meanings superficially because it can only recognize and count word forms, regardless of their variant meanings. Indeed, a form-meaning issue arises when it appears to calculate the frequency of words with multiple related meanings (polysemes) and unrelated meanings (homographs). There is a likelihood of “[overestimating] the true coverage of word form” (Gardner, 2007, p. 253) by virtue of relying solely on the frequency of word form rather than word meaning. Moreover, Wang et al.'s (2008) Medical Academic Word List and Hsu's (2013) Medical Word

List, which are purely lists of word forms, do not provide learners with further semantic annotations of the words they contain. The indication of the most frequently occurring word forms in the two wordlists is beneficial in terms of narrowing down the number of words to a manageable level. However, in the case of polysemous words, learners will gain very few clues about how many meanings a word has and which one(s) is/are used in the medical context. It is essential to realize that statistical evidence (*range* and *frequency*) may play a significant role in showing learners which words should be learned but not how words can be interpreted, and thus the pedagogical applications of the lists are limited.

4.3 The study

4.3.1 Aims and research questions

Within the scope of this paper, a semi-technical wordlist in the medical field, Hsu's (2013) Medical Word List, will be re-evaluated for two reasons. First, compared to Wang et al.'s (2008) Medical Academic Word List, it overlaps less with Coxhead's (2000) AWL, whose words were already semantically examined by Wang and Nation (2004), so Hsu's (2013) Medical Word List is expected to provide richer input for the study. Second, it is assumed that Hsu's (2013) Medical Word List may contain methodological enhancements because it was more recently developed than Wang et al.'s (2008) Medical Academic Word List, which may make it more pedagogically useful. There are three main research questions.

- a. Where does semi-technical medical vocabulary sit on the vocabulary continuum?
- b. What words in Hsu's (2013) Medical Word List (MWL) can be identified as possessing multiple meanings?
- c. What are the main disadvantages of semi-technical medical wordlists based on word form frequency?

4.3.2 Methodological framework

Polysemy and homography: These two linguistic phenomena indicate, respectively, a word having multiple related meanings and two words sharing the same form (Murphy, 2010). In essence, polysemes are etymologically related. They are rooted in a lexical source and extend their original meaning over time so that old words are, according to Murphy (2010), used in new ways. Homographs are accidentally orthographically identical due to language change or borrowing, for example; therefore, unlike polysemes, they are etymologically unrelated. Most dictionaries use etymological evidence to differentiate polysemy from homography. Nevertheless, the etymology-based distinction “is not always straightforward, especially since words that are etymologically related can, over time, drift so far apart that the original semantic relation is no longer recognizable” (Ravin & Leacock, 2000, p. 2).

The scale of semantic relatedness (pre-pilot): Owing to the limitations of an etymology-based approach, native speaker intuition was proposed as a means of distinguishing between polysemy and homography. In the simplest terms, the relatedness or unrelatedness of meanings can be judged by native speakers and the distinction may be not a dichotomy but rather a continuum (Cruse, 2000; Klepousniotou, 2002; Murphy, 2010). Wang and Nation (2004), in their study examining homography in Coxhead’s (2000) AWL, developed a rating scale that permits evaluators to rely on their intuitive judgment to measure the degree of semantic relatedness. The scale has six levels (from *Level 0: Close relation* to *Level 5: No relation*) and the cut-off point is at *Level 3*, indicating that any meanings ranked at *Level 4* or *5* are homographs. For instance, five meanings of the word family *issue* (AWL Sub-list 1) were intuitively ranked by an evaluator, as shown in Table 4.2. According to Wang and Nation (2004), meanings 1, 3 and 4 (which are indicated in bold in Table 4.2) relate to each other at either *Level 4* or *5*, so they are three groups of homographs (*an important topic, flowing, and children*). As meanings 2 and 5 are more closely related to *flowing* (at *Level 3* and above) than

an important topic and *children* (below *Level 3*), they are placed in the *flowing* group (see Table 4.3 for our summary of Wang and Nation’s homograph groups of the word family *issue*).

Table 4.2 Wang and Nation’s (2004, p. 302) scale of semantic relatedness for *Issue*

Level of semantic relatedness	Dictionary definitions				
	1) an important topic	2) the action of distributing	3) children	4) the action of flowing	5) a result or outcome
<i>Level 0:</i> The same					
<i>Level 1:</i> Slightly different	5) a result or outcome 4) the action of flowing				
<i>Level 2:</i> Related with some changes					
<i>Level 3:</i> Substantially different but related	4) the action of flowing 2) the action of distributing				

Level 4:	4) the action	3) children;	2) the action	1) an	1) an
Very	of flowing;	5) a result	of	important	important
distantly	5) a result or	or outcome	distributing;	topic;	topic;
related and	outcome		4) the action	3) children	2) the action
almost			of flowing;		of
totally			5) a result or		distributing;
different			outcome		3) children
Level 5: No	2) the action	1) an	1) an		
relation	of	important	important		
	distributing;	topic	topic		
	3) children				

Table 4.3 Summarized interpretation of meanings for the word *Issue* using Wang and Nation’s scale of semantic relatedness

<i>An important topic</i>	1) an important topic
<i>Flowing</i>	2) the action of distributing
	4) the action of flowing out
	5) A result or outcome
<i>Children</i>	3) Children

Core-meaning theories: We perceived that the scale of semantic relatedness is potentially relevant to our study, so we pre-piloted a re-evaluation of 25 MWL headwords, adopting Wang and Nation’s scale to confirm whether it applies to the identification of polysemy and homography in the MWL. The ranking of the 25 headwords elicited considerable

disagreement among our three evaluators, leading to a further step: re-doing two headwords using *Think Aloud Protocol* (TAP) (Charters, 2003), in which the evaluators recorded their reflections on how they came up with their final evaluation. The two main points retrieved from the TAP are summarized below:

- a. The scale requires subjective judgment; therefore, when more than one evaluator was involved, both inter- and intra-evaluator reliability were difficult to maintain. Not only did we have some discrepancies when comparing our findings, but each of us had different responses to the same evaluated headword at different times. Significantly, two evaluators experienced an asymmetric evaluation, placing a pair of meanings at two different levels. As can be seen in Table 4.2, Wang and Nation's pair of meanings 2 and 4 was placed at the same level, showing a symmetric relation: 2 relates to 4 at *Level 3* and 4 relates to 2 at *Level 3*. Our two evaluators sometimes put, for example, 2 and 4 at one level and 4 and 2 at another level.
- b. The unsatisfactory inter- and intra-evaluator reliability scores obtained from the 25-headword pre-pilot do not imply that the scale is unreliable. In fact, Wang and Nation (2004) emphasized their scale benefited an analysis of homography, while our study's primary focus is on both polysemy and homography. Lehrer (1974) reasons that there is more stability in the judgment formed by native speakers on homography than polysemy, so inconsistent pre-pilot results are inevitable.

Wang and Nation (2004) state that Ruhl's (1989) Monosemic Bias frames their scale. The Monosemic Bias is in line with studies (Caramazza & Grober, 1976; Geeraerts, 2010; Klein & Murphy, 2001; Klepousniotou et al., 2008) which acknowledge the existence of a core/central (part) of the meaning that is present in almost all senses of a word. We observed that even though the polysemy-homography continuum underpins the scale development, in the end, meanings were put in each distinctive group representing a shared core meaning.

Hence, to enhance the reliability, we simplified Wang and Nation's analysis procedure and requested evaluators to identify only core meaning(s). If all senses of a word shared a common core meaning, the word would be deemed polysemous. If a core meaning could not be located, any senses that created a new core meaning would represent homographs. Additionally, we found etymology helpful in informing our judgment. Knowing the meaning of a Greek root, for example, helped us to establish a word's core meaning. We piloted the core meaning theory-driven method in polysemy-homography analyses of 10 headwords in the MWL. The pilot findings revealed that the three evaluators reached agreement on eight out of 10 words (80%), which is slightly higher than Wang and Nation's reliability score of two evaluators (75%). We therefore considered this satisfactory. The entire procedure of analyzing polysemy and homography phenomena in the MWL in the light of core meaning theories is described below.

4.3.3 Research procedures

Hsu's (2013) MWL showcases headwords of 595-word families with their range and frequency presented alongside. In what follows, an individual headword is listed to represent related members in a word family; for example, *diagnosis* stands alone in the list as a representative of its derived and inflected forms (*diagnosable*, *diagnose*, *diagnoses*, *diagnosing*, and *diagnosed*). The analysis involved three stages.

Stage 1. The 595 MWL headwords were looked up in the *Oxford English Dictionary* (OED), a historical dictionary which comprehensively captures meanings of a word and systematically groups word meanings based on their semantic relatedness. In this way, the OED promotes inclusiveness and systematization in the presentation of word meanings.

Stage 2. To ensure the viability of the 595-word semantic re-evaluation, the researchers then put OED definitions in an input file and, where necessary, merged them. Only current definitions were used; obsolete definitions were discarded. *Fistula*, for example, has two obsolete meanings (3 and 4) among four originally listed meanings (Figure 4.1); therefore, the

researchers decided to discard the obsolete meanings and only retain the first two meanings. Within meaning 1, three sub-meanings (a, b, and c) were merged into one because sub-meaning *a* seemingly covers *b* and *c*. *Fistula* eventually appeared in our input file with two meanings (see Figure 4.1 and Figure 4.2). 302 MWL headwords were identified as having more than one meaning.

Stage 3. Three evaluators (one with English as an additional language and two with English as their first language) undertook the re-evaluation of these 302 MWL headwords (Table 4.4). One evaluator, who comes from a medical background, worked collaboratively with the other two ESL/EFL/ESP lecturers having expertise in lexical semantics to identify core meaning(s) shared by listed dictionary definitions of an examined word. Such a collaboration is considered reciprocal, because evaluator 3's insight into medical knowledge assisted evaluators 1 and 2 with terminological complications in certain medical meanings. Evaluators 1 and 2, in their turn, provided scaffolding from which evaluator 3 shaped her conceptualization of the two linguistic phenomena (polysemy and homography).

Table 4.4 Evaluator details

Evaluator	English language background	Professional background	Working experience
Evaluator 1	Non-native speaker	University lecturer	Over 10 years' EFL and ESP teaching experience
Evaluator 2	Native speaker	University lecturer	28 years' experience in teaching EFL and EAP
Evaluator 3	Native speaker	University lecturer and physiotherapist	Over 40 years' experience of teaching physiotherapy at the tertiary level and working as a physiotherapist

Fistula, n.

1.

a. *Pathology*. A long, narrow, suppurating canal of morbid origin in some part of the body; a long, sinuous pipe-like ulcer with a narrow orifice.

b. In animals, birds, etc.

c. Also applied to certain passages in the body made surgically.

2. A natural or normal pipe or spout in cetaceous animals, insects, etc.

3. *Ecclesiastical*. A tube through which in early times communicants received the consecrated wine; now used by the Pope only.

4. *Music*. A reed instrument or pipe of the ancient Romans.

Figure 4.1 The word entry *Fistula* in the OED

Fistula	<i>BNC: Band 14</i>	<i>Range: 27</i>	<i>Frequency: 1,394</i>
a) n. (Pathology) A long, narrow, suppurating canal of morbid origin in some part of the body; a long, sinuous pipe-like ulcer with a narrow orifice.			
b) n. A natural or normal pipe or spout in cetaceous animals, insects, etc.			
Core meaning(s)	a)	b)	
Core meaning 1: Pipe-like	✓	✓	

Figure 4.2 The semantic re-evaluation of *Fistula*

Each evaluator was expected to undertake three steps: (1) read through all listed dictionary definitions of a word, (2) identify (how many) core meanings a word had and write them down, (3) point out dictionary definitions that share the same core meaning. By way of illustration, after skimming two dictionary definitions of *fistula*, it can be perceived that definitions *a* and *b* converged in the sense of “something that has a pipe-like shape”. Thus,

“pipe-like” was the only core meaning identified, and the re-evaluation result was tabulated as in Figure 4.2.

To ensure the 3-step evaluation was consistently executed, the three evaluators initially worked on ten pilot words independently of each other and then discussed their questions about core-meaning identification. After reaching a mutual understanding of the entire process, they worked at their own pace to evaluate around 96-98 words per week for three weeks. Weekly discussions were scheduled to compare preliminary results. The last follow-up discussion centered on words that caused disagreement among the evaluators. In these cases, we sometimes referred to etymological roots to help resolve the disagreement. The process is outlined in Figure 4.3.

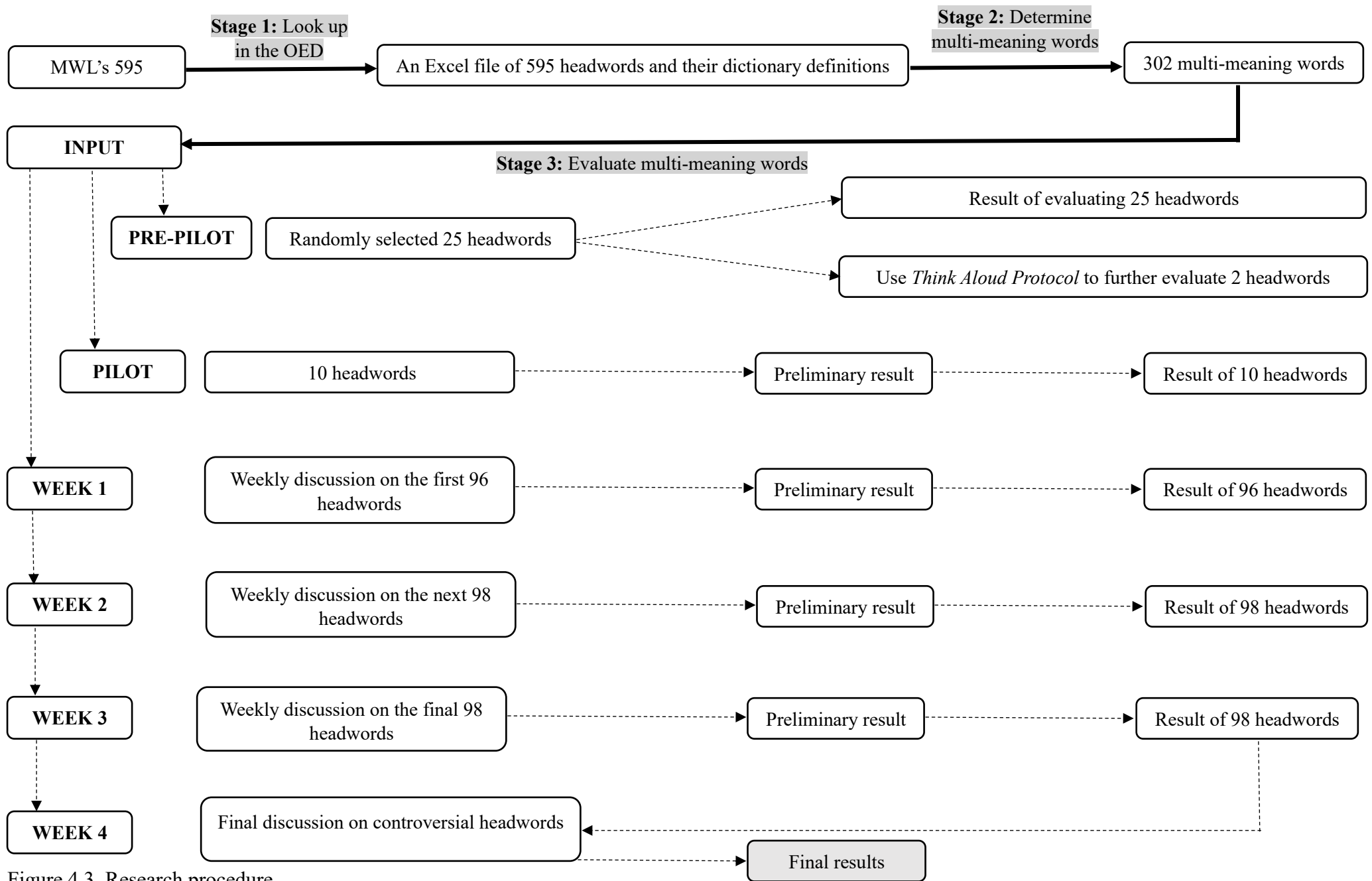


Figure 4.3. Research procedure

4.4 Results and discussion

4.4.1 Examining the boundary of semi-technical medical vocabulary

According to Hsu (2013), the MWL headwords range from the BNC 4th to 10th bands, and 76 are found in Coxhead's (2000) AWL. It is noteworthy that the creation of the MWL goes beyond 3,000 high-frequency words, and we acknowledge Quero and Coxhead's (2018) findings that high-frequency words may become semi-technical vocabulary in the medical context. Taking this into account, the MWL establishment and our review of literature mark the boundary of semi-technical medical vocabulary, which is limited to Schmitt and Schmitt's (2014) categories of high-, mid-, and low-frequency words and overlaps with academic words. Hsu (2013) also divides the MWL into six lexical groups. Although her attempt to specify subsets of semi-technical medical vocabulary with examples taken from the MWL is noteworthy, there is room for discussion.

Hsu (2013, p. 467) categorizes 76 academic words into Group 1, whereby she re-confirms that academic words are semi-technical. She perceives this type of semi-technical vocabulary as words “express[ing] some academic notions, approaches or procedures, and can be found across a wide range of disciplines”. Therein lies a problem that she has not explicitly articulated: whether the “academic notions, approaches or procedures” are similarly interpreted in various disciplines. Our findings from the re-examination of the MWL's headwords in the OED reveal that 42 of the 76 academic words have more than one meaning, implying that some academic words have different meanings in different disciplines. This is evident in the case of *primary* (Figure 4.4). The word *primary* has six meanings sharing the core meaning of “first and original”. In various fields, *primary* may refer to meanings *a*, *b*, *c*, and *d* while meanings *e* and *f* are usually activated in medical settings.

Primary	<i>BNC: Band 4</i>	<i>Range: 30</i>	<i>Frequency: 12,165</i>			
a) adj. First in time						
b) adj. Of the highest rank						
c) adj. Original						
d) adj. Designating a main branch of a ramifying structure						
e) adj. Designating the earliest symptoms of certain chronic infectious diseases						
f) adj. Of a neoplasm: located in the organ or tissue of origin						
Core meaning(s)	a)	b)	c)	d)	e)	f)
Core meaning 1: First and original	✓	✓	✓	✓	✓	✓

Figure 4.4 The evaluation of *Primary*

Hsu's (2013) Group 2 encompasses words in general use but expressing specialized meanings within medicine-related areas. She states, "words in this category are usually not difficult for medical students to guess their technical meaning, as the hidden technical sense is closely related to their core meaning and can be viewed as a derivative of their general meaning" (Hsu, 2013, p. 467). Hsu's (2013) example for Group 2 is *acute* (in *acute pain*), which, she reasons, can be effortlessly acquired because it derives from a core meaning common in general contexts. It is true that "sharp" and "extreme", the two distinct core meanings in Figure 4.5, may not pose a problem to learners in distinguishing a type of pain. However, if learners have previously only encountered *acute angle* and *acute accent* – two phrases derived from the core meaning "sharp" – they would not readily guess the medical core meaning of "extreme" (in *acute pain*). The semantic re-evaluation of *acute* thus far anticipates that some of Group 2's words could become troublesome due to multiple core meanings, and this issue should not be underestimated.

Acute	<i>BNC: Band 7</i>	<i>Range: 30</i>	<i>Frequency: 13,801</i>	
a) adj. Extreme				
b) adj. Accurate/ clever				
c) adj. Relating to an angle (less than 90 degrees)				
d) n. A symbol written above a letter (in some languages)				
Core meaning(s)	a)	b)	c)	d)
Core meaning 1: Extreme	✓			
Core meaning 2: Sharp		✓	✓	✓

Figure 4.5 The evaluation of *Acute*

Groups 3 and 4 are respectively identified as words “used equally with general and specialized meanings [,] . . . invisibly [slipping] out of the medical field and into other specialized fields or everyday conversation” (Hsu, 2013, p. 467) and “words with a medical dress [that] may undergo a semantic transfer when used in general language” (Hsu, 2013, p. 468). While Hsu (2013) clarifies that general and medical meanings of Group 4’s words are distantly related, e.g., *cataract* in “removing a cataract” and “cataracts of rain”, she does not mention if words in Group 3 go through a meaning shift. It is still inferred from her example of *plasma* in blood transfusion (e.g., blood plasma) and in modern appliances (e.g., plasma TV) that Group 3’s words have distinct meanings when they are outside the medical field. Our re-evaluation results (Figures 4.6 and 4.7) are aligned with Hsu’s (2013) and we postulate that these two example words are problematic because their general and specialized meanings are homographs regardless of which group they belong to.

Cataract	<i>BNC: Band 9</i>	<i>Range: 23</i>	<i>Frequency: 1,744</i>
a) n. A waterfall			
b) n. An opacity of the crystalline lens of the eye, or of the capsule of the lens, or of both, ‘producing more or less impairment of sight, but never complete blindness’			
Core meaning(s)		a)	b)
Core meaning 1: Waterfall		✓	
Core meaning 2: Lens impairment			✓

Figure 4.6 The evaluation of *Cataract*

Plasma	<i>BNC: Band 10</i>	<i>Range: 26</i>	<i>Frequency: 5,002</i>
a) n. More fully blood plasma: the clear, protein-rich liquid in which the cells of the blood are suspended. Also: the liquid component of lymph			
b) n. An ionized gas containing free electrons and positive ions [...]			
Core meaning(s)		a)	b)
Core meaning 1: Liquid in which blood cells are suspended		✓	
Core meaning 2: Ionized gas			✓

Figure 4.7 The evaluation of *Plasma*

Words in Group 5 are primarily associated with “anatomical, biochemical, demographic, epidemiological, semiological and topographical medicine” and “easily understood by the layperson” (Hsu, 2013, p. 468). Hsu (2013) illustrates by providing a derived form of the headword *secrete*, that is, *secretion*. Undoubtedly, *secretion* causes zero confusion for readers because it conveys a single meaning of “releasing substances”. However, the re-evaluation of *secrete* uncovers that it is more problematic than its derived form because it has homographs. As shown in Figure 4.8, meaning *a* has no semantic overlap with meanings *b* and

c, so the layperson is less likely to guess the medical meaning of *secrete* despite their prior knowledge of its general meanings. Words in Group 6 are, according to Hsu (2013), exclusively used in the medical register, and we believe that they are straightforward as they are single-meaning words.

Secrete	<i>BNC: Band 7</i>	<i>Range: 29</i>	<i>Frequency: 4,562</i>
a) v. To produce by means of secretion/ to perform the act of secretion			
b) v. To place in concealment, to hide out of sight, to keep secret			
c) v. To remove secretly, to appropriate (the possessions of another) in a secret manner			
Core meaning(s)	a)	b)	c)
Core meaning 1: To release	✓		
Core meaning 2: To do something out of sight		✓	✓

Figure 4.8 The evaluation of *Secrete*

By and large, Hsu’s (2013) adaptation of Baker’s (1988) and Fraser’s (2009) classification seems to contradict her viewpoint that the MWL lies along a continuum of speciality “from the vocabulary of which the technical sense is frequently an extension of the general meaning, to the vocabulary of which the technical sense is primarily used” (Hsu, 2013, p. 467). The discussion above shows that a number of MWL words can fit into multiple groups, e.g., *primary* can be a candidate for Group 1 and Group 2, and the ongoing attempt to group semi-technical medical vocabulary on the continuum seems to be unviable. This explains why the borderline between groups is not as well defined as expected.

Moreover, semi-technical medical words may be academic words (e.g., *primary*); at other times they are mid-frequency words (e.g., *acute*) or low-frequency words (e.g., *plasma*). This finding appears to support the current literature on semi-technical vocabulary (e.g., Fraser, 2007, 2009, 2012; Quero & Coxhead, 2018) which suggests that the boundary between semi-technical and other types of vocabulary is far from clear-cut. From the pedagogical perspective,

it is less important to establish either a comprehensive taxonomy of semi-technical medical vocabulary or a clear-cut boundary between semi-technical medical and other types of vocabulary. Rather, we would suggest that since semi-technical medical vocabulary flexibly stretches along what Hsu (2013) called a continuum of speciality and overlaps with other types of vocabulary, identifying troublesome words with polysemes and/or homographs, such as *primary*, *acute*, *cataract*, *plasma*, and *secrete*, is an attainable goal that has pedagogical significance.

4.4.2 Identifying semi-technical medical vocabulary with multiple meanings

The re-evaluation of the MWL identifies 302 words with polysemes and/or homographs, making them harder to learn. From now on, these 302 words are consistently referred to as “semi-technical medical vocabulary” and we propose that the semi-technical medical vocabulary identified within the scope of this study belongs to a stand-alone lexical category.

Of the 302 words, 218, approximately 72%, cover the BNC from Bands 4 to 9, implying that a fair proportion of semi-technical words are situated in Schmitt and Schmitt’s (2014) mid-frequency category. The remaining 84 (28%) are at 9,000+ levels, which according to Schmitt and Schmitt’s (2014) categorization makes them low-frequency words. The finding that semi-technical medical words are located on the BNC continuum, which provides clear starting and end points, might be more concrete than Cowan’s (1974), Flowerdew’s (1993), and Huizhong’s (1986) viewpoint that semi-technical vocabulary lies in an area between non-technical and technical vocabulary. It is easy to perceive the concepts of non-technical and technical vocabulary but much harder to separate them. It thus may be inferred that the previously researched continuum with two non-specific ends is less plausible, as is the identification of semi-technical words supposed to lie on this continuum.

The finding of 84 semi-technical medical words belonging to Schmitt and Schmitt's (2014) low-frequency category reaffirms the need to give substantial direct instruction on low-frequency words, which resonates with Nation's (2013) recommendations for vocabulary teaching. Nation (2013, p. 29) did emphasize that "teachers should teach low-frequency words only when they are essential to the understanding of the text or when they are in a relevant technical vocabulary". We agree on the fact that low-frequency words form a modest proportion of academic discourse, yet low-frequency words (in this case, semi-technical words) still need to be taught to learners to help them precisely perceive the medical meanings of these words, which are unrelated to and, consequently, hard to guess from their widely known general meanings.

The identification of polysemes and homographs of 42 academic words, in line with Wang and Nation's (2004) findings, questions the context-independent characteristic of academic words advocated by Li and Pemberton (1994). Forty-two of the academic words we examined put forward a counterargument that the meaning activation is affected by contextual relevance. For example, meaning *b* of *resolve* is activated in a medical context, meaning *f* in mathematics, meaning *g* in music, meaning *h* in chemistry, and meaning *j* in computing (Figure 4.9).

The hybrid nature of semi-technical words, a source of confusion for learners, as discussed in the literature review, results from polysemy and homography. This can be illustrated briefly through anticipated problems caused by *primary* and *resolve*. *Primary* is a polysemous word and even if medical meanings (*e* and *f*) share the same core meaning, learners still need additional contextual clues to determine which meaning is activated. *Resolve* is very challenging for learners because of having not only more meanings but also more core meanings. Learners may familiarize themselves with such generally encountered meanings as *a*, *c*, *d*, and/or *e* but it does not necessarily guarantee they will work out meaning *b*, which is in

the medical context and from a core meaning different from core meanings of *a*, *c*, *d*, and *e*. Therefore, we reiterate that our identification of 302 problematic semi-technical medical words is pedagogically significant in tailoring teaching instruction and learning strategies to clear up confusion due to polysemy and homography.

Resolve	<i>BNC: Band 4</i>	<i>Range: 30</i>	<i>Frequency: 1,606</i>								
a) v. To cause to melt or dissolve; to reduce from a solid to a liquid or fluid state											
b) v. To bring (a disease, pathological process, etc.) to resolution											
c) v. To break up or separate (a material thing) into constituent parts or elements; to disintegrate (something)											
d) v. To reduce (a subject, statement, phenomenon, etc.) by analysis into more elementary forms, principles, etc.; to consider or demonstrate (something) to be divisible or analysable into											
e) v. To convert, transform, alter, render (a material or immaterial thing) into some other thing or form											
f) v. To analyse (a force or other vector quantity) into two or more components acting in different directions but collectively having the same effect as the original vector											
g) v. To alter or transform (a discord, or relatively dissonant harmony) so as to form a concord, or relatively more consonant harmony											
h) v. To separate (a racemic compound or mixture) into optical isomers											
i) v. To translate (a readable, alphanumeric domain name) into a numerical IP address, typically by means of the domain name system											
j) v. To untie; to answer, solve; to decide, determine											
k) v. To determine or fix upon a course of action											
Core meaning(s)	a)	b)	c)	d)	e)	f)	g)	h)	i)	j)	k)

Core meaning 1: To transform	✓				✓		✓		✓		
Core meaning 2: To break into separate parts			✓	✓		✓		✓			
Core meaning 3: To bring to resolution		✓								✓	✓

Figure 4.9 The evaluation of *Resolve*

4.4.3 Disadvantages of the word form frequency-based list of semi-technical medical vocabulary

The most concerning issue in the creation of the MWL is the automatic calculation of written forms of word families, which seems unable to cope with polysemy and homography. When it comes to a polysemous word, for example, *primary* as an adjective, the statistical information about its range (30) and frequency (12,165) poses several questions: What do the range and frequency of word form occurrence indicate? Do the figures inform EMP learners of which medical meaning(s) is (are) so frequent that they should be learned intensively? Can EMP learners deduce possible meanings *primary* may convey in medical material from the range and frequency information? Very little information about the semantic properties of *primary* is de facto manifested in the statistical figures of its word form occurrence. As discussed earlier, *primary* has meanings *e* and *f* which are normally expected to be medical meanings, yet it is likely that the rest of the meanings will also be activated in medical contexts. We therefore feel learners may be confused when they are encouraged to learn *primary*, because the word is ranked in Hsu's (2013) top ten of the most frequently occurring semi-technical medical words, but learners will not be confident whether all or just some meanings of *primary* should be learned to help them deal with medical materials.

Besides polysemes, homographs, which cause deeper problems, are also left untouched. In the case of an identical word form used in different parts of speech, like *disorder* (n, v), the range and frequency resulting from the automatic word form calculation may be misleading.

The homograph of *disorder*, whose core meaning is “to give a contrary instruction”, must be calculated and presented separately from *disorder* (n, v) in the sense of “(to put) out of order” because the core meanings are different. Another example of derived form(s) of a headword having homographs is *acute*. Unlike *disorder* (n, v), *acute* (adj) and *acuity* (n) are treated as two headwords in the MWL, which questions the consistency in the MWL’s headword presentation. Hsu (2013) states that a headword is chosen to appear in the list as a representative of its family only when knowing what the headword means can guarantee understanding of the meanings of other members it represents. If so, *disorder* (n. out of order, v. to put out of order) and *disorder* (v. to countermand) should have been listed as two headwords. From these two illustrations, we observe that the automatic calculation of word forms can recognize homographs of derived words whose spellings are dissimilar to the spelling of their headword (e.g., *acute* and *acuity*) but cannot make any differentiation in the case of derived words and headwords sharing an identical spelling (e.g., *disorder*).

Disorder (n)	<i>BNC: Band 5</i>	<i>Range: 30</i>	<i>Frequency: 11,877</i>	
a) n. Absence of order; confusion				
b) n. An irregularity				
c) n. Disturbance, commotion, tumult				
d) n. Disturbance of the bodily (or mental) functions; a disease				
Core meaning(s)	a)	b)	c)	d)
Core meaning 1: Out of order	✓	✓	✓	✓
Disorder (v)				
a) v. To put out of order				
b) v. To derange the functions of; to put out of health				
c) v. To countermand				

Core meaning(s)	a)	b)	c)
Core meaning 1: To put out of order	✓	✓	
Core meaning 2: To give a contrary instruction			✓

Figure 4.10 The evaluation of *Disorder* (n, v)

Despite the usefulness of wordlists (e.g., MWL) in general, automatic word form calculation can lead to inconsistent headword presentation and problematic statistical figures, thus restricting its pedagogical potential. We highlighted from the literature review that the act of learning a word form with no comprehension of its meaning has limited value, and the finding that 302 of the MWL words (accounting for more than 50%) have multiple meanings indicates that the MWL alone is less pedagogically useful for EMP learners. Additionally, the inclusion of acronyms (e.g., GI) may puzzle learners as there is no semantic explanation. Although Hsu (2013) suggests accompanying activities allowing learners to see how MWL words, especially words with more than one meaning, are used in sentences extracted from the medical textbook corpus, the corpus is not publicly available. It is thus vital to make the MWL, a potentially ready-to-use lexical resource, more inclusive, i.e., include information about word form, meaning, and usage, instead of a list of word forms ranked according to their range and frequency of occurrence. Otherwise, unless recommended teaching/learning resources such as the medical textbook corpus are attached, learners (and teachers) will not benefit so much from wordlists, because even though teachers can create their own corpus, not many of them are willing or have the time to do so.

4.5 Potential implications of the research results for teaching semi-technical vocabulary

The core meaning-based findings (see Appendix 1), although not envisaged as a ready-to-use resource, could be used as a supplementary reference for teachers to use along with Hsu's (2013) MWL. Although the MWL can provide a long-term vocabulary goal, classroom

time is frequently limited, making the teaching of MWL's 595 frequency-ranked headwords less practical. We would therefore suggest that for a short-term vocabulary goal when writing lesson plan objectives, the core meaning list of 302 headwords could be a starting point for prioritizing words to teach, thereby making the most of limited classroom time. In addition, rather than introducing MWL headwords and leaving them for learners to self-study, teachers are recommended to select words with multiple core meanings from the core meaning list to directly teach to their learners in the classroom. Teachers could also usefully devote classroom time to delivering explicit instruction on core meanings to give learners an insight into polysemy and homography. Such an insight into the shared core meaning among polysemous words and the distinct core meanings of homographs may ease the learning of multi-meaning semi-technical medical vocabulary and help learners correctly interpret word meanings in both technical and non-technical contexts.

Since the list of 302 words is not yet in the form of a ready to use resource, teachers would need to expand on the core meanings of any word they intend to teach so that learners are provided with context-based meanings. In the case of *benign*, for example, teachers could present the core meaning to learners together with other related meanings derived from the core meaning (Figure 4.11), emphasizing that this word is polysemous. Teachers could then help students to find examples of each meaning in context by using publicly available concordancers such as Sketch Engine for Language Learning (SKELL) facility (2014-2021) (Figure 4.12). This would help to consolidate many instances of the word in its technical sense. Three instances of the word *benign*, for example, appear with their medical meaning in the first seven concordances in SKELL, and all collocate with the word *tumor*. Students could then do a more extended search for other words collocating with the technical meaning, such as *mole*, *disorder* and *condition*, and write their own sentences based on these examples, paying attention to collocations and sentence structure.

Benign, adj.

Core meaning: Mild

Meaning 1: (Of weather) pleasant

E.g. Valencia is one of the most benign climates in Europe.

Meaning 2: (Of disease) not harmful

E.g. Benign tumors are not cancer.

Figure 4.11 An example of the core meaning of *Benign* for in-class teaching

benign 3.69 hits per million

1. The 2000s was an unusually **benign** decade.
2. Too many obvious areas suffered from **benign** neglect.
3. These tumors are either **benign** or cancerous.
4. Even the **benign** shows are completely artificial.
5. She is generally considered a **benign** goddess.
6. There are many kinds of **benign** soft tissue tumors.
7. They are considered **benign** and slow growing tumors.

Figure 4.12 The first seven examples of *Benign* in SKELL (2014-2021)

4.6 Limitations and directions for future research

The current study has two limitations. First, the selection of the MWL as an input resource might exclude the examination of other high-frequency words. Second, the core meaning findings have not yet been developed as reference material. However, as mentioned earlier, previous studies have investigated high-frequency words within the medical context (Quero & Coxhead, 2018); therefore, the MWL still serves as a good resource for studies on semi-technical medical vocabulary. Additionally, the MWL presents a finite number of words on which to focus, despite its word form frequency-related issues.

A possible area of future studies would be to investigate how to develop a pedagogically helpful and ready-to-use resource of semi-technical vocabulary from existing wordlists like the MWL. The core meanings of 302 words may also play a useful role in the initial stages of

future studies. These findings should be elaborated and followed by corpus-based analyses to ensure that any newly developed resource is produced on the basis of context-based evidence. Corpus-based studies which aim to improve the MWL should consider the frequency of word meaning in addition to word form frequency. A process for calculating the frequency of word meanings should rely on human involvement; in other words, it should be semi-automatic rather than automatic, to enhance reliability and validity. Findings relating to word meaning frequency should be transferred into a teachable lexical resource in which word meanings and their interrelation are explicitly presented.

4.7 Conclusion

In response to the controversy around semi-technical words, the study took a closer look at such words in the medical field, particularly, those in the MWL. Although the selection of the MWL excluded high-frequency words from the semantic analysis, we refer to the relevant literature to substantiate our findings. Accordingly, we relocate this under-researched type of vocabulary onto the BNC continuum and assert that its hybrid nature is rooted in the phenomena of polysemy and homography. Due to this, we propose core meaning-based analysis to identify words with polysemes and/or homographs that potentially cause trouble for learners and teachers. We believe this approach is more feasible and has more pedagogical benefits than attempting to establish a clear-cut boundary between semi-technical medical and other types of vocabulary, because semi-technical medical words are heterogeneous. Moreover, it is eventually neither the clear-cut borderline around semi-technical medical words nor a comprehensive taxonomy of such words that matters. It is, in fact, semi-technical medical words themselves, especially potentially problematic ones, that need full attention from learners and teachers.

Refocusing learners' and teachers' attention back onto semi-technical medical words raises concerns over the potential difficulties of using word form frequency-based lists, e.g.,

the MWL, in learning and teaching this type of vocabulary. The difficulty is attributable to the semantic variation, through which words have multiple related and unrelated meanings (polysemes and homographs), which is overlooked in developing such lists. The automatic calculation of word form frequency only facilitates identifying commonly encountered words that are single-meaning. For multi-meaning vocabulary like semi-technical medical words, studies in which word forms are automatically calculated without taking word meanings into account lack rigor. The pedagogical uses of such studies, i.e., a list of 595 headwords with their statistical figures but with no semantic annotation, are undoubtedly limited.

In light of the research-based evidence (302 words identified with polysemes and/or homographs, equal to 51% of the total number of MWL words), we suggest that word form calculation is a necessary yet insufficient condition. This means the creation of the MWL in particular, or any wordlists in general, should consider meaning frequency together with form frequency so that outcomes will be well-rounded. We recommend that to make good use of the MWL or any other similar word form frequency-based lists, accompanying corpus resources should also be attached alongside publicly available lists. Unfortunately, it may be that learners are able to gain only restricted access to contexts where MWL words appear, due to copyright issues, and they may therefore fail to understand their meanings.

Finally, further research on the MWL is welcomed in order to address the dearth of semantic information resulting from existing word form-related issues. It is hoped that any improvement of semantic aspects in the MWL will promote the development of an even more pedagogically helpful vocabulary resource.

CHAPTER 5: THE DEVELOPMENT OF SEMIMED

Statement of Authorship

Title of paper	Developing a pilot version of SemiMed—A corpus-based resource of semi-technical medical words
Publication status	Ready to submit to Journal of English for Specific Purposes
Conference presentation	Le, C. N. N. (2023, April 18-22). <i>When lexical semantics meets lexicography : Hits and misses</i> [Paper presentation]. IATEFL Conference, Harrogate, England.

Principal Author

Name of principal author (Candidate)	Chinh Ngan Nguyen Le		
Contribution to the paper	Researched and developed conceptual framework, performed all data collection and analysis, interpreted data, developed first draft, wrote and revised manuscript, and acted as corresponding author.		
Overall percentage (%)	80%		
Certification	This paper reports on original research I conducted during the period of my Higher Degree Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis.		
Signature		Date	21/11/2023

Co-author contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of co-author	Julia Miller		
Contribution to the paper	Supervised development of work, helped in data analysis and interpretation and manuscript revisions.		
Signature		Date	21/11/2023

Developing a pilot version of semimed—A corpus-based resource of semi-technical medical words

Abstract This study presents a possible solution to pedagogic challenges of multiple-meaning semi-technical medical vocabulary that have received insufficient attention from current wordlists. The target of the investigation was Hsu’s (2013) Medical Word List (MWL). The list specifies a set of frequently encountered words that is worth learning. However, it is based solely on word form frequency; word meaning frequency is not addressed. This study used mixed methods to propose a remedy for this semantic deficiency. First, 40 MWL words were consulted in the *Oxford English Dictionary*. Next, their definitions underwent a core meaning-based analysis (Le & Miller, 2023) to identify the relationships between them. Then, Cantos and Sanchez’s (2001) Lexical Constellation (LC) model was used to visualize the relationships between word meanings. Lastly, a follow-up corpus-based analysis was conducted to validate the qualitatively established LCs. The final 40 LCs provided the pilot version of a new resource named SemiMed.

Keywords: English for Medical Purposes; wordlists; semi-technical vocabulary; word form frequency; word meaning frequency

5.1 Introduction

In English for Medical Purposes (EMP), there is a type of vocabulary known as semi-technical vocabulary that is challenging to learn and teach (Le & Miller, 2023; Li & Pemberton, 1994; Shaw, 1991; Thurston & Candlin, 1998). These challenges arise because semi-technical vocabulary usually has more than one meaning and is interpreted differently depending on context. For example, the meaning of *orbit* as “the eye socket” mostly appears in the medical context, while its other two meanings, “the path something in space follows round something bigger” and “to follow a path in space round something bigger”, are often used in the general context and are only distantly related to “the eye socket”.

Conventional resources do not seem to address this multi-meaning phenomenon comprehensively. Dictionaries (usually general ones), which may be the first option for learners (and teachers), can create confusion, as learners are unlikely to locate a correct sub-entry when they look up a headword with multiple meanings (Nesi & Haill, 2002; Winkler, 2001). More specifically, they tend to refer to the first sub-entry and neglect the others (Boonmoh et al., 2006). Discipline-specific meanings, such as those used in semi-technical medical vocabulary, however, tend to be more unusual (Nesi & Haill, 2002), and thus do not always appear in the first sub-entry. A learner consulting a dictionary could thus fail to locate medical meanings of a semi-technical medical word.

Wordlists such as Hsu's (2013) Medical Word List (MWL) are considered an alternative pathway to learning and teaching semi-technical medical vocabulary. The MWL is useful because the number of semi-technical medical words is limited to a manageable size (595 words). The idea of narrowing the focus of learning and teaching down to a definite set of more frequently used semi-technical medical words is worth acknowledging. However, the absence of word meanings offered by the MWL, due to its word-form-frequency-based development, may restrict its usefulness in resolving the multi-meaning-related challenges.

This study considers the multi-meaning issues raised by learning and teaching semi-technical medical vocabulary which are not adequately addressed in resources such as dictionaries and wordlists. It uses the MWL as a starting point for the development of a new resource that minimizes the confusion caused by dictionary sub-entries and compensates for the lack of word meanings in current wordlists.

5.2 Literature review

5.2.1 Wordlists

A wordlist is defined as a list of all the different words in a text or corpus. It provides information about the number of times each word occurs and is arranged either alphabetically

or in order of frequency (Hunston, 2005, Lüdeling & Kytö, 2008). Studies on wordlists have thrived since the advent of a prevalent assumption stressing the critical role of frequency information in vocabulary acquisition. It is thought that frequently used words should receive more learning and teaching time because they will be more useful to learners (Gardner & Schmitt, 2015).

Following this insight, the creation of wordlists has tended to focus on identifying a set of target words that frequently occur in particular contexts. A widely adopted approach to creating such wordlists “relies on empirical evidence from language use (corpora) to select and/or rank words based on frequency and other quantitative criteria” (Brezina & Gablasova, 2017b, p. 765). Under this approach, Miller and Biber (2015, p. 31) describe a step-by-step methodological procedure as follows:

... (i) design and construct a representative corpus; (ii) identify the full set of [words] found in that corpus; (iii) analyze the distributions (frequency and range) for each [word]; and (iv) select the [words] with the highest frequencies and widest dispersions in the corpus ...

Although frequency is considered an important criterion for generating wordlists, Paquot (2007, p. 127) states that “it is only half of the story”. Wordlists created on the basis of frequency have an undeniable pedagogical potential, which is to set an explicit, attainable goal for vocabulary learning and teaching. In other words, these wordlists help inform teachers how many words learners should know (i.e., the breadth of their vocabulary knowledge). Nevertheless, it has been argued that “vocabulary learning is not simply remembering a list of words but rather a complex process” (Yu & Trainin, 2022, p. 235). Yamamoto (2014) states that there is another equally important aspect of vocabulary knowledge, that is, the depth of vocabulary knowledge, which should be considered simultaneously with the breadth of vocabulary knowledge. In wordlists, the depth of vocabulary knowledge (e.g., word meanings) is, however, only superficially treated, because wordlists showcase very little information

about the different meanings of a word. Wordlists may thus “provide targets for eventual achievement, but say nothing about how those targets are to be reached” (Todd, 2017, p. 32).

The inadequacy of the depth of vocabulary knowledge in wordlists may stem from the operationalization of the word construct. In the development of corpus-based wordlists, word forms are usually selected as a unit of counting in frequency-based analyses. Gardner (2007), however, highlights the fact that a word may have more than one meaning and that this is a problem that should be fully addressed. Todd (2017, p. 32) also criticizes “the use of surface forms as the basis for distinguishing between words” because he casts doubt on the validity of word form frequency-based results for multi-meaning words. He reasons that some words may have different meanings in different contexts, so wordlists that rely heavily on the written form of the word may fail to distinguish meanings sharing the same word form.

The phenomenon of multiple meanings has led to growing concern because words with high frequencies usually have more than one meaning (Todd, 2017). This means there is a likelihood that frequency wordlists may contain a number of multi-meaning words. Furthermore, the fact that a word may have multiple meanings may make it harder to learn (Laufer, 1997, as cited in Fraser, 2012). Laufer reasons that learners have a tendency to erroneously rely on a word meaning that they already know and persist with that meaning regardless of the different contexts in which a word appears. Hence, multiple meanings may cause significant difficulties for learners and this difficulty should not be underestimated in word form frequency-based wordlists.

5.2.2 Word sense disambiguation

To fully address the phenomenon of multiple meanings in word form frequency-based wordlists, extensive work is necessary to “[disambiguate] a word that can have many senses based on its usage context [e.g., a corpus]” (Vidhu Bhala & Abirami, 2012, p. 159). This kind of work is situated in the area of word sense disambiguation (WSD), which has arisen since

computers have been involved in building solutions for human language problems. The scope of WSD can be vast, but within this paper, we focus on polysemy and homography. Polysemes and homographs are two types of identically spelled words with related and unrelated multiple meanings, respectively.

In word form frequency-based wordlists, disambiguating meanings of polysemes and homographs has been a perennial problem because, as mentioned earlier (Grabe, 1991, p. 392),

each word form is counted as a single word, though in reality, each word form may represent a number of distinct meanings, some of which depend strongly on the reading context, and some of which are quite different from each other in meaning.

This is exemplified in the case of *bear*. Gardner (2007) notes that when appearing in different parts of speech, the word *bear* can be perceived as two homographs (*bear* as a noun meaning “an animal” and *bear* as a verb meaning “to carry”). *To bear* is also a polyseme having 13 meanings across contexts. Two meanings listed in Gardner’s (2007, p. 251) work include “to move while holding up and supporting” and “to hold in the mind”. Gardner (2007, p. 253) anticipates that machine-based frequency counts of the written form of *bear* link all of its meanings together and thus incur some risks:

(a) they will overestimate the true coverage of the word forms; (b) they will underestimate the actual user knowledge required to negotiate the word forms; and/or (c) they will underestimate the actual number of meanings inherent in the word forms.

The fact that one or more of these risks may exist in word form frequency-based lists raises questions about their usefulness, especially when teaching is based on word form frequency without taking account of multiple meanings (Gardner, 2007). From a practical standpoint, Biemiller and Slonim (2001, as cited in Gardner, 2007, p. 252) state that “general print frequency of word forms is a poor predictor of learners’ root word knowledge”. Moreover, since Ravin and Leacock (2000, p. 1) remind us that “the most commonly used words tend to be the most polysemous”, the construct of the word in corpus-based lists of frequently used words might need to be re-operationalized.

From this premise, Knowles and Mohd Don (2004) suggest that individual word meanings be considered as the basis for frequency-based analyses. Biemiller and Slonim (2001, p. 510) advocate this by further reasoning that “frequencies of word meanings rather than word forms might lead to better predictions of learners’ root word knowledge”. Although Gardner (2007, p. 253) comments that “such frequencies would be very hard to produce”, the task may not be impossible. There have been several approaches to WSD that are believed to facilitate the count of word meaning frequency. For reasons of space, we only elaborate on the one sense per collocation method (Yarowsky, 1993) within the scope of this paper.

The method is named after the one sense per collocation heuristic, which was first introduced by Yarowsky (1993). This heuristic places focus on word collocations, since “nearby words provide strong and consistent clues to the sense of a target word” (Mihalcea, 2007, p. 124). In regards to polysemy, Yarowsky’s (1993) one sense per collocation hypothesis resonates closely with Hoey’s (2005, p. 13) Lexical Priming, which postulates that “when a word is polysemous, the collocations ... of one sense of the word differ from those of its other senses”.

The underlying mechanism of this hypothesis used to disambiguate polysemous meanings is demonstrated in Cantos et al.’s (2009, p. 79) diagram (see Figure 5.1), where three meanings of a polysemous word are discriminated by their sets of collocations:

... assume we have a polysemous word *w* with three different meanings *m*₁, *m*₂ and *m*₃. If we take for granted that each actual sense of a word is lexically codified in the forms of its syntagmatic environment, we find that each meaning (*m*₁, *m*₂ and *m*₃) has accordingly a number of associated collocates. That is, for meaning 1 (*m*₁) we find two collocates (*m*-*c*₁ and *m*-*c*₂); for meaning 2 (*m*₂), three collocates (*m*₂-*c*₁, *m*₂-*c*₂, and *m*₂-*c*₃); and for meaning 3 (*m*₃), seven collocates (*m*₃-*c*₁, *m*₃-*c*₂, *m*₃-*c*₃, *m*₃-*c*₄... *m*₃-*c*₇).

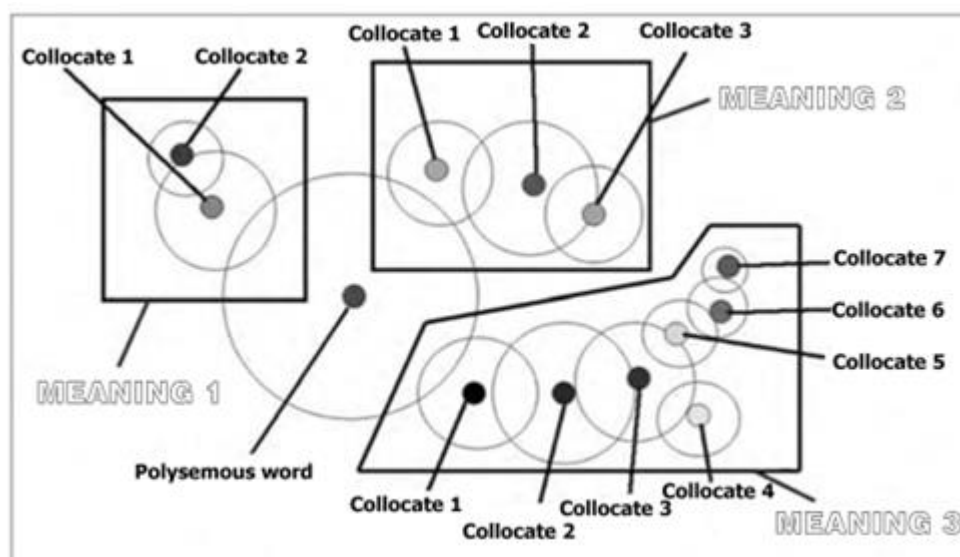


Figure 5.1 Cantos et al.'s (2009) illustration of word meaning disambiguation based on collocations

It is widely agreed by corpus linguists that “the meaning of a word is dependent on the other words associated with it in a particular text (cotext), and that words are only ambiguous when isolated from their cotext” (Sinclair, 2004, as cited in Gardner, 2007, p. 252). Conceivably, putting multi-meaning words in wordlists back into a particular context (a corpus) and investigating their collocations may provide valuable disambiguating clues. Hence, the one sense per collocation method is promising and may be achievable as corpus analysis software now allows the automatic export of collocational data.

5.2.3 Lexical constellations

According to Widdowson (2003, p. 115), pedagogically useful teaching and learning materials “[have] to be specified along two parameters: in terms of the objectives to be eventually achieved, and in terms of the process that has to be activated to get there”. Wordlists, as discussed above, only fulfil the former parameter because their current formats appear not to provide a venue for elaborating meaning interrelations, especially in the case of multi-meaning words.

Todd (2017) describes this type of words as opaque. These are high-frequency words that take on technical (or unusual) meanings in discipline-specific contexts. Semi-technical vocabulary is considered to belong to this category. These words, if they are in word form frequency-based lists, may cause problems for learners (Todd, 2017), as the lists contain very few details of the context-dependent meanings of a word. Additionally, even though learners could opt for dictionaries to look up multi-meaning words presented in wordlists, relying on dictionaries may result in comprehension problems because the unusual meanings learners look for are not always the first meanings given and students frequently do not read past the first sub-entry (Boonmoh et al., 2006; Nesi & Hail, 2002; Winkler, 2001).

To maximize the pedagogical usefulness of frequency wordlists, greater effort should be devoted to achieving the second parameter—facilitating the learning process. A possible direction is “to bring more pedagogical value to corpus-based research on wordlists” (Dang 2019, p. 300) by fleshing out (or even altering) the default setting of wordlists (i.e., the vertical presentation of headwords with their frequency and range statistics in parallel columns).

One such approach is Cantos and Sanchez’s (2001) Lexical Constellation (LC) model. The term LC, which is often used in astronomy, is interpreted within the field of lexical semantics, according to Cantos and Sanchez (2001), as a visualized network of word meanings.

Underpinning the development of the LC model is the notion that word meanings interact in a multidimensional rather than a linear way. Consequently, there is a need to visualize complex semantic connectivity within a multi-meaning word to better understand it across various contexts. In this sense, Cantos and Sanchez (2001, p. 109) perceive a word as “a hierarchical structure whereby each element [each meaning] is directly or indirectly dependent on other elements [meanings]”, which is likened to a constellation of stars.

The generic LC model has a core meaning (meaning C) placed at the centre and surrounded by multiple, related meanings (meanings D and E, etc.) located in outer layers,

which showcase the degree of interconnectivity (see Figure 5.2). The distance and proximity between meanings indicate how close their relation is.

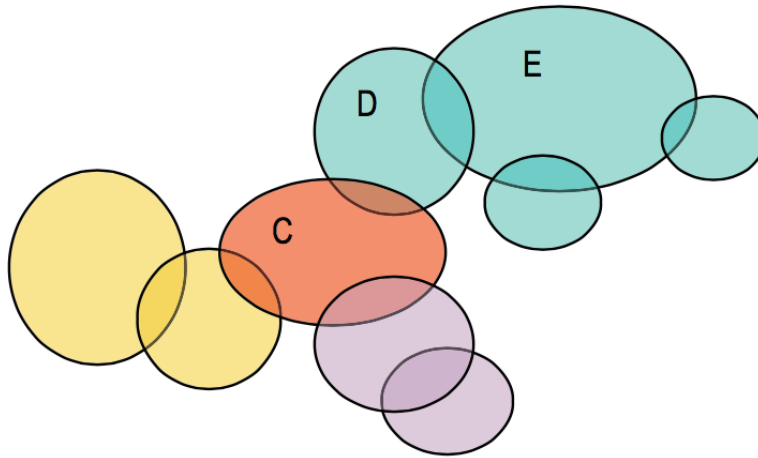


Figure 5.2 Generic pattern of an LC (Adapted from Rizzo & Sanchez, 2010, p. 110)

As an illustration, Figure 5.3 presents an LC of *heart* in which the three meanings in the first layer (central part; thoughts, emotions, feelings; shape of a heart) are more directly related to the core meaning (physical organ in persons/animals) than the meanings on the second layer (core/centre/essence; lover, devotion, sympathy; courage; card with figure of heart). Additionally, the intersection between meaning clusters in the outer layers indicates their inter-connection. For example, “central part of anything” closely links with “core/centre/essence” as they intersect with each other.

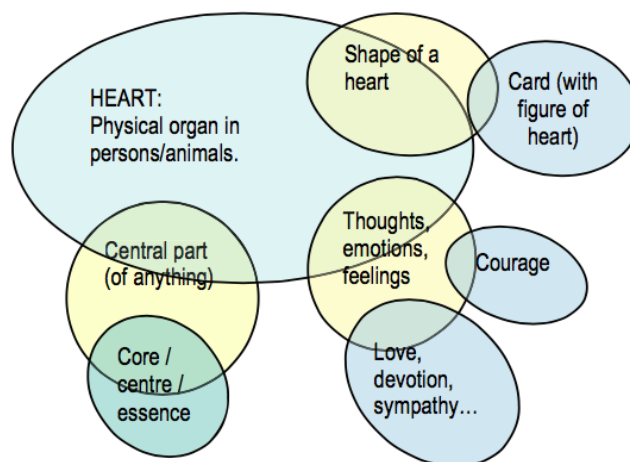


Figure 5.3 An LC of *Heart* (Adapted from Rizzo & Sanchez, 2010, p. 112)

Cantos and Sanchez (2001) claim that their model is capable of describing the intricacies of meanings that have evolved throughout the word’s history (i.e., how (new) meanings are rooted in some core semantic properties and take additional semantic features over time regardless of their unchanged written form) which is seen as “a permanent source of ambiguity, hence of possible misunderstanding” (Rizzo & Sanchez, 2010, p. 110). Thus, Cantos and Sanchez’s (2001) LC model has, according to Perez (2013), not only the advantage of analysing semantic complexity, but also the potential to explain semi-technical words (alternatively called sub-technical words in Perez’s (2013) work) through the clear visualization of interconnectivity existing among usual (generally used) and unusual (discipline-specific) meanings of these words.

5.3 The study

In response to the concern raised in the literature review about the inadequacy of the depth of vocabulary knowledge promoted by word form-based wordlists, this study suggests semantic improvements to one particular wordlist: Hsu’s (2013) 595-word Medical Word List (MWL) (Figure 5.4).

MWL	Range	Frequency	BNC
PRESENCE	30	7,635	BNC4
RECUR	30	7,607	BNC6
INFLAME	30	7,554	BNC5
BIOPSY	29	7,487	BNC8
PULMONARY	30	7,301	BNC11
PATHOLOGY	30	6,949	BNC7
VAGINA	24	6,941	BNC8
INHIBIT	29	6,840	BNC5
MUTATE	27	6,780	BNC9
ANAESTHESIA	28	6,662	BNC12
HEPATIC	24	6,660	BNC12
MALIGN	30	6,556	BNC6
ABDOMEN	29	6,507	BNC9
DEFECT	30	6,496	BNC5

Figure 5.4 An excerpt from Hsu’s (2013) Medical Word List

The MWL was selected because, first, it characterizes features of a list developed by calculating word forms in a corpus (see more details about the MWL’s development in Table

5.1). Secondly, Le and Miller’s (2023) evaluation of the list revealed that of the 595 words, just over half (302 words) have multiple meanings (polysemes, homographs or both). The reliance solely on word frequency as a basis for list compilation indicates an urgent need to conduct an extensive study to enrich semantic information absent in the MWL.

Table 5.1 Details of the development of Hsu’s (2013) Medical Word List

A representative corpus		Word selection criteria	
<i>Name</i>	Medical Textbook Corpus	<i>Specialized occurrence</i>	Outside the first BNC 3,000 words
<i>Size</i>	15 million words	<i>Range</i>	Occur in more than half of 31 medical subject areas
<i>Source</i>	155 medical textbooks across 31 medical subject areas	<i>Frequency</i>	Occur at least 863 times in the Medical Textbook Corpus

The methodological approach taken in this study is a mixed method based on the theories mentioned in the literature review. Initially, Cantos and Sanchez’s (2001) LC model shed light on a qualitative analysis in which different meanings of a word were visualized in a learnable manner. Then, a quantitative analysis was conducted with a focus on word meaning frequency to substantiate and validate results from the qualitative analysis. The step-by-step methodological procedure is illustrated via the case of the word *defect*.

5.3.1 Qualitative analysis

Step 1. Look up each word in the *Oxford English Dictionary* (OED)

The word was looked up in the OED because the MWL provides no semantic information other than range and frequency statistics (Figure 5.4). As in Le and Miller’s (2023) examination of *defect* in the OED, the word was looked up and then the results were refined

by removing obsolete and merging overlapping meanings to prepare the word for further analysis.

Defect	<i>BNC: Band 5</i>	<i>Range: 30</i>	<i>Frequency: 6,496</i>
n. Lack or absence of something necessary or desirable; a deficiency, a want. Also: the state or fact of being deficient or falling short.			
n. An imperfection in a person or thing; a shortcoming, a failing; a fault, flaw, or abnormality.			
v. To abandon or desert a person, party, organization, or cause, esp. in favour of an opposing one.			

Figure 5.5 OED definitions of *Defect* used in Le and Miller (2023)

Step 2. Simplify OED definitions

Next, the original definitions from the OED were simplified to ensure all learners of English, particularly those at lower levels, could fully understand individual definitions. Moreover, using simply reworded OED definitions optimizes the space of text bubbles in LCs and avoids copyright infringements. Figure 5.6 shows OED definitions of *defect* after the simplification process.

Defect	<i>BNC: Band 5</i>	<i>Range: 30</i>	<i>Frequency: 6,496</i>
a) n. A condition where there is not enough of something			
b) n. An imperfection in a person or thing			
c) v. To leave (and join the other side)			

Figure 5.6 Simplified OED definitions of *Defect*

Step 3. Group simplified definitions under core meanings

Once the OED definitions were simplified, they were classified into polysemes and homographs. Polysemes tend to share a mutual core meaning, while homographs do not (Le & Miller, 2023).

Defect	<i>BNC: Band 5</i>	<i>Range: 30</i>	<i>Frequency: 6,496</i>
a) n. A condition where there is not enough of something			

b) n. An imperfection in a person or thing			
c) v. To leave (and join the other side)			
Core meaning(s)	a)	b)	c)
Core meaning 1: A lack	✓	✓	
Core meaning 2: To leave (and join the other side)			✓

Figure 5.7 Core and other related meanings of *Defect* used by Le and Miller (2023)

Following this observation, Le and Miller (2023, p. 258) carried out a core meaning-based analysis consisting of three steps: “(1) read through all listed dictionary definitions of a word, (2) identify (how many) core meanings a word [has] and write them down, (3) point out dictionary definitions that share the same core meaning”. Figure 5.7 showcases Le and Miller’s (2023) core meaning-based analysis of *defect*. This result was used as an input for the next step involving the visualization of the meanings of *defect* in LCs.

Step 4. Visualize core and other related meanings in LCs

As can be seen in Figure 5.7, *defect* had two distinct core meanings, originally stated in the work of Le and Miller (2023) as “relating to a deficiency” and “to leave”. However, the former was reworded to ensure that the defining words were simpler than the word *defect* itself and more information was added to the latter to make it more specific:

<p>Defect</p> <p>Core meaning 1: A lack</p> <p>a) (n) A condition where there is not enough of something</p> <p>b) (n) An imperfection in a person or thing</p> <p>Core meaning 2: c) (v) To leave (and join the other side)</p>
--

Figure 5.8 Description of *Defect* resulting from the qualitative analysis

After arriving at the meanings of *defect* (Figure 5.8), the process of visualizing the meanings in LCs was carried out. Each LC represents a (polysemous) word and if a word has homographs, its homographs are presented in separate LCs. *Defect* has two polysemous meanings derived from core meaning 1 and one homograph representing core meaning 2. The word thus has two LCs, which are featured in Figure 5.9.

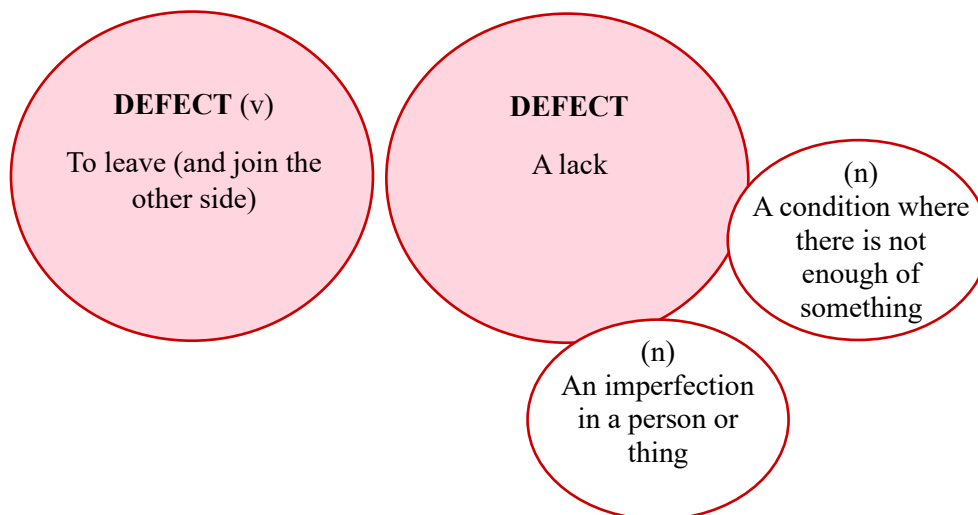


Figure 5.9 LCs of *Defect* resulting from the qualitative analysis

The first LC (on the right) illustrates a polysemous word with core meaning 1 placed at the centre surrounded by meanings *a* and *b*. The remaining LC (on the left) is of a homograph with core meaning 2 placed at the centre; no other related meanings were derived.

5.3.2 Quantitative analysis

LCs produced from the qualitative analysis were then validated through a corpus-based analysis. At this stage, frequencies of word meanings visualized in qualitatively established LCs were examined in general and specialized corpora to (a) assess whether each particular meaning was frequent enough to keep it in the LCs and, if yes, (b) specify in which context(s) the meaning is more likely to appear.

The analysis tool used in the quantitative analysis was Sketch Engine (Word Sketch <https://ske.li/ufw>). Two corpora, English Web 2020 and Medical Web Corpus (Table 5.2),

freely accessible for Sketch Engine subscribers, were selected to represent general and specialized corpora. The one sense per collocation method was applied to disambiguate multi-meaning words examined in the two corpora. The unit of analysis was word meaning frequency and the unit of counting was collocate frequency.

Table 5.2 Description of English Web 2020 and Medical Web Corpus

	Tokens	Words	Sentences	Documents
<i>English Web 2020</i>	43,125,207,462	36,561,273,153	2,008,143,278	78,373,887
<i>Medical Web Corpus</i>	42,054,011	33,961,786	1,545,862	526

Step 5. Examine top 15 collocates in English Web 2020 and the Medical Web Corpus

Word Sketch in Sketch Engine exported a list of collocates, from which the top 15 most frequent collocates were identified. After multiple trials, a cut-off line of fifteen was found to be an optimal window to retrieve collocates with high frequency of occurrence and practical significance in disambiguating their node words.

Fifteen collocates were selected on the basis of their frequency and typicality scores, both of which were automatically computed by Word Sketch. The primary selection criterion was frequency. Next, typicality was considered to check whether frequent collocates are strong collocates (i.e., ones that do not often co-occur with many other words). Typicality ensures that frequent collocates are typical of particular meanings. Knowing the typicality score facilitates the follow-up process of assigning meanings to collocates.

After the identification of the top 15 frequent collocates, meanings were assigned to individual collocates. Tables 5.3 – 5.5 indicate the results of the examination of the top 15 collocates of *defect* in English Web 2020 and Medical Web Corpus (a search of *defect* as a verb in the Medical Web Corpus returned no results):

Table 5.3 Top 15 most frequent collocates and meanings of *Defect* (n) in English Web 2020

Collocate	Frequency	Typicality	Meaning of <i>defect</i> in relation to collocate
1. birth	58,021	10.7	n. Something wrong with part of the body
2. heart	14,079	7.9	n. Something wrong with part of the body
3. congenital	12,062	9.6	n. Something wrong with part of the body
4. genetic	10,309	7.8	n. Something wrong with part of the body
5. neural tube	6,533	8.2	n. Something wrong with part of the body
6. manufacturing	6,359	7.2	n. Something that is not perfect
7. structural	4,923	7.0	n. Something that is not perfect
8. construction	3,058	5.7	n. Something that is not perfect
9. design	2,612	4.8	n. Something that is not perfect
10. physical	2,605	4.3	n. Something wrong with part of the body
11. visual field	2,556	5.4	n. Something wrong with part of the body
12. product	2,497	4.7	n. Something that is not perfect
13. ventricular septal	2,449	7.4	n. Something wrong with part of the body
14. atrial septal	2,176	7.2	n. Something wrong with part of the body
15. surface	2,144	5.5	n. Something that is not perfect

Table 5.4 Top 15 most frequent collocates and meanings of *Defect* (n) in Medical Web Corpus

Collocate	Frequency	Typicality	Meaning of <i>defect</i> in relation to collocate
1. birth	275	9.8	n. Something wrong with part of the body
2. congenital	123	9.4	n. Something wrong with part of the body
3. genetic	85	8.6	n. Something wrong with part of the body
4. heart	75	6.3	n. Something wrong with part of the body
5. neural tube	72	9.4	n. Something wrong with part of the body

6. ventricular septal	56	9.3	n. Something wrong with part of the body
7. valvular	37	8.3	n. Something wrong with part of the body
8. visual field	36	7.6	n. Something wrong with part of the body
9. cardiac	32	6.5	n. Something wrong with part of the body
10. mental	31	6.0	n. Something wrong with part of the body
11. physical	27	5.7	n. Something wrong with part of the body
12. afferent pupillary	20	7.7	n. Something wrong with part of the body
13. perfusion	20	7.4	n. Something wrong with part of the body
14. speech	18	6.6	n. Something wrong with part of the body
15. atrial septal	17	7.2	n. Something wrong with part of the body

Table 5.5 Top 15 most frequent collocates and meanings of *Defect* (v) in English Web 2020

Collocate	Frequency	Typicality	Meaning of <i>defect</i> in relation to collocate
1. soldier	484	4.3	v. To leave (and join the other side)
2. officer	387	3.2	v. To leave (and join the other side)
3. member	365	1.8	v. To leave (and join the other side)
4. pilot	225	4.3	v. To leave (and join the other side)
5. player	184	1.1	v. To leave (and join the other side)
6. official	159	1.4	v. To leave (and join the other side)
7. customer	158	1.6	v. To leave (and join the other side)
8. voter	155	3.4	v. To leave (and join the other side)
9. general	151	4.3	v. To leave (and join the other side)
10. army	145	3.2	v. To leave (and join the other side)
11. troop	136	3.1	v. To leave (and join the other side)
12. councillor	134	5.5	v. To leave (and join the other side)

13. leader	111	1.2	v. To leave (and join the other side)
14. agent	108	2.2	v. To leave (and join the other side)
15. commander	107	4.6	v. To leave (and join the other side)

Step 6. Rate meanings on a 4-level technicality scale

The examination of the top 15 most frequent meanings resulted in five possibilities.

- Possibility 1: Found in the top 15 meanings in English Web 2020, but not found in the Medical Web Corpus
- Possibility 2: Found in the top 15 meanings in English Web 2020, and outside the top 15 meanings in the Medical Web Corpus
- Possibility 3: Found in the top 15 meanings in both English Web 2020 and the Medical Web Corpus
- Possibility 4: Found in the top 15 meanings in the Medical Web Corpus, and outside the top 15 meanings in English Web 2020
- Possibility 5: Found in the top 15 meanings in the Medical Web Corpus, but not found in English Web 2020

A technicality scale (Table 5.6) was devised from these five possibilities to specify the degree of technicality of a particular meaning. The scale has four levels ranked in ascending order of technicality (from Level 0: purely general to Level 3: highly technical).

Table 5.6 Descriptors of four technicality levels

Level	Descriptor	Possibility
Level 0	<i>This meaning is solely used in general contexts</i>	Possibility 1
Level 1	<i>This is the generally used meaning</i>	Possibility 2
Level 2	<i>This meaning is used in both general and medical contexts</i>	Possibility 3
Level 3	<i>This meaning is used only in medical contexts</i>	Possibility 4 or Possibility 5

Note: Level 0 is not indicated in LCs.

As can be seen from Tables 5.3, 5.4, and 5.5, *defect* has two purely general meanings (Level 0): “something that is not perfect” and “to leave (and join the other side)”, and one meaning used in both general and medical contexts (Level 2): “something wrong with part of the body”.

Step 7. Modify LCs resulting from the qualitative analysis (if necessary) and add the technicality level to the LCs

The technicality rating of the meanings of *defect* was incorporated into the word’s quantitative analysis results. The analysis of *defect* in the two corpora revealed that meaning *a* was not found in either English Web 2020 or the Medical Web Corpus, leading to its removal from the final LCs. Meaning *b* was broken down into two sub-meanings indicated by distinctive sets of collocates and ranked at two different levels of technicality (Level 0: Something that is not perfect, Level 2: Something wrong with part of the body). Below is the finalized description of *defect*:

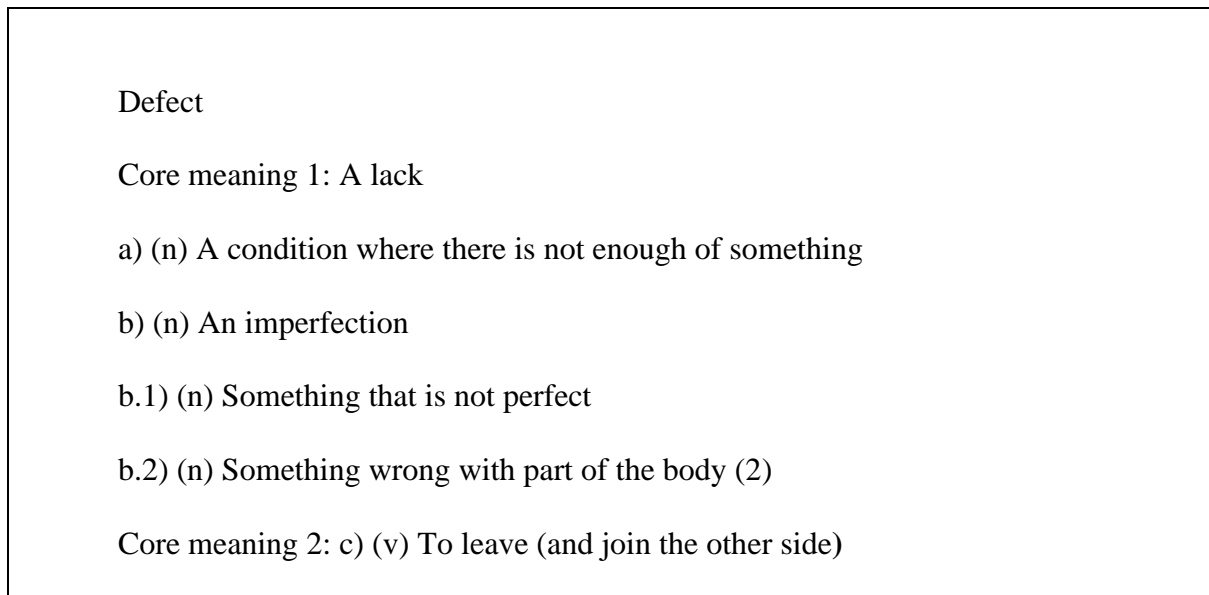


Figure 5.10 Description of *Defect* resulting from the quantitative analysis

Accordingly, the qualitatively established LCs were modified to exclude meaning *a* and form a cluster of sub-meanings *b.1* and *b.2* (Figure 5.11). Numbers (1, 2, and 3) were placed at the bottom of each LC text bubble to indicate the technicality levels of each meaning. In the case of *defect*, only one sub-meaning had a technicality level.

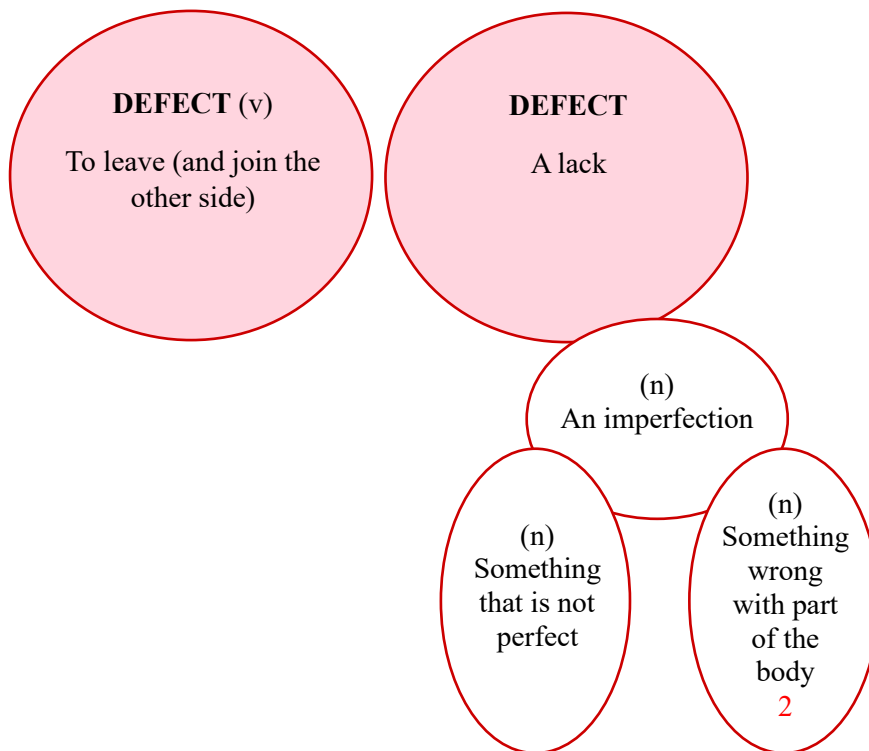


Figure 5.11 LCs of *Defect* resulting from the quantitative analysis

5.4 Findings and discussion

5.4.1 Selection of words from Hsu's (2013) MWL to create a pilot version of SemiMed

The main focus of the study was 302 words with polysemes and/or homographs. These words, according to Le and Miller (2023), are more challenging to learn than the rest of the words in the MWL and thus deserve intensive investigation. Of these 302, 40 words (approximately 13%) were selected to create pilot LCs (Table 5.7).

We ensured that the 40 words covered a full range of word types to represent a good sample of the 302 words. They include

- words with different parts of speech:
 - nouns
 - verbs
 - adjectives
 - adverbs
- words with
 - homographs (single-meaning words)
 - polysemes (multi-meaning words with one core meaning)
 - polysemes and homographs (multi-meaning words with more than one core meaning)

Detailed descriptions of these 40 words after the qualitative and quantitative analyses are available in Appendix 2.

Table 5.7 Forty sampled words used to develop SemiMed

Single-meaning words	Multi-meaning words with single core meaning	Multi-meaning words with more than one core meaning
Acute	Absorb	Arch
Cardiac	Benign	Diffuse
Cataract	Compound	Defect
Chronic	Conduct	Moderate
Colon	Circulate	Orbit
Disorder	Degenerate	Peel
Induce	Fascia	Radical
Intern	Inferior	Reflex
Liver	Lobe	Resolve
Palsy	Migrate	Stem
Secrete	Parallel	
Stool	Predispose	
Tumor	Primary	
	Prior	
	Radiate	
	Sedate	
	Shunt	

5.4.2 SemiMed template

A generic template of the three types of words (single-meaning words, multi-meaning words with one core meaning and multi-meaning words with more than one core meaning) was designed with key elements:

- Headword
- Core meaning
 - Homograph: Meaning of a homograph
 - Polysemous word: Shared meaning from which other polysemous meanings are derived
- PoS: Part of speech
- Meaning cluster
 - Meaning 1, 2: Polysemous meanings
 - Meaning 2.1, 2.2: Sub-meanings of meaning 2
- Technicality level
 - Level 1: This is the generally used meaning
 - Level 2: This meaning is used in both general and medical contexts
 - Level 3: This meaning is used only in medical contexts

Generic LC designs accompanied by specific examples are shown in Figures 5.12 – 5.17.

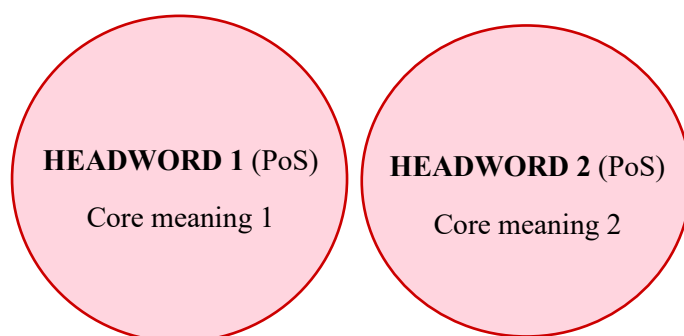


Figure 5.12 A generic LC for single-meaning words

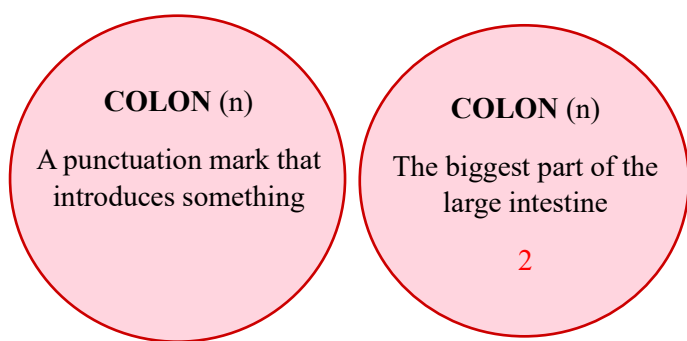


Figure 5.13 An example LC of a single-meaning word

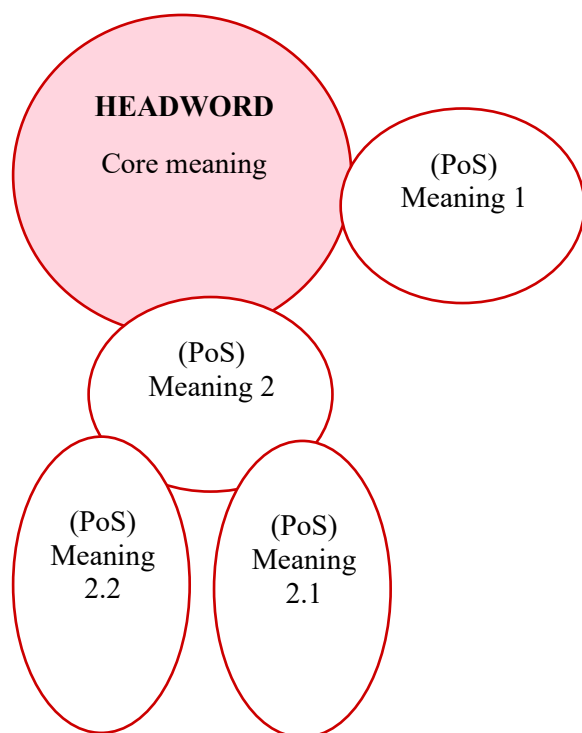


Figure 5.14 A generic LC for multi-meaning words with a single core meaning

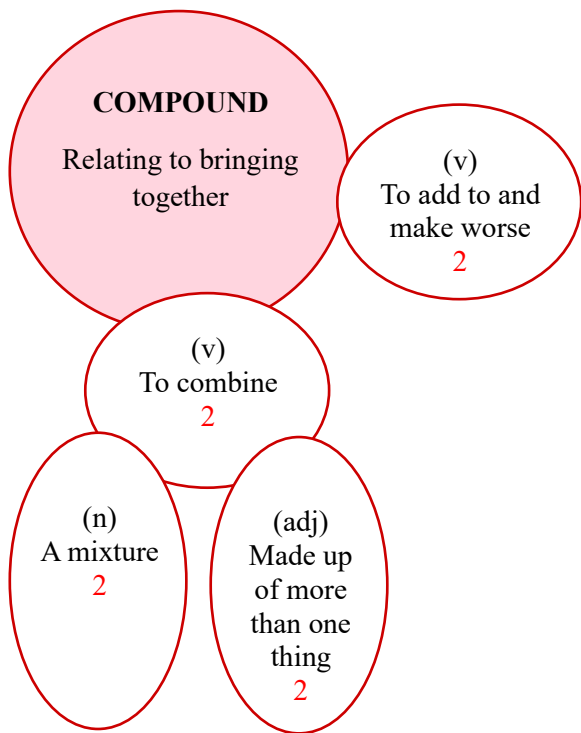


Figure 5.15 An example LC of a multi-meaning word with a single core meaning

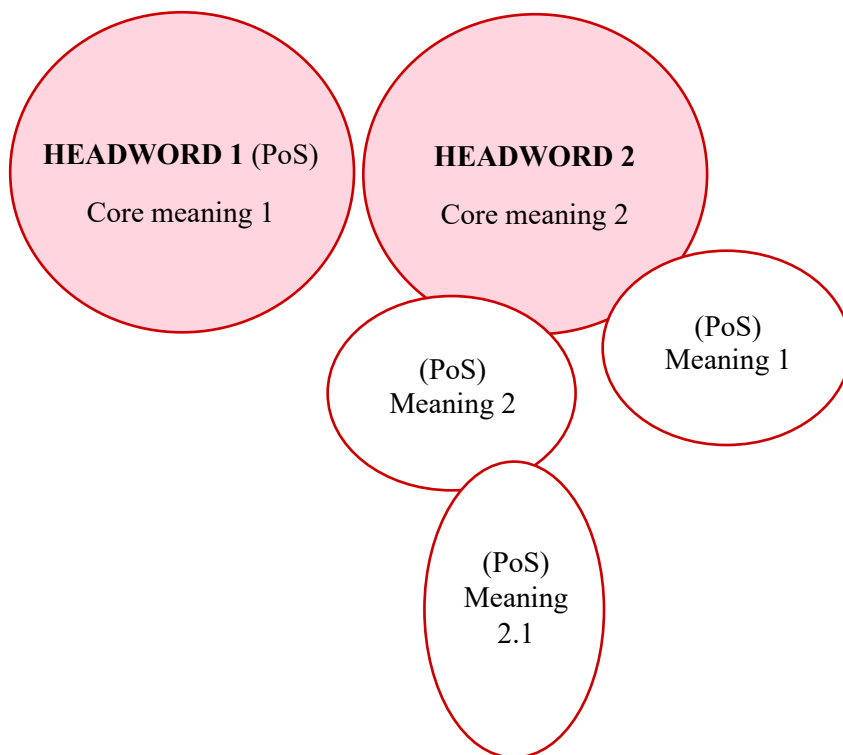


Figure 5.16 A generic LC for multi-meaning words with more than one core meaning

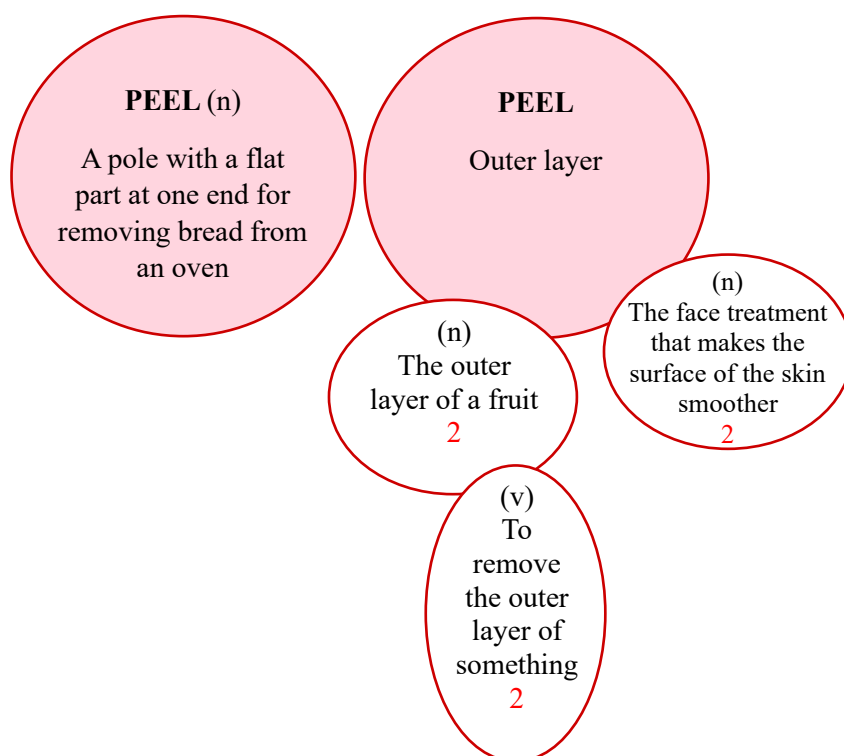


Figure 5.17 An example LC of a multi-meaning word with more than one core meaning

5.4.3 Pedagogical potential of SemiMed

The design of SemiMed, which visualizes semantic relations existing in the MWL’s multi-meaning words, has potential to optimize Widdowson’s (2003) second parameter—SemiMed’s LCs provide an extended version of Hsu’s MWL that improves the pedagogical usefulness of this resource from simply informing learners and teachers of how many words to focus on to guiding how these words can be learned and taught.

First, SemiMed increases the semantic input in MWL by detailing the meanings of words in addition to their range and frequency. Word meanings are not vertically listed as in conventional lexical resources (e.g., dictionaries). Rather, they are placed in a radial structure that permits an explicit illustration of how generally used meanings interact with discipline-specific ones.

For example, *primary* illustrated in Figure 5.18 has five meanings (relating to earliest symptoms of a disease, not linked to a previous disease, culture of cells from the tissue where a disease started, found in the tissue or organ where it started and a neoplasm found in the tissue or organ where a disease started) that are used in a restricted way in the medical context and are closely related to the remaining meanings via the core meaning of “first”. SemiMed users can understand general and medical meaning interactions after a first quick look at the LC of *primary*. This wider understanding is much less achievable in other resources like dictionaries. *Cambridge Dictionary*, for example, treats each meaning of *primary* as a separate entity having no relation with other meanings listed under the word entry (Figure 5.19). SemiMed users would also identify medical meanings with ease because the radial structure does not have the concept of ranked sub-entries that are often misleading for dictionary users. In other words, the comprehension problems raised in previous studies (Boonmoh et al., 2006; Nesi & Hail, 2002; Winkler, 2001) might be eliminated while using SemiMed to look up multi-meaning words.

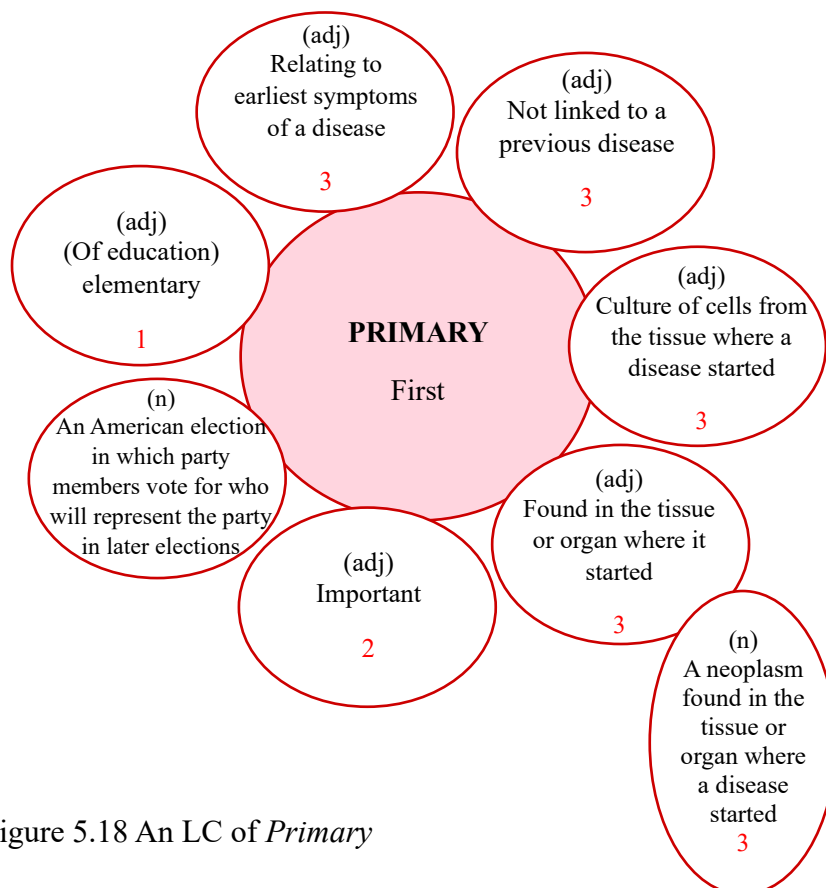


Figure 5.18 An LC of *Primary*

Primary

(adj) more important than anything else; main

(adj) (education) of or for the teaching of young children, especially those between five and eleven years old

(adj) happening first

(n) in the US, an election in which people choose who will represent a particular party in an election for political office

Figure 5.19 *Primary* in *Cambridge Dictionary* (<http://dictionary.cambridge.org>)

Second, SemiMed LCs not only showcase interactions between general and medical meanings but also specify these semantic interactions, whether general and medical meanings are in a relation of polysemy, homography or both. Providing the meanings of each headword is a must, but it seems insufficient on its own to resolve the issue around depth of vocabulary knowledge in Hsu's (2013) MWL. A grasp of polysemy and homography gained through LCs may help deepen and facilitate the understanding and interpretation of word meanings across different contexts.

Returning to the word *primary*, for example, it would be pedagogically useful to know if the word is polysemous. Let us assume, for the sake of argument, that SemiMed users already know the general meaning(s) of *primary*. They would be able to link their prior knowledge of the general meaning(s) with their new knowledge of the medical meaning(s) once they knew that the general and medical meanings are derived from the core meaning of "first". Such links may assist users in consolidating known general meaning(s) and retaining newly acquired medical meaning(s). Furthermore, SemiMed offers a shortcut to learning multi-meaning words like *primary*. Instead of trying to remember all the meanings listed under *primary* in the *Cambridge Dictionary*, SemiMed users, after understanding the word's general and medical

meanings, can remember the word's core meaning and interpret other meanings depending on future contexts.

Another example of SemiMed's polysemy- and homography-related benefits is *peel*. *Peel* is more complex than *primary* because it has polysemous meanings used in both contexts (a face treatment that makes the surface of the skin smoother, the outer layer of a fruit, and to remove the outer layer of something) and a homograph whose meaning is purely general (a pole with a flat part at one end for removing bread from an oven). SemiMed users are expected, after examining the LC of *peel* (Figure 5.17), to perceive links between the polysemous meanings via the core meaning of "outer layer" and become aware that "a tool to remove bread from an oven" is the meaning of a word that only shares the same written form. Such interpretations of SemiMed LCs can inform a new vocabulary learning strategy that involves storing core meanings of polysemes and homographs separately and relying on a context to decide which core meaning should be activated and which meaning from the activated core meaning should be used.

Third, the incorporation of word meaning frequency information into SemiMed LCs consolidates the word form frequency findings inherited from the MWL. The calculation of word meaning frequency in the two corpora proved that it would only be necessary in some cases to know all the meanings of an MWL word. This was seen in the cases of *cardiac*, *cataract*, *chronic*, *disorder*, *induce*, *liver*, *palsy* and *tumour* (see in the Appendix 2). In the OED, *cardiac* has two meanings: "relating to the heart" and "relating to part of the stomach". However, after the corpus-based calculation, the second meaning was found to be much less frequent in both general and medical contexts and was thus discarded from the LC. In other words, SemiMed filters frequently encountered meanings and increases the pedagogical usefulness of the MWL by highlighting which words and which word meanings are worth learning and teaching.

The technicality levels also enhance the pedagogical potential of SemiMed by providing contextual clues for users to arrive at appropriate interpretations of word meanings, especially unusual ones. Together with the radial structure, the technicality levels may make the process of looking for unusual meanings in SemiMed more straightforward than in dictionaries because users can effortlessly identify which ones are medical by scanning LCs to find meanings ranked at Levels 2 and 3. The presence of technicality levels in SemiMed suggests that corpus-based results of word meaning frequency could be transferred to a teachable element in a lexical resource and this element is pedagogically practical in use.

5.5 Future work and conclusion

The present study was designed to address semantic issues around word form frequency-based wordlists such as Hsu's (2013) MWL. A novel approach combining semantic and corpus-based analyses was implemented to remove the root cause of the lack of semantic depth in the MWL. This issue was uncovered from the review of current literature, which observed that the development of wordlists still relies heavily on word form frequency to deal with multi-meaning words. The outcome of these analyses is 40 LCs, named SemiMed (in this piloted version), and an LC template. Two significant features of SemiMed that make this resource an improvement over word form frequency-based lists like the MWL, and other resources, like conventional dictionaries, are its radial structure and technicality levels. The visualization of general and medical meaning interrelations in the LCs is believed to deepen the knowledge of vocabulary and eliminate the confusion caused by sub-entry-structured dictionaries, which both have significant effects on learning and teaching multi-meaning words, especially those with unusual meanings found in discipline-specific contexts like semi-technical medical vocabulary. With the four-level degree of technicality, SemiMed is one of the very few resources that considers and transfers word meaning frequency into a learnable and teachable aspect of vocabulary learning and teaching. This finding suggests that the

calculation of word meaning frequency is an achievable task that may remedy the shortcomings of wordlists that rely entirely on word form frequency. Future studies on currently available wordlists with due consideration given to word meaning frequency are therefore recommended. Moreover, since the pedagogical usefulness of SemiMed is theoretically based, empirical evidence will need to be gathered through pilot investigations on the practical use of SemiMed in classroom settings. Last but not least, additional studies will need to be undertaken to develop a full version of SemiMed.

CHAPTER 6: THE SEMIMED PILOT

Statement of Authorship

Title of paper	Piloting SemiMed – a mini semantic visualization dictionary of semi-technical medical vocabulary: A response to semantic deficiencies in a medicine-related wordlist
Publication status	Published – revised for this thesis for stylistic consistency
Publication details	Le, C. N. N., & Miller, J. (2022). Piloting SemiMed – a mini semantic visualization dictionary of semi-technical medical vocabulary: A response to semantic deficiencies in a medicine-related wordlist (Report on A.S. Hornby Dictionary Research Award Project). https://www.hornby-trust.org.uk/projects
Conference presentation	Le, C. N. N. (2022, July 12-16). <i>Piloting SemiMed – a mini semantic visualization dictionary of semi-technical medical vocabulary: A response to semantic deficiencies in Hsu's (2013) Medical Word List</i> [Paper presentation]. Euralex Conference, Mannheim, Germany.

Principal Author

Name of principal author (Candidate)	Chinh Ngan Nguyen Le		
Contribution to the paper	Performed all data collection and analysis stages, interpreted data, developed first draft, wrote and revised manuscript, and act as corresponding author.		
Overall percentage (%)	90%		
Certification	This paper reports on original research I conducted during the period of my Higher Degree Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis.		
Signature		Date	21/11/2023

Co-author contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of co-author	Julia Miller		
Contribution to the paper	Supervised development of work, helped in data analysis and interpretation and manuscript revisions.		
Signature		Date	21/11/2023

Report on A.S. Hornby Dictionary Research Award Project

Title: Piloting SemiMed – a mini semantic visualization dictionary of semi-technical medical vocabulary: A response to semantic deficiencies in a medicine-related wordlist

Country: Australia/Vietnam

Dates: July 2021 – October 2022

Lead researcher: Chinh Ngan Nguyen Le

6.1 Project summary

The project aimed to pilot SemiMed – the final product of a larger-scale project developing a mini dictionary where meanings of semi-technical medical vocabulary are visualized in semantic networks. The compilation of SemiMed stems from a demand for a reference source mainly designed for teaching/learning semi-technical medical vocabulary, because this type of vocabulary usually brings about pedagogical challenges. The starting point was Hsu's (2013) list of semi-technical medical words, whose creation and presentation incur semantic deficiencies (Le & Miller, 2023). Multi-meaning words in Hsu's list, which are anticipated to cause difficulties in learning and teaching, were semantically analysed with reference to theories in lexical semantics. Cantos and Sanchez's (2001) model of Lexical Constellations (LCs) was adopted as a means of showcasing intricate interrelations between general and specialized meanings of semi-technical medical words. A corpus-based analysis followed to quantify the word meaning frequency. To examine the practicality of SemiMed, a pilot study was conducted in which 18 EFL medical students were provided with lexicographic resources, including a sample of SemiMed as well as conventional dictionaries, to help them use appropriate vocabulary while role playing targeted medical scenarios. Focus groups were conducted to gain their feedback on the usefulness of the materials, informing improvements to SemiMed's design to better meet user needs.

6.2 Background and objectives

6.2.1 Statement of research problem

In teaching and learning English for specific purposes (ESP), semi-technical vocabulary has long been downplayed, as greater attention has been paid to technical vocabulary. The acquisition of only technical words, however, is inadequate for a full understanding of specialized readings (Cohen et al., 1988), and the body of literature contains several studies which underscore the importance of semi-technical words and their complicated nature (Baker, 1988; Farrell, 1990; Fraser, 2009, 2012; Higgins, 1966; Le & Miller, 2023; Li & Pemberton, 1994). As its name suggests, semi-technical medical vocabulary is hybrid in nature, i.e., conveying general and medical meanings, and sometimes activating additional meanings in a specialized context that differ from those in the general context. By analysing 302 semi-technical medical words, Le and Miller (2023) elucidate that a root cause of learning and teaching difficulties lies in polysemy and homography. Semi-technical medical words are subject to meaning variation. This type of vocabulary has multiple related (polysemic) and unrelated (homographic) meanings across different contexts and this, according to Fraser (2012), “provide[s] learners with the greatest difficulty” (p. 135)

Given that semi-technical medical vocabulary has a hybrid nature, that is to say, it is found in both general and medical contexts, learners of English for medical purposes (EMP) may need to consult both general and specialized dictionaries to gain an adequate interpretation of semi-technical medical words. Moreover, most dictionaries are structured in a unidimensional format, with senses vertically listed under a dictionary word entry. This makes it harder to retrieve polysemous words, which are multidimensional in structure (Geeraerts, 2006). Therefore, this study was conducted to examine the feasibility of an alternative lexicographic resource developed to deal exclusively with semi-technical medical vocabulary and address the semantic intricacies of polysemy and homography.

6.2.2 Literature review

6.2.2.1 Polysemy and homography in lexical semantics

Polysemy refers to a word having multiple related meanings. Homography is a reverse phenomenon where two words with different meanings share the same written form. In lexical semantics, attention has been paid to the distinction between polysemy and homography, and the mental representation of polysemy.

There are two approaches to distinguishing polysemy from homography: etymology-based and intuitive judgment (Lyons, 1977, as cited in Klepousniotou, 2002). The former approach traces the word origin to distinguish polysemy from homography – homographs are derived from distinct roots, while polysemous words are not (Croft & Cruse, 2004; Klepousniotou, 2002). The latter approach rests on the native speaker's intuition to judge the relatedness of meaning and then determine whether meanings are closely related enough to be polysemous. Each approach has its own shortcomings, including uncertainty about the historical derivation of words (Klepousniotou, 2002) and the undesirably high level of subjectivity resulting from the existence of arbitrariness (Lyons, 1969, as cited in Atkins, 1991). Combining the two approaches may remedy the shortcomings. For example, intuitive judgment can be informed by etymology-based evidence to minimize subjectivity.

Regarding the mental representation of polysemy, it is worth mentioning that only the structural nature of polysemy is discussed within the scope of this section because polysemous meanings intertwine in a more complicated manner than homographs and thus need elaboration. Two standpoints that merit discussion in the study context are Ruhl's (1989) *monosemy* and Lakoff's (1987) *radial category*. Ruhl (1989, 2002) argues that despite having many meanings, only one abstract meaning is stored in the brain; other meanings of a polysemous word are constructed via semantic and pragmatic context clues. By contrast, Lakoff (1987) maintains that a polysemous word is a conceptual category and we store "a

category of distinct polysemous senses rather than a single abstract monosemous sense” (as cited in Evans & Green, 2006, p. 330).

The development of Lakoff’s proposed radial categories was later parameterized by Tyler and Evans’s (2003a, 2003b) principles (also known as the *Principled Polysemy* approach). However, in essence, a radial category does not change its nature. It remains “a conceptual category in which the range of concepts are organised relative to a central or prototypical concept” (Evans & Green, 2006, p. 331). In other words, the radial category visualizes how different meanings of a word interact vis-à-vis a central meaning, the one that typically presents mutual semantic properties of other meanings. Polysemy under this perspective is structured in a “highly complex” way (Evans & Green, 2006, p. 328) and has “multidimensional structural relations” (Geeraerts, 2006, p. 351).

6.2.2.2 Polysemy and homography in lexicography

Turning now to polysemy and homography from the perspective of lexicography, ways in which these two phenomena are handled in two lexicographic resources (wordlists and dictionaries) will be discussed.

Wordlists

A wordlist is a list that indicates a finite number of words learners need to master for their particular learning purposes. For example, Hsu’s (2013) Medical Word List (MWL) contains 595 semi-technical words that appear so frequently that learners of EMP are advised to spend their time learning the listed words to gain adequate comprehension of what they hear or read.

Having the frequency of word forms as an underlying basis for the selection of candidate words, frequency wordlists come at a price, that is, the wordlist creation and presentation do not pay due attention to semantic relations. The wordlist creation rests upon the automatic corpus-based distinction of word forms rather than word meanings, thereby

disregarding the phenomena of polysemy and homography (Watson-Todd, 2017), and consequently failing to include them in the presentation. Although wordlists play a significant role in delimiting vocabulary size and thus letting learners know which words they should focus on, Le and Miller (2023) express a growing concern over the absence of semantic explanation in wordlists, especially wordlists of semi-technical words like the MWL, in which 51% of words are polysemes or homographs or both.

Dictionaries

Compared with wordlists, dictionaries have more sufficient and elaborated presentation of polysemy and homography. There are several ways to order related and unrelated meanings within a dictionary entry. Still, for reasons of space, only two internal structures are discussed in this report because they are the ones most commonly used in conventional dictionaries. These are *linearization* and *hierarchy*.

In a linear structure, “all [meanings] have equal status ... [and] are presented on one level” (Moerdijk, 2003, p. 285). A hierarchical structure, on the other hand, has “two or more levels on which related [meanings] are grouped” (p. 286). It has been argued, however, that these internal structures do not fully capture the semantic intricacies of polysemy and homography.

Given that meanings are all listed on the same level, linearization may not imply semantic inter-relatedness and thus, dictionary users, when they look at linearly organized meanings, may tend to treat each meaning as a discrete element that has no relation to remaining meanings. For example, *Cambridge Dictionary* (<https://dictionary.cambridge.org/>) does not flag (a) the distinction between homography and polysemy, and (b) the relation among polysemous meanings due to its linear structure of presenting homographs (e.g., *colon*) and polysemous meanings of words (e.g., *benign*) at the same level (Figure 6.1).

colon	benign
n. (body part) the lower and bigger half of the bowels in which water is removed from solid waste	adj. (person) pleasant and kind
n. (sign) the symbol: used in writing, especially to introduce a list of things or a sentence or phrase taken from somewhere else	adj. (disease) a benign growth is not cancer and is not likely to be harmful

Figure 6.1 The linear structure of *Colon* and *Benign*. Definitions from *Cambridge Dictionary online* (<https://dictionary.cambridge.org/>) in the order in which they appear

In a hierarchical structure, by contrast, meaning groupings determined from their relatedness seem to provide more straightforward indications than the linear structure does. The distinction between homography and polysemy is drawn because homographs and polysemous words are grouped in separate entries. Polysemous meanings are grouped within an entry in a hierarchical order (Figure 6.2). Nevertheless, although the hierarchy of polysemous meanings establishes their relation, how the different meanings relate to each other is not explicitly showcased. In other words, from the standpoint of lexical semantics, particularly the multidimensional structural relations of polysemy (Geeraerts, 2006), a hierarchical structure still has a minimal capacity for showcasing polysemous relations.

colon (n)	benign (adj)
Entry 1: the part of the large intestine that extends from the cecum to the rectum	1a: of a mild type or character that does not threaten health or life
Entry 2	b: having no significant effect
1 plural cola: a rhythmical unit of an utterance	2: of a gentle disposition
2 plural colons:	3a: showing kindness and gentleness
a: a punctuation mark	b: favourable, wholesome
b: the sign	
Entry 3: a colonial farmer or plantation owner	

Figure 6.2 The hierarchical structure of *Colon* and *Benign*. Definitions from *Merriam-Webster Dictionary online* (<https://www.merriam-webster.com/>) in the order in which they appear

6.3 Description of research

A review of the literature, then, indicates that the semantic structures in lexical semantics are not fully observable in lexicographic practices, raising questions as to whether the conventional format of lexical resources does full justice to the intractable nature of linguistic phenomena. To begin to address this issue, this study aimed to develop a non-conventional lexical resource of semi-technical medical vocabulary that takes into account theories of polysemy and homography in lexical semantics. The study had two phases:

- Developing a pilot version of SemiMed, an exclusive resource of semi-technical medical vocabulary that considers polysemy and homography from the perspective of lexical semantics
- Piloting SemiMed to test its usefulness in comparison with current conventional resources

6.3.1 Developing SemiMed

The MWL was a starting point for the development of SemiMed. The study took advantage of Le and Miller's (2023) findings and conducted a semantic analysis of 302 multi-meaning semi-technical medical words in the MWL. The MWL was chosen for two reasons:

- A wordlist, unlike a dictionary, usually has a finite number of words. This would ensure the feasibility of the study. More importantly, although the MWL has semantic issues due to its reliance on word form frequency, it still informs us of semi-technical words that frequently occur in medical contexts.
- A semantic analysis of words in a wordlist is more pedagogically significant than analysis of words in a dictionary. Words in the MWL are chosen selectively on the basis of frequency, which means they occur so frequently in medicine that EMP learners should devote time and effort to learning them. In comparison, not every word in a general/medical dictionary is worth learning. Additionally, a wordlist has minimum semantic features, so the semantic improvement of the MWL may be expected to compensate for the shortcomings of word form-based wordlists and so pave the way for the development of a resource containing frequently occurring semi-technical medical words with sufficient semantic explanation.

6.3.1.1 Qualitative analysis

The qualitative analysis of 302 MWL words was rooted in the theories of lexical semantics reviewed above. First, the analysis used a combined approach that considered both etymology and speaker intuition to distinguish polysemy from homography. Second, although Ruhl (1989) and Lakoff (1987) hold contradictory views on how a polysemous word is mentally stored, at the heart of monosemy and radial category, a shared concept can be observed of a core meaning (variously named an abstract, central or prototypical meaning) – the one from which polysemous meanings are derived. Following this observation, a

visualization of polysemous relations was proposed in response to the hierarchical structure’s minimal capacity to showcase how each polysemous meaning interrelates with others. This allows a higher level of hierarchical structure, where polysemous relations are not implicit or implied but explicitly visualized. Rather than vertically listing polysemous meanings under a word entry such as *benign* in Figure 6.2, the qualitative analysis further visualizes how polysemous meanings interact vis-à-vis a core meaning. The highly complex structure of polysemy in Lakoff’s radial category is acknowledged, and his idea that “the range of concepts are organized relative to a central or prototypical concept” (Evans & Green, 2006, p. 331) helps to explain the semantic visualization.

6.3.1.2 Quantitative analysis

To address the word form frequency-related issues, a corpus-based analysis was carried out to examine how frequently word meanings presented in our semantic visualization appear in a general and a medical corpus. Two corpora were selected – English Web 2020 (36 billion words) and the Medical Web Corpus (34 million words) (Table 6.1). The unit of analysis was word meaning frequency and the unit of counting was collocate frequency. The analytical method was based on an approach to determining meaning by collocation (Cantos, Sanchez, & Almela, 2009; Hoey, 2012; Perez, 2013). Simply put, the meaning interpretation of a word in a corpus is retrieved from an extensive investigation into its collocations. The collocational data were computed using the online corpus analysis tool Sketch Engine.

Table 6.1 Details of English Web 2020 and Medical Web Corpus

	English Web 2020 (enTenTen20)	Medical Web Corpus
Tokens	43,125,207,462	42,054,011
Words	36,561,273,153	33,961,786

Sentences	2,008,143,278	1,545,862
Documents	78,373,887	526

6.3.1.3 Procedural demonstration of *diffuse*

Step 1: Oxford English Dictionary (OED) definition adaptation and simplification

An MWL headword (e.g., *diffuse*, see Figure 6.4) was prepared by adapting the procedure for looking up MWL headwords in the OED used by Le and Miller (2023). OED definitions were then simplified to:

- Make OED definitions shorter and easier to understand for learners at a minimum upper-intermediate level of English proficiency
- Ensure the use of simply reworded OED definitions in the semantic visualization does not infringe copyright

Step 2: Identification of core and other related meanings

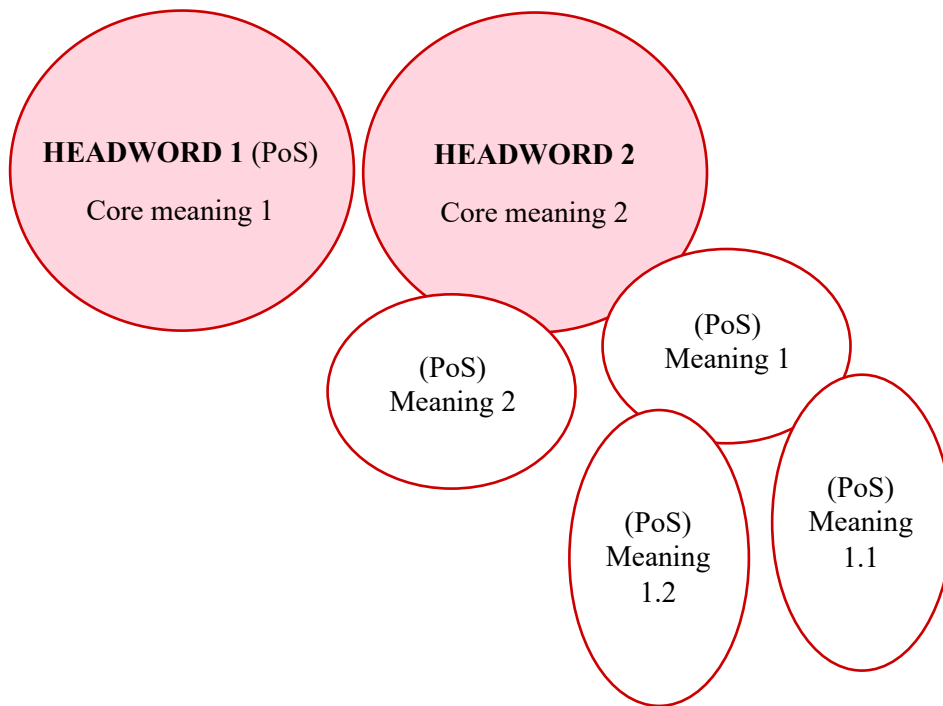
The Principled Polysemy approach informed the identification of core and other related meanings. Criteria to determine a core meaning in this study were derived from Evans (2005). A core meaning needs to fulfil at least one, and preferably more than one, of three criteria: “(1) [closely relates to the] historically earliest attested meaning, (2) predominance in the semantic network, [...] (3) predictability regarding other senses” (Evans, 2005, p. 44). The study adopted Le and Miller’s (2023) identified core meanings of 302 troublesome semi-technical medical words from the MWL. Briefly, Le and Miller evaluated the first criterion by looking at the etymological reference in the OED. The remaining criteria were based on the intuitive judgment of three evaluators. From the core meaning identification using this combined approach, Le and Miller reasoned that they would classify a word as polysemous if all its senses shared a core meaning; new senses creating new core meanings would be classified as homographs. This reasoning was used to distinguish polysemy from homography.

Hierarchy of other related meanings

Non-core meanings were further analysed by putting closely related meanings into a cluster and establishing a meaning hierarchy within a cluster.

Step 3: Visualization of semantic relations in Lexical Constellations

The study adapted Cantos and Sanchez's (2001) Lexical Constellation (LC) model to visualize how related meanings interact vis-à-vis a core meaning. The generic pattern of an LC has a core meaning placed at the centre and surrounded by multiple, related meanings located in outer layers, which showcase the degree of interconnectivity (Figure 6.3). Each LC represents a (polysemous) word and if two words are homographs, they have two separate LCs.



POS: Part of Speech

Level 1: Meanings 1 and 2

Level 2: Meanings 1.1 and 1.2

Figure 6.3 Generic pattern of LCs of a polysemous word and a homograph (Adapted from Rizzo & Sanchez, 2010)

Step 4: Quantification of the meaning frequency of occurrence

Sketch Engine (Word Sketch) was used to export collocates of a searched headword and select the top 15 most frequent collocates in two corpora (English Web 2020 and the Medical Web Corpus). Meanings were assigned to collocates and then divided into four levels of technicality (Table 6.2). Level 0 indicated that no technicality information was shown for a meaning, the meaning is not found in the Medical Web Corpus, and it is considered a purely general meaning. Technicality levels 1 – 3 were embedded in LCs (Step 4 in Figure 6.4).

Table 6.2 Technicality level description

Level 0	This meaning is solely used in general contexts
Not indicated in LCs	Found in the top 15 meanings in English Web 2020, but not found in the Medical Web Corpus
Level 1	This is a generally used meaning
	Found in the top 15 meanings in English Web 2020, and outside of the top 15 meanings for the Medical Web Corpus
Level 2	This meaning is used in both general and medical contexts
	Found in the top 15 meanings in both English Web 2020 and the Medical Web Corpus
Level 3	This meaning is used only in medical contexts
	Found in the top 15 meanings in the Medical Web Corpus, and outside of the top 15 meanings for English Web 2020; or Found in the top 15 meanings in the Medical Web Corpus, but not found in English Web 2020

Step 1: Simplify OED definitions of <i>diffuse</i>	DIFFUSE (adj) Spread out (adj) (Of disease) in more than one place (v) To (make something) spread (v) To make something weaker
---	---

<p>Step 2.1: Identify 2 core meanings and 3 other meanings relating to core meaning 1</p> <p>Step 2.2: Put the 3 other meanings in clusters and indicate the hierarchy</p>	<p>DIFFUSE</p> <p><i>Core meaning 1: Widespread</i></p> <p>Meaning 1: (adj) Spread out</p> <p>Meaning 1.1: (adj) (Of disease) in more than one place</p> <p>Meaning 2: (v) To (make something) spread</p> <p><i>Core meaning 2: To make something weaker</i></p>
<p>Step 3: Develop 2 LCs of <i>diffuse</i> (with 3 polysemous meanings) and its homograph</p>	<p>The diagram illustrates two meaning clusters for the word 'diffuse'. The first cluster, highlighted in light blue and labeled 'A meaning cluster', contains three items: '(adj) Spread out', '(adj) (Of disease) in more than one place', and '(v) To (make something) spread'. The second cluster contains two items: 'DIFFUSE (v) To make something weaker' and 'DIFFUSE Widespread'.</p>

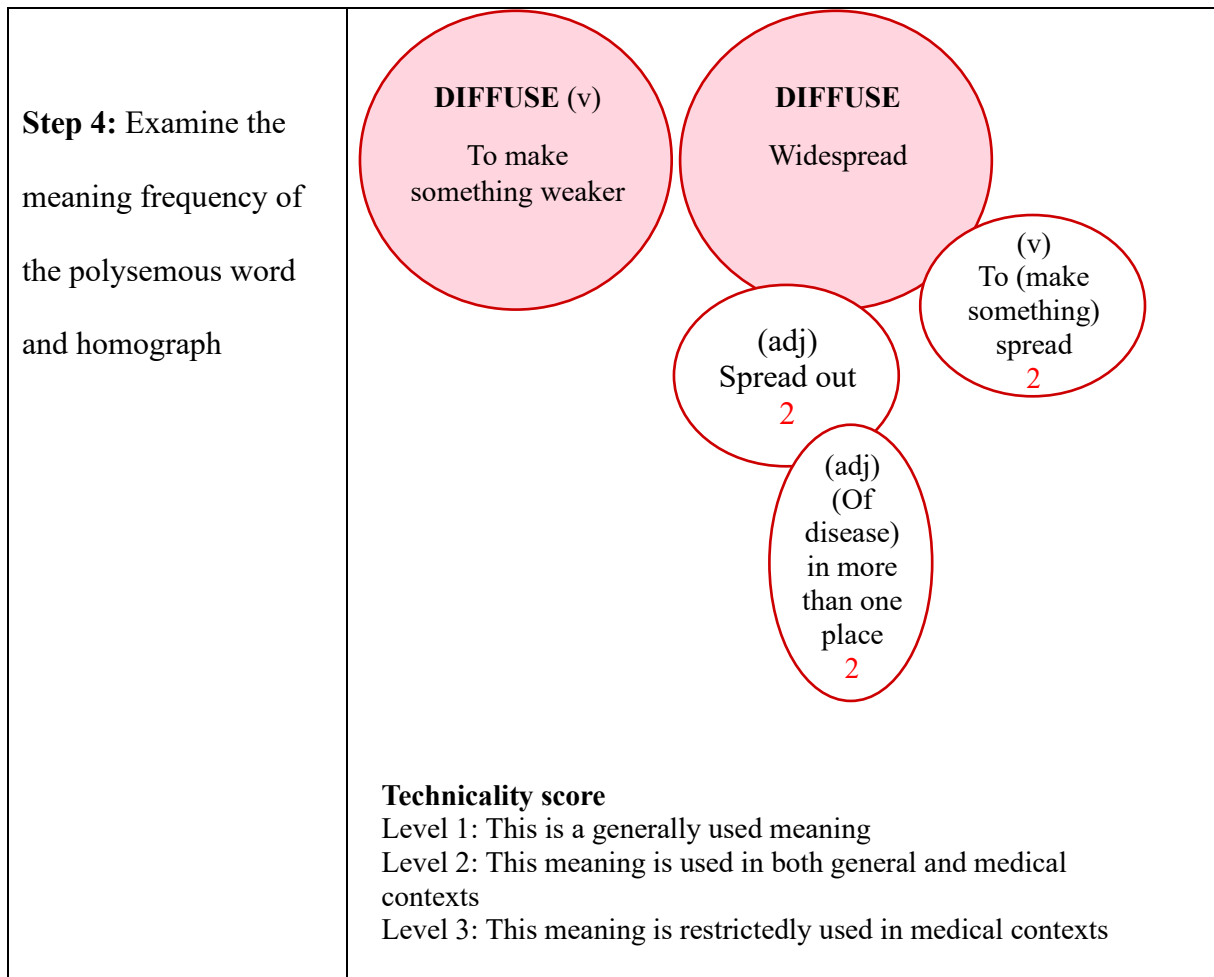


Figure 6.4 Procedural demonstration of *Diffuse*

6.3.2 Piloting SemiMed

6.3.2.1 Sampling

Forty LCs were selected for the pilot study. A wide range of LC constructs were taken into consideration during the sampling process to ensure pilot words closely reflected the characteristics of SemiMed LCs. The sample included LCs of (a) single-meaning words (e.g., *colon*, Figure 6.5), (b) multi-meaning words with a single core meaning (e.g., *benign*, Figure 6.6) and (c) multi-meaning words with more than one core meaning (e.g., *diffuse*, Figure 6.7).

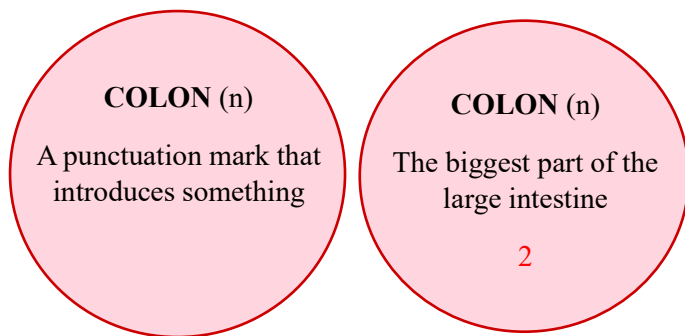


Figure 6.5 Two homographs

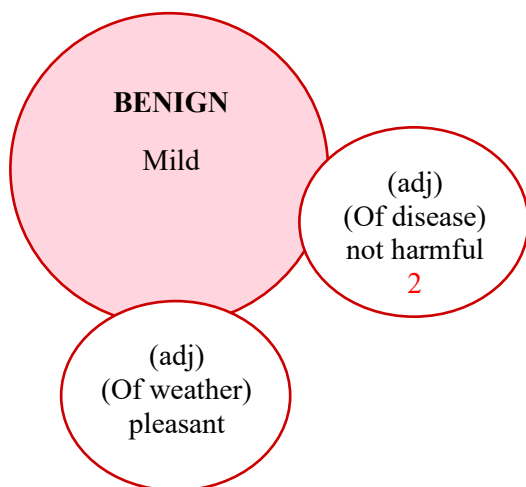


Figure 6.6 A polysemous word

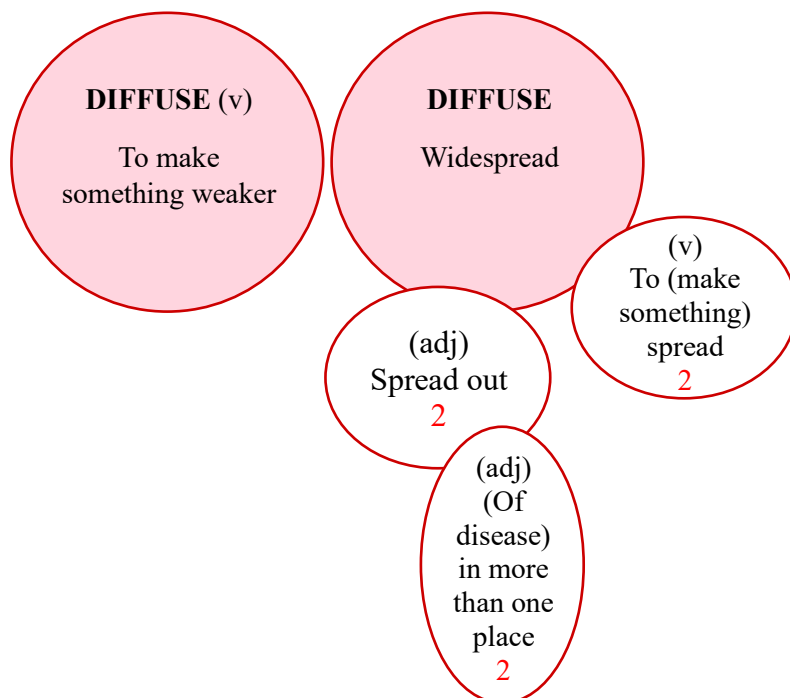


Figure 6.7 A polysemous word and a homograph

6.3.2.2 Participants

Eighteen EFL medical students from a University of Medicine and Pharmacy (UMP) in Vietnam were participants in the pilot study. They were recruited based on their English proficiency. Eligible participants were students who majored in medical fields and possessed an upper-intermediate or higher level of English.

6.3.2.3 Lexicographic resources

Participants were allowed to use three resources:

SemiMed which presents word meanings in the format of an LC.

Two designated dictionaries

- A general dictionary: *Cambridge Dictionary* (<https://dictionary.cambridge.org/>), in which definitions are presented in the linear format.
- A specialized dictionary: *Merriam-Webster Medical Dictionary* (<https://www.merriam-webster.com/medical>), in which definitions are presented in the hierarchical format.

6.3.2.4 Online platform

All forty pilot LCs were drawn using Inkscape software, then uploaded onto H5P (<https://h5p.org/>) and finally embedded in the UMP's Moodle system for participants to access. The LCs were alphabetically ordered and presented in four 'books' for ease of access (SemiMed A-C, SemiMed D-I, SemiMed L-P, and SemiMed R-T) (Figure 6.8). A pop-up box was designed to show detailed information of the technicality level (Figure 6.9).

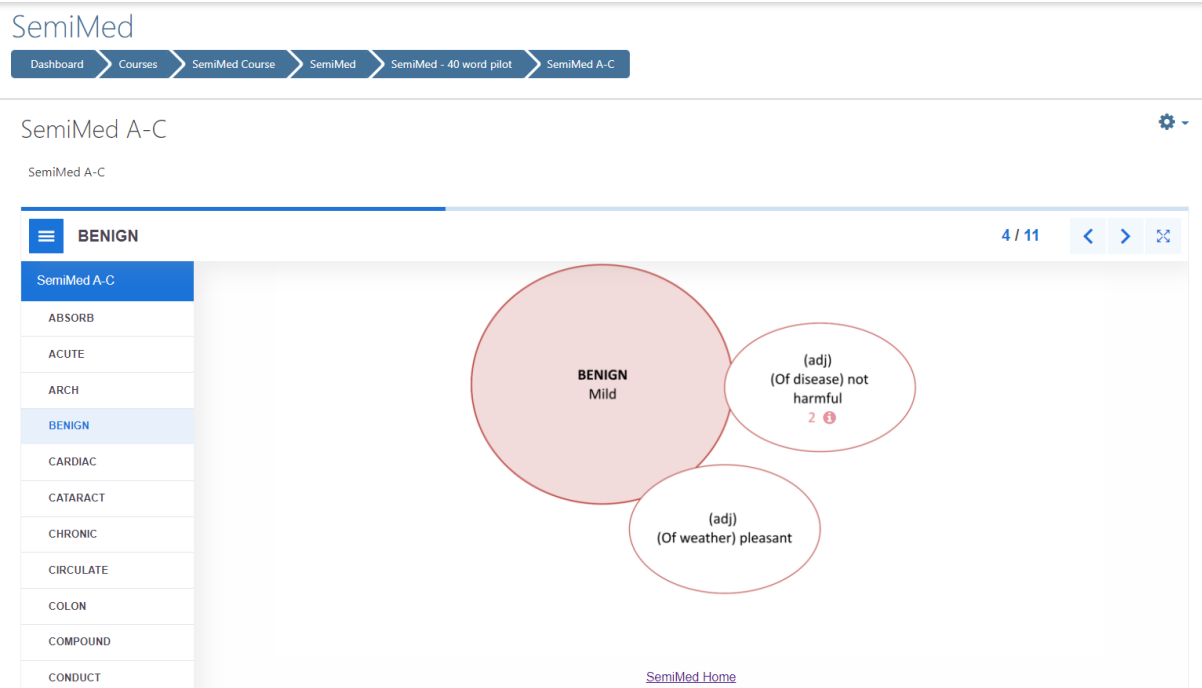


Figure 6.8 A Moodle interface of the LC of *Benign* (in Book 1: SemiMed A-C)

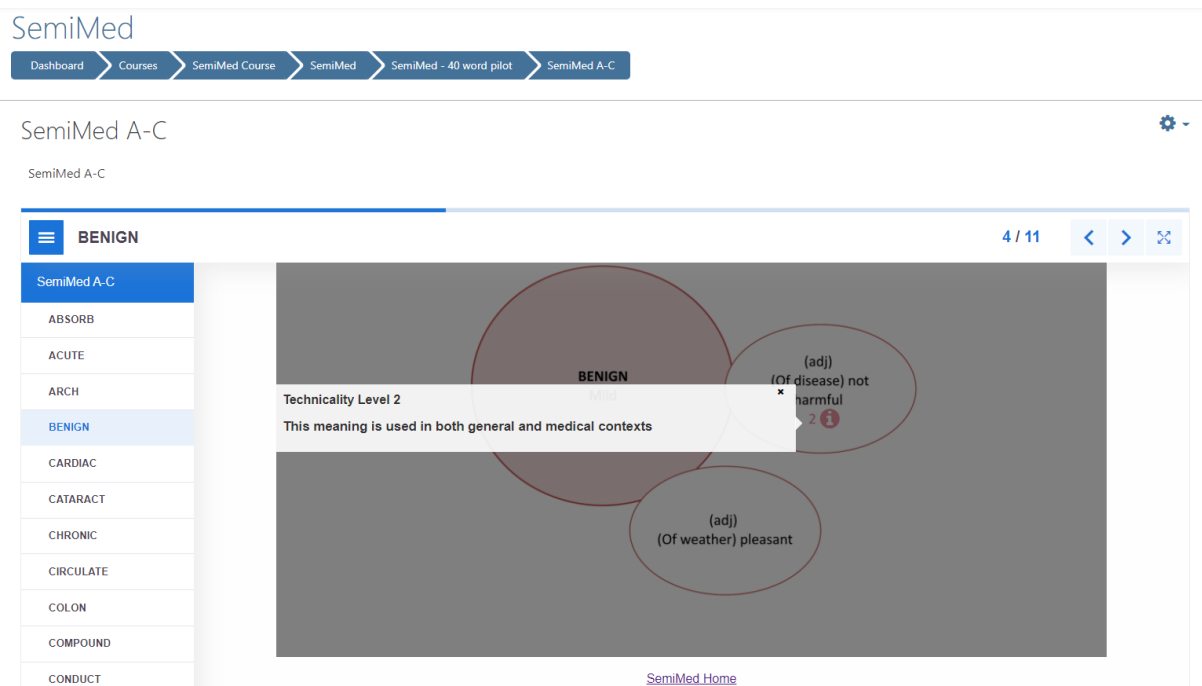


Figure 6.9 A pop-up box indicating the detailed technicality level

6.3.2.5 Medical scenarios

Five scenarios were written around topics that closely mimicked real-life situations that participants might experience. Each scenario targeted six pilot words, making a total of 30. The

remaining ten words were example words shown to participants in the Induction phase of the study (see the Meeting structure). In essence, scenarios set the scene to stimulate participants to use the pilot words in meaningful and relevant contexts. Gaps, indicated by ellipses, were left in the scripts to prompt participants to explain the target words to the ‘patient’ in their role-play.

6.3.2.6 Grouping

Eighteen participants were randomly divided into six groups (three people per group). The researchers scheduled a separate online Zoom meeting with each group.

6.3.2.7 Meeting structure

Induction: Participants were introduced to SemiMed and instructed to use this new resource, especially to interpret information presented in LCs. Participants were also informed of the dictionaries they were requested to use.

Activities: Participants chose their roles in scenarios and acted out the scenarios. They were encouraged to consult the lexicographic resources provided to use the pilot words as appropriately as possible. Specifically, they were requested to look up the first pilot word in the *Cambridge Dictionary* online, the second one in *Merriam-Webster Dictionary* online, and the third one in SemiMed. The rest of the pilot words could be looked up in any dictionary, allowing participants to choose which format they preferred. The researchers observed and facilitated as needed.

Focus group: Participants then engaged in a follow-up focus group where they shared their experiences of using SemiMed and the conventional dictionaries.

6.4 Results and evaluation

Thematic analysis was used to analyse the focus group data. The themes which emerged centred around participants' experiences of using SemiMed in the pilot study and also extended to their experiences of using other conventional resources prior to the pilot study.

6.4.1 Conventional resources

The majority of participants reported that, before the study, their two most frequently used general dictionaries were those published by Cambridge and Oxford (titles and editions were not given). In addition to monolingual dictionaries, they sometimes referred to bilingual dictionaries (SOHA¹ and TFLAT²) to search for Vietnamese meanings. Surprisingly, they seldom used medical dictionaries. Some mentioned *The language of medicine* (Chabner, 2020) as the only resource formally introduced in classrooms that provided them with topic-based medical terminology and learning strategies (e.g., morphemic analysis). Many participants used Google Search and Google Translate, which according to Participant O was a strategy passed down from senior to freshman students. Participants exploited these two functionalities of Google in various ways, ranging from looking up words to checking meanings of a known word.

Participant O shared that she usually put *what does word X mean?* in the search box and emphasized that “a strength of Google [Search] is that it provides you with images and some kinds of videos so that it helps you understand the word more clearly”. Several participants also considered Google Search engine as a medium for seeking related visual aids to assist them in understanding and learning a word. Participant B, for example, stated:

I think the most problem I get when I try to find meanings of the English medical terms is that there are some rare medical words I don't find on the Internet so I have to look up [a word] on the Google Images and I see the picture of it and I will have to try to guess [its] meaning.

¹ An e-dictionary available at http://tratu.soha.vn/dict/en_vn/Dictionary

² An English Dictionary App developed by TFLAT, a mobile application development team based in Vietnam

Another student went straight to video searching:

I prefer Youtube [videos] so I can learn more about [a] medical word. (Participant D)

Another common strategy shared among Participants B, C and M was doing Google searches for articles containing a specific word. They revealed that the retrieved articles offered contextual clues by which they could guess a meaning of the word. Google Translate was also used to get an instant Vietnamese translation of an English article (Participant C) or the Vietnamese equivalent of an unknown word (Participants A, N and R). Participant F used Google Translate for “fast-checking” whether she had correctly understood a word definition in the *Oxford Dictionary*.

Besides looking for and checking meanings of a word, Participant H added that she searched for the etymology of a word via Google. She also took advantage of Google to further learn about roots, prefixes, or suffixes from which a searched word is built. For Participant H, knowing the constituent parts of a searched word somehow made possible the guessing of the meaning of a new word made up of the same parts.

As previously stated, most participants reported little experience of using medical dictionaries. There are many possible reasons for this. Participant H reasoned that although she had been informed about medical dictionaries, she had never used one, as she could not afford the subscription fee. Participant K admitted that “I am afraid of being not fully understand [sic] words [in medical dictionaries]”. Participant R had used medical dictionaries but thought that the definitions of a word were sometimes more complex than the word itself. Participant M asserted that he had no intention of finding a medical dictionary:

I am a visual learner so I think that for medical dictionaries just [containing] words, they are not just for me.

Participants appeared to rely heavily on general dictionaries to look for medical meanings. However, they reported low satisfaction with the use of general dictionaries because they did not always find what they were seeking.

When I look up the meaning [of a word] in the [Cambridge and Oxford] dictionaries, they normally show the general use of the word and sometimes that word doesn't have the ... sometimes I cannot find the technical meaning. (Participant C)

For me, when I [try to] find some medical words in Cambridge or Oxford [dictionaries], there is no result so I have to use Google to find the meanings of the medical words that I want to figure out. (Participant I)

When I [used] Cambridge Dictionary, some of technical words didn't appear. (Participant K)

The possible inference of this feedback is that non-specialized Cambridge and Oxford dictionaries are not ideal for searching for medical meanings. Moreover, Participant N commented that the two general dictionaries occasionally led to homographs irrelevant to the medical context and this distracted her. In the case of medical meanings found in the general dictionaries, Participant O revealed that she found definitions in the *Oxford Dictionary* too lengthy to arrive at appropriate Vietnamese equivalents.

The challenges faced by participants while searching for medical meanings in general dictionaries may have contributed to their preference for *Google Search and Translate*. Nevertheless, the Google tools raise some concerns. Participant F admitted that although Google Search helped her target relevant medical articles or books, it was relatively time-consuming to understand a page of the books or even a paragraph of the articles where a searched word appeared. Worse still, she sometimes failed to double-check meaning(s) as they still did not make sense to her after reading through the translated parts. Participant L recalled that he sometimes had to read up to three documents but could not work out word meanings by himself, so he eventually consulted his teachers. Participant Q was concerned that understanding a segment of articles (or books) retrieved from Google Search was exhausting

because she needed to do other searches to be sure that she fully understood the entire segment containing the searched word.

The participants' experiences of conventional resources may highlight the concerning issue mentioned at the start of this report regarding the search for semi-technical vocabulary in general and specialized dictionaries. Participants seemed to get limited benefits from general (Cambridge and Oxford dictionaries) and medical (*The language of medicine*) dictionaries and this eventually drove them to use Google tools which caused them even more trouble. This finding supports the early stated need to develop a lexical resource of semi-technical vocabulary with an aim of creating a better experience in looking up this type of vocabulary.

6.4.2 SemiMed

SemiMed was developed to serve the practical need for a semi-technical vocabulary resource with dual foci on the logical presentation of general and medical meanings, and explicit guidance on polysemy and homography. In the pilot study, feedback from participants on the usefulness of SemiMed compared to the designated dictionaries was expected. However, the focus group data uncovered that participants also reflected on resources they used beyond the pilot study (as listed in the above section) and compared them with SemiMed. This newly emerging theme intertwined with the expected theme and provides a much better insight into the usefulness of SemiMed in comparison with current conventional resources; therefore, the two themes have been reported simultaneously rather than separately.

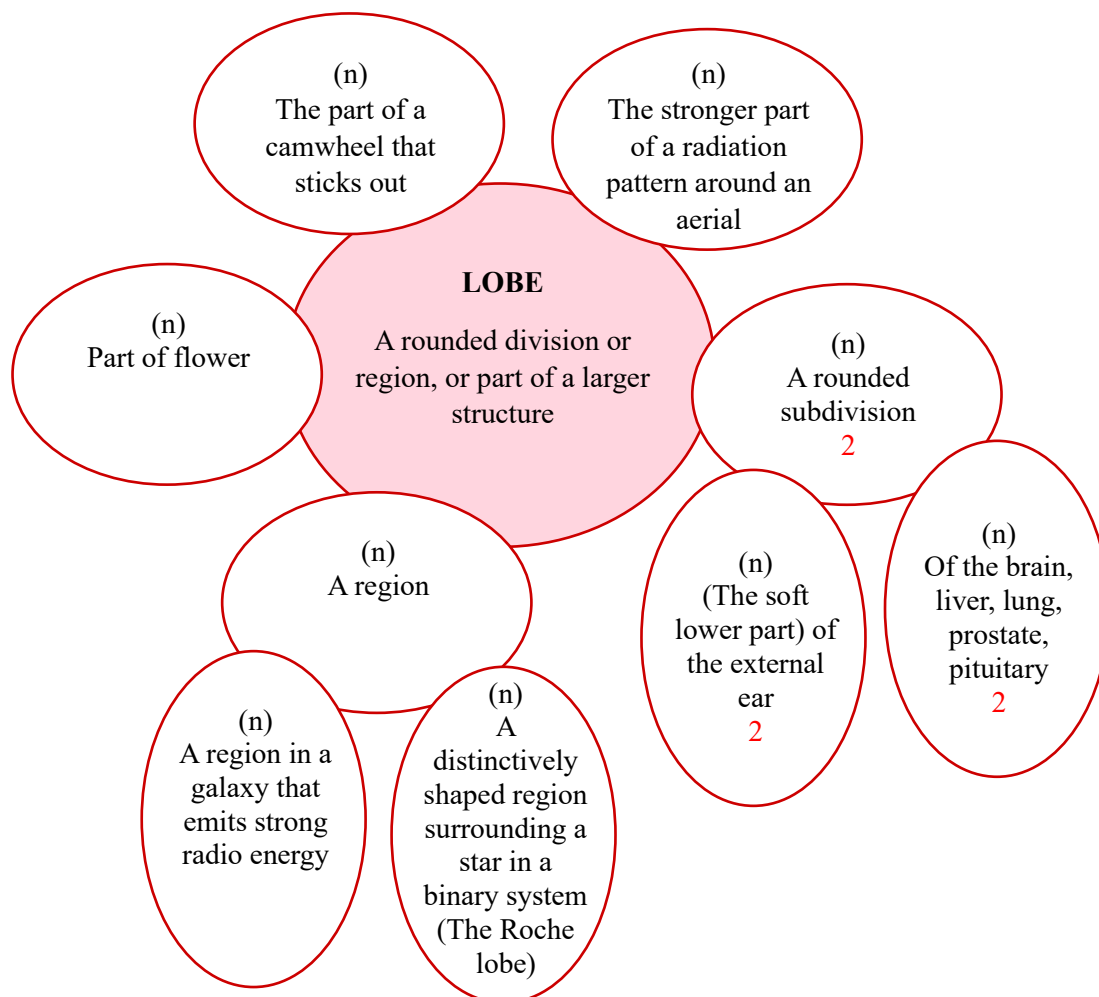
Participants identified three main advantages of SemiMed over conventional resources.

Concise and simplified definitions

Participants O and M gave feedback on the way word definitions are written in SemiMed compared to the two designated dictionaries. Participant O said when she looked up the last three pilot words in the *Cambridge Dictionary*, she found their definitions lengthy. She thus anticipated that if she had acted out her role as a specialist and explained the pilot words

using definitions from the *Cambridge Dictionary*, the group member who played the role of a patient might have become confused. When she looked up the same pilot words in SemiMed, she said that “the word is explained in a very short and simplified way, so I suppose that [SemiMed] will be applicable in real context when we have [a] conversation with our patients”. Participant M agreed and gave the pilot word *lobe* as an example of a word being more concisely defined in SemiMed than in the *Cambridge Dictionary* or *Merriam-Webster Dictionary* (see Figure 6.10).

SemiMed Dictionary



Cambridge Dictionary

Merriam-Webster Medical Dictionary

lobe

lobe (n)

n. (anatomy) any part of an organ that seems to be separate in some way from the rest,

A curved or rounded projection or division:

as

especially one of the parts of the brain,	a: a more or less rounded projection of a body
lungs, or liver	organ or part
n. (ear) an earlobe	b: a division of a body organ (as the brain,
n. (biology) a rounded or pointed part on a	lungs, or liver) marked off by a fissure on the
leaf that sticks out from the main part	surface

Figure 6.10 *Lobe* in the three dictionary formats

Non-conventional format

The participants' overview of the SemiMed format was that it is systematic, neat, and simple. Several participants (B, C, D, J and O) perceived SemiMed as a "mind map", with which medical students, according to Participant C, are familiar because they usually use mind maps for lesson revision. Participant B was impressed that SemiMed followed a mind map-like design to present word meanings and this design permitted the systematic learning of semi-technical medical words. Participant Q, agreeing with Participant B, reasoned that the inclusion of both general and medical meanings helped her to form "a general view of all meanings". This is considered a prominent advantage over conventional resources participants had used, such as Cambridge or Oxford dictionaries and Google tools, because four participants (B, C, F and J) could retrieve and understand general and medical meanings in SemiMed using only a single search.

Moreover, Participant A pinpointed that SemiMed laid out meaning interrelations in a logical manner, facilitating his ability to see how general meanings interact with medical ones. Participant H said that by knowing the relationship between general and medical meanings, she might expand her knowledge about the general meanings through the learning of medical meanings and vice versa. In the role-play activity, the systematic visualization of relationships between general and medical meanings might have created a better experience compared to the two designated dictionaries; many participants reported that it was more convenient

(Participant I), more helpful (Participant N), and faster (Participant Q) to search for and find appropriate medical meanings of pilot words in SemiMed than in the Cambridge and Merriam-Webster dictionaries.

Participants acknowledged the explicit distinction between polysemous words and homographs, which Participant I admitted he had not observed in the Cambridge and Oxford dictionaries. The feedback from participants revealed that the polysemy and homography presentation in SemiMed was easy to understand (Participants G, M and P) and more importantly, led them to the medical meanings of a word they were looking for, not the irrelevant meanings of its homographs (Participants E and F). In addition to reducing the distraction from unwanted homographs, the non-linear format, particularly the radial visualization, assisted participants in learning polysemous meanings of a word. Participant D commented that it was good to know a core meaning was shared among related meanings, and reiterated that if he understood the core meanings the mind map-like structure would enable him to memorize polysemous meanings faster.

Another significant advantage of SemiMed is its clear and neat display. Participants D, J and K liked the fact that SemiMed used a one-page display view, which helped them stay focused. Participant D stated that information in the Cambridge and Merriam-Webster dictionaries was so detailed that he sometimes lost his focus. Additionally, because they show word definitions in the form of linear lists, these dictionaries require participants to scroll up and down to read through search results. Participants J and L explained that it was fairly time-consuming to scroll through the entries to find the meanings used in the scenarios. In contrast, they felt SemiMed saved considerable time as the semantic visualization of a word was designed to fit the screen. In other words, participants were likely to spend less time manipulating displayed contents and thus their focus on finding meanings was enhanced.

Although SemiMed had a non-conventional format, no participants reported challenges in familiarizing themselves with it. Rather, the findings showed consensus among participants, emphasizing that the SemiMed interface is simple and user-friendly. Participant P clarified:

a strength [of SemiMed] is its format ... it's simple and clear ... so it's kind of easy to understand ... suitable for beginners and when people use it, we don't need to [have] a lot of technical and literacy skills.

For this reason, Participant K said she could manipulate SemiMed with ease after being guided through its functions in the Induction. The easy-to-use design seems to offer participants quick access to pilot words (Participant I) and then provide scaffolding for their understanding of word meanings (Participant E).

Technicality level

The level of technicality is a feature peculiar to SemiMed which attracted positive feedback from participants. One benefit of the technicality level is that it informs users of the context in which a certain meaning is more likely to appear. As explained earlier, semi-technical medical words can be used across different contexts, so the contextual details provided for each meaning are important (Participants J and O). Since medical meanings are central to learning semi-technical medical words, participants mainly commented on how the technicality level facilitated their search for medical meanings. When relating the difficulties in finding medical meanings using conventional general dictionaries (Cambridge and Oxford), Participant C shared that he had a more pleasant experience using SemiMed, especially its technicality function. Participants A and R were impressed by the technicality of meanings, which made the search for pilot words used in the medical scenarios much easier than searches in the designated dictionaries (Cambridge and Merriam-Webster). Furthermore, Participants C, F, and L agreed that the technicality level information significantly reduced the time allocated to searching for medical meanings. Thanks to the technicality information, Participant C was certain he spent less time finding medical meanings of pilot words, Participant F said she could

know immediately which meanings fitted in medical scenarios, and Participant L stated that his focus was quickly directed to medical meanings.

The three advantages of SemiMed are related to features absent in conventional dictionaries. These results are likely to further support the idea of developing SemiMed. Three implications were accordingly drawn from the findings.

First, participants positively reacted to the simplified definitions in SemiMed and this finding underscores the importance of well-written word definitions in medical dictionaries. Here, a “well-written” definition is understood to be one that has been constructed so that it is as easy as possible to understand by learners at all language levels. Although issues around writing a definition of a word have long been situated at the heart of the dictionary-making procedure, it is still believed that the issues deserve more attention, especially in the compilation of medical dictionaries, as this study indicated that difficulties in understanding definitions may make learners hesitant to use medical dictionaries.

Second, the radial structure seems more advantageous than a hierarchical format in terms of leveraging insight into relations (i.e., polysemy and homography) between general and medical meanings. SemiMed’s non-conventional format, which adheres to theories in lexical semantics, is de facto the mental representation of polysemy and homography. That may explain why participants considered the SemiMed format beneficial in facilitating mental processes such as understanding and memorizing general and medical meanings of semi-technical vocabulary. This implication provides some support for the consideration of lexical semantic theories in the development of lexical resources.

Third, participants’ appreciation of the technicality function indicates that word meaning frequency results can be transferrable into the four technicality levels, enhancing the pedagogical usefulness of SemiMed and showing that improvement of word form frequency-

based wordlists can be achieved. This has potential for resolving word form frequency-related issues in the MWL and other wordlists.

Despite advantages in word definitions, presentation format and technicality function, however, SemiMed nevertheless has some disadvantages.

Time-inefficient platform manipulation and not-so-attractive interface

As explained previously, SemiMed was uploaded onto H5P and this online platform supported four ‘books’ of alphabetically ordered words (SemiMed A-C, D-I, L-P and R-T). Technically speaking, to look up *benign*, for example, participants had to access the first book (SemiMed A-C) and scroll down until they retrieved the word (see Figure 6.8). This manual method of looking up a new word, which closely mimics the traditional method used with paper-based dictionaries, created a little confusion for participants. Participant P said that even though he kept a searched word in mind, he sometimes lost his train of thought and couldn’t decide which book he should select to find the word. He admitted to singing the ABC song to himself to aid his memory. Participant P added:

the weakness [of SemiMed] is that it has no finding tool so maybe sometimes it’s very time-consuming when I have to scroll down and search for the word.

Agreeing with Participant P, Participants I, J and L reported that this manual search of pilot words in SemiMed took more time than when using the two designated dictionaries.

Regarding SemiMed’s interface, the minimal design was intentionally chosen, and this was evaluated as simple and user-friendly by many participants. However, a few participants (B, F and R) still viewed the SemiMed design as less attractive and would have liked to see additional visual features. Participants K, L and R added that the absence of illustrative pictures in SemiMed not only made the interface look monotonous but also meant that new words could not be learned by looking at pictures.

Insufficient pronunciation and examples

Another disadvantage of SemiMed is the paucity of pronunciation guidance and of examples. Participants A, I, K and Q expressed their need to know how to pronounce a word in addition to its meanings. Participant Q reasoned that maybe because SemiMed did not show her how to pronounce a word, it was not of much benefit when speaking. In addition, Participant N said the lack of examples stopped SemiMed users from seeing a word in context. Participant E, when comparing SemiMed with the conventional dictionaries she used, stated that Cambridge and Oxford dictionaries gave her examples which enabled her to better understand what a word meant and how to use it. Participant Q added that some SemiMed word definitions were too concise to be readily understandable, so it was difficult to gain an adequate understanding by reading definitions with no examples. For instance, she could not adequately understand *conduct*, whose definition was *to do* in SemiMed, until she searched for relevant examples in the *Cambridge Dictionary*.

6.4.3 Suggestions for future improvement

Suggestions were made around potential features which participants believed should be added to SemiMed to mitigate its current disadvantages. First, Participants J, L and P suggested that the online platform should be upgraded with a search bar to automate the word searching process. Rather than manually looking up a new word in the four books, typing the word in the search bar and then clicking a search button to retrieve search results seemed to be more time-efficient and thus might create a more pleasant experience for users. Second, to maximize benefits, many participants recommended the incorporation of pronunciation aspects into SemiMed so that they could both read and pronounce a word correctly. Third, they recommended that images and pictures should be added where necessary to aid the comprehension of words such as those naming parts of the body (Participants E and R) and to accommodate the needs of visual learners (like Participant M). The use of visual illustrations might also improve the SemiMed interface, making it more vivid and attractive (Participants L

and R). Fourth, the inclusion of examples was highly recommended, as the majority of participants stressed the importance of seeing a word in context to better understand it and use it correctly. A sentence example would be “just fine” for this (Participant H). Furthermore, participants F, L and Q said that they spent a considerable amount of time reading longer texts (a paragraph or page of relevant documents) retrieved from a Google search. From their experiences it can be inferred that example sentences would both save time and satisfy the need to learn words from context.

6.5 Overall reflections and future plans

This study set out to pilot SemiMed, a new lexicographic resource of semi-technical medical vocabulary which is being developed in response to semantic deficiencies resulting from the reliance on word form frequency in the MWL. The development of SemiMed is based on theoretical premises of lexical semantics that have not been observed in current resources. The results of the pilot study show that SemiMed has some significant advantages over other resources, especially the radial visualizations of semantic relationships (polysemy and homography), which are the fruit of the consideration of lexical semantic theories during its development. SemiMed also addresses semantic deficiencies in the MWL because it takes into account word meaning frequency together with word form frequency. This methodological approach may pave the way for future studies which attempt to improve word form frequency-based wordlists. However, due to the limited timeframe of the study, and limited resources, SemiMed is not without flaws. The enhancement of its platform and provision of visuals, examples and pronunciation aspects are key areas that deserve further study. It is strongly believed that if these shortcomings are addressed, SemiMed has the potential to be of great benefit to EMP learners.

CHAPTER 7: CONCLUSION

7.1 Summary

This study revisited polysemy and homography, where a word form has multiple related and unrelated meanings, in lexical semantics, lexicography and corpus linguistics. It focused on investigating semi-technical medical vocabulary, a type of vocabulary situated in an area between medical terminology and general vocabulary that is characterized by having polysemes and/or homographs. Semi-technical medical vocabulary usually has multiple related and unrelated meanings which are differentially activated in general and medical contexts, potentially creating pedagogical difficulties. Issues of polysemy and homography in semi-technical medical vocabulary have not been fully addressed in current lexicographic resources like conventional general and medical dictionaries and corpus-based wordlists, leaving a gap that the study aimed to fill.

The study started by examining a list of semi-technical medical words, Hsu's (2013) MWL, which has a limited capacity to provide sufficient polysemy and homography annotations due to excessive reliance on an automatic process of counting word forms that fails to disambiguate word senses. This initial phase proposed a core meaning-based analysis, which was a convergence of two contrasting methods (etymology and native speaker's judgement) and approaches (monosemy and polysemy) in lexical semantics, to identify and distinguish polysemes and homographs in the 595 MWL words. This novel analysis resulted in 302 multi-meaning words that were anticipated to pose significant learning and teaching problems because of having polysemes and/or homographs dependent on the various contexts they are found in. These 302 problematic words (whose polysemes and homographs were not explicitly clarified in the MWL), accounting for over half the MWL words, sparked concerns about the increased level of automation in dealing with WSD. Automatic processes of calculating word frequency, which are capable of distinguishing word forms rather than word meanings, only

superficially address the task of WSD and thus fail to solidly underpin the development of corpus-derived lists of words with multiple context-dependent meanings like the MWL. The identification of 302 words also uncovered unique characteristics of semi-technical medical vocabulary, i.e., stretching along the vocabulary continuum and overlapping with other types of vocabulary. These characteristics portray the elusiveness of semi-technical medical vocabulary, placing this type of vocabulary in a grey area to which few lexicographic resources pay due attention.

In response to the lack of polysemy and homography indication in Hsu's (2013) MWL (possibly a consequence of automatic WSD), the study developed SemiMed, one of very few lexicographic resources that exclusively deals with semi-technical medical vocabulary and takes polysemy and homography into consideration. A pilot version SemiMed was developed based on the findings of the previous phase, i.e., featuring 40 sampled words from the MWL's 302 problematic words. The developmental procedure of SemiMed used a novel methodological approach, which consisted of semantic and corpus-based analyses, to address issues of polysemy and homography in dictionaries and wordlists. The semantic analysis adapted theories in lexical semantics (Lakoff's (1987) radial categories, Tyler and Evans's (2004) Principled Polysemy and Cantos and Sanchez's (2001) Lexical Constellation model) to resolve lexicographic constraints on word sense distinction and polysemy and homography presentation. The following corpus-based analysis addressed WSD-related issues in the MWL by considering word meanings in addition to word forms, i.e., quantifying meanings of 40 sampled words in general and medical corpora using the WSD method of one-sense-per-collocation. The outcomes of the twofold analyses were SemiMed's pilot version of 40 semantic networks in the form of lexical constellations (LCs) and its template. In each LC, different general and medical meanings of a word (its polysemes and/or homographs) were structured in a radial format, enabling explicit visualization of their interrelations. The

frequency degree of word meanings was specified through a four-level scale of technicality, created through the corpus-based analysis and incorporated into 40 LCs, to indicate in which context (general, medical or both) a word meaning is more likely to be activated.

To test the practicality and usefulness of SemiMed in comparison with conventional resources, a pilot study was conducted with 18 user participants, who were EFL medical students. They provided positive feedback on two salient features of SemiMed resulting from the semantic and corpus-based analyses, i.e., the radial format and scale of technicality. The radial format was more advantageous than the entry-structured format in conventional general and medical dictionaries in illustrating relationships between polysemy and homography. This non-conventional format demonstrated how different meanings of a word radiate from core meaning(s), allowing participants to understand how general and medical meanings of a word relate to one another in terms of polysemy and homography. The understanding of relationships between polysemy and homography in semi-technical medical vocabulary gained through the radial format facilitated the participants in acquiring different meanings of a semi-technical medical word more easily. The scale of technicality helped participants to navigate the search more efficiently, especially for medical meanings of semi-technical medical vocabulary, than in conventional general dictionaries. It also specified a context in which a particular meaning of semi-technical medical vocabulary tends to be activated. The contextual information that SemiMed offered informed participants about how to arrive at appropriate interpretations of a word in accordance with the contexts it appears in. With these features, SemiMed was therefore considered to have advantages over conventional general and medical dictionaries. Nevertheless, user participants recommended that SemiMed still needed improvements in other aspects, such as visuals, examples and pronunciation.

7.2 Implications

7.2.1 Lexical semantics

The study may contribute to the current literature in lexical semantics because its findings suggest that the methods of distinguishing polysemy and homography and approaches to polysemy are not entirely contradictory but, rather, complementary. It would be worth combining etymology and native speaker judgement in distinguishing polysemy and homography in order to mitigate the disadvantages of each method. It can be assumed from the consistent results of the core meaning-based analysis that etymological evidence is valuable in minimizing the level of subjectivity arising from native speaker judgment alone. When etymological evidence is untraceable by native speakers, their judgement has a role to play in distinguishing polysemy and homography. Moreover, this study also highlighted the possibility of involving non-native speakers in making judgements on the (un)relatedness of meanings and recommended that distinctions between polysemy and homography can rely on non-native speaker judgement in conjunction with that of L1 speakers.

The (un)relatedness of meanings should be judged in relation to a core meaning, a concept which exists in both monosemy and polysemy. Despite their differing views on the mental representations of words, approaches to monosemy and polysemy agree on a core meaning from which word meaning extension stems. This may form a theoretical premise to distinguish polysemy from homography, which, together with Evans's (2005) criteria to determine central meanings, potentially provides principled guidelines to inform native and non-native speaker judgement. Rather than judging the degree of (un)relatedness, native and non-native speakers could examine whether there is a core meaning shared by different meanings of a word to determine if they are polysemous and/or meanings of the word's homograph(s). Both methods embrace subjectivity; however, the latter has the potential to yield more stable and meaningful outcomes because the judgment on (un)relatedness of meaning is

based on defined principles, i.e., judging the meaning against a core meaning using Evans's (2005) criteria. The former relies purely on intuition, which may produce an indefinite continuum of results. Core meaning-based judgment may therefore be promising for achieving consistency in distinguishing polysemy and homography.

7.2.2 Lexicography

The study has implications for practical lexicography, i.e., the compilation of monolingual general and medical dictionaries. The development of SemiMed confirms and suggests that lexicographic practices could benefit from and be underpinned by theories in lexical semantics, more particularly, cognitive lexical semantics. Among cognitive lexical semantic theories, Tyler and Evans's (2004) Principled Polysemy and Lakoff's (1987) radial categories could be adapted to alleviate polysemy-and-homography-related constraints on word sense divisions and presentations in dictionaries. It is suggested that Principled Polysemy should be exploited fully in the lexicographic task of distinguishing word senses to tackle polysemy, especially in irregular cases that exhibit prototypical effects (e.g., *bank* as "financial institution" and as "bank of blood"). The utilization of Principled Polysemy may reduce subjective intuition in performing the task as it provides lexicographers with parameters to distinguish more prototypical from less prototypical senses and distinguish word senses from their instances. Furthermore, this study also reveals that Principled Polysemy applies not only to prepositions, abstract nouns and verbs, but also to adjectives and adverbs. This means Principled Polysemy would assist lexicographers in systematically and consistently capturing senses of words from different parts of speech. Besides Principled Polysemy, Lakoff's (1987) idea of considering words as radial categories should also be acknowledged as an underpinning theory to guide how word senses can be presented to avoid the problems of polysemy and homography in entry-structured dictionaries. Lakoff's idea could be realized in lexicographic practices using Cantos and Sanchez's (2001) model of Lexical Constellations. The flexibly

radial format of LCs can showcase relations (e.g., polysemy and homography) existing among different senses of a word in a more transparent and logical manner than the vertically numbered sense format of dictionary entries. Additionally, feedback from participants in the pilot study highlights the learnability and teachability of this non-conventional format, i.e., facilitating users in navigating and understanding polysemous meanings and homographs of a word, indicating its pedagogical usefulness. All of these suggest that considering the presentation of word senses from the cognitive perspective, i.e., mimicking their representation in the mental lexicon, could pave the way for non-conventional formats that may become an alternative or even a replacement for the sense enumeration format in conventional dictionaries.

7.2.3 Corpus linguistics

The study contributes to WSD in the field of corpus linguistics by suggesting a method to undertake WSD tasks using corpora with due consideration being given to polysemy and homography. Since corpus-based data offer rich, contextual clues such as collocations, which are viewed as valuable sources to disambiguate word senses, a one-sense-per-collocation heuristic could be exploited to underlie a WSD method that addresses polysemy and homograph in corpus linguistics. The method would be best implemented semi-automatically, because semi-automation allows automatically computed results to be triangulated with human evaluations, producing in-depth outcomes. Human involvement (to disambiguate node words based on their collocates) should occur after corpus analysis software has automatically generated a list of collocates. The semi-automatic one-sense-per-collocation method, although requiring significant human investment, could introduce learnable and teachable elements to lexical resources like corpus-based wordlists and dictionaries. This method addresses polysemy-and-homography-related problems in corpus-derived lists of frequently occurring words by considering the frequency of word meanings and forms. In this way, word form frequency-based lists are pedagogically improved, as semantic information, which is usually

scarce in these lists, could be enriched and sufficiently presented to learners (and teachers) to be learned (and taught). This method is also helpful for the compilation of corpus-based dictionaries in the sense that it offers evidence-based frequency information which can be shown together with word senses. Although the indication of meaning frequency is not a new feature, cross-context meaning frequency (e.g., the frequency degree of word meanings in general and specialized contexts) could be valuable information that should be made available for words with contextually varied meanings, such as the semi-technical vocabulary in dictionaries.

7.2.4 Learning and teaching semi-technical medical vocabulary

First, the identification of 302 potentially confusing words in Hsu's (2013) MWL casts some doubt on corpus-derived lists of semi-technical medical words. That these lists tend to enumerate frequently occurring words, which are usually multi-meaning, without explanatory provision of word meanings, casts doubt on the pedagogical benefits of the lists in their current format of only presenting word forms and frequency statistics. Their development appears to rely excessively on automatic WSD analyses that exclude word meanings, posing concerns over the validity and reliability of these lists. Corpus-derived semi-technical medical wordlists like the MWL should thus be used in the classroom with caution and, where necessary, adapted (e.g., by searching and incorporating word meanings in teaching instructions) to augment their pedagogical effectiveness. More importantly, comprehensive re-evaluation with a focus on advancing WSD processes to account for word meanings is recommended to address the root of problems and bring about significant pedagogical improvement to these lists.

Another issue that emerges from the identification of the 302 words is that contemporary attempts to establish clear-cut boundaries around semi-technical medical vocabulary are only marginally successful. Due to their elusive nature, it is suggested that identifying specific words with differing meanings in general and medical contexts is

pedagogically sound rather than striving for the taxonomy that categorizes semi-technical medical vocabulary into groups distinct from other types of vocabulary. Core meaning-based analysis is a potential method to identify such kinds of semi-technical medical vocabulary. The outcome of this method is twofold: (a) a particular set of targeted semi-technical medical words that deserves learning and teaching attention and (b) their core meanings. The core meanings of the 302 MWL words, for example (see Appendix 1), albeit not yet a full-fledged lexical resource, already have pedagogical implications, especially in facilitating the incorporation of polysemy and homography instruction into semi-technical medical vocabulary teaching practices (as previously detailed in 4.5).

Secondly, SemiMed (the pilot version of 40 LCs) provides support for the idea of developing a resource of semi-technical medical vocabulary that fully accounts for polysemy and homography. Since they are situated in the grey area between technical and general vocabulary, semi-technical medical words have yet to be well treated in dictionaries, i.e., either their general or medical meanings are the focus of general and medical dictionaries, but not both meanings together. More resources like SemiMed that exclusively concentrate on semi-technical medical vocabulary would therefore be welcome to help learners acquire this type of vocabulary more easily. An interdisciplinary approach in which knowledge from lexical semantics and WSD in corpus linguistics underpins the development process is recommended to ensure that the treatment of polysemy and homography is theory-based.

SemiMed also suggests a new method of learning and teaching semi-technical medical vocabulary. Explicit instructions on polysemy and homography when semi-technical medical vocabulary is introduced to learners could be made available. Learners could be encouraged to learn about polysemy and homography to become more aware of and familiar with a word's polysemes and/or homographs. In this way, they can equip themselves with sufficient semantic knowledge of the word and there is less potential for confusion in understanding and

interpreting its different meanings in general and medical contexts. Equally importantly, the general and medical meanings of semi-technical medical vocabulary should be presented, learned and taught in parallel, as knowing general meanings (and how they are related to medical meanings) can assist the acquisition of medical meanings and vice versa. This method of learning and teaching semi-technical medical vocabulary may, like SemiMed and its template, be potentially transferrable to specializations other than medicine.

7.3 Recommendations for future research

To address the methodological limitations (discussed in 3.5), it is suggested that future research should increase the number of evaluators, plus the involvement of non-native speakers, to see whether the reliability remains stable. Future lexical research could also consider the remaining words (262 of 302 potentially problematic words) in the MWL and their core meanings that this study was unable to address to develop a full version of SemiMed. Further work is also needed to scale up SemiMed by increasing the number of words this resource features to cover the large number of semi-technical medical words users comprehensively need to look up. There is also abundant scope for further research on incorporating aspects other than word meanings (e.g., collocations, pronunciation, examples, etc.) into the full version of SemiMed.

Moreover, as SemiMed's format is not well suited to printed resources, future studies intending to utilize this format in printed versions would need to optimize the bubble-shaped layout to ensure a neat, space-saving presentation. Also, the e-version of SemiMed on the H5P platform in the pilot study partly requires manual manipulation to search for words. In future investigations into e-resources that reuse SemiMed's format, it might be worth creating advanced, automatic functions to accelerate searching. Lastly, future lexicographic studies on other types of vocabulary intending to reduplicate the methods of developing SemiMed should

examine the feasibility of these methods in working with grammatical words and multi-word items.

To conclude, SemiMed has provided a promising direction for the unresolved issues of polysemy and homography in current dictionaries and wordlists. This has been the first among very few resources that (a) exclusively deal with semi-technical medical vocabulary, (b) comprehensively incorporate theories in cognitive lexical semantics into the developmental procedure to make its content and structure efficient, and (c) thoroughly examine word meanings in corpus-based WSD to transfer word meaning frequency-focused results into a learnable and teachable lexical resource. It is hoped that future research that considers the above-mentioned recommendations will develop a well-rounded version of SemiMed that better serves users.

REFERENCES

A. Dictionaries

Cambridge Dictionary. <https://dictionary.cambridge.org/>

Collins English Dictionary. HarperCollins Publishers, UK.

Longman Dictionary of Contemporary English. Pearson Education Ltd, UK.

Merriam-Webster Medical Dictionary. <https://www.merriam-webster.com/medical>.

Online Etymology Dictionary. <https://www.etymonline.com/>

Oxford English Dictionary. <https://www.oed.com>.

Oxford Dictionary of English. Oxford University Press, UK.

B. Other literature

Abed, S. A., Tiun, S., & Omar, N. (2016). Word sense disambiguation in evolutionary manner. *Connection Science*, 28(3), 226-241.

Agirre, E., & Edmonds, P. (2006). *Word sense disambiguation: Algorithms and applications*. Springer.

Aitchison, J. (2003). *Words in the mind: An introduction to the mental lexicon*. Blackwell.

Apresjan, J. D. (1974). Regular polysemy. *Linguistics*, 14(2), 5-32.

Anderson, R. C., & Nagy, W. E. (1991). Word meanings. In R. Barr, M. L. Kamil, P. B. Mosenthal & P. D. Pearson (Eds.), *Handbook of reading research* (pp. 690-794). Longman Publishing Group.

Atkins, B. T. S. (1991). Building a lexicon: The contribution of lexicography. *International Journal of Lexicography*, 4(3), 167-204.

Atkins, S., Rundell, M., & Sato, H. (2003). The contribution of FrameNet to practical lexicography. *International Journal of Lexicography*, 16(3), 333-357.

- Atkins, S. (2008). Theoretical lexicography and its relation to dictionary-making. In T. Fontenelle (Ed.), *Practical lexicography: A reader* (pp. 31-50). Oxford University Press.
- Atkins, B. S., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.
- Ayto, J. (1983). On specifying meaning. In R. R. K. Hartmann (Ed.), *Lexicography: Principles and practice* (pp. 89-98). Academic Press.
- Baker, M. (1988). Sub-technical vocabulary and the ESP teacher: An analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language*, 4(2), 91-105.
- Başkaya, O., & Jurgens, D. (2016). Semi-supervised learning with induced word senses for state of the art word sense disambiguation. *Journal of Artificial Intelligence Research*, 55, 1025-1058.
- Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253-279.
- Bensoussan, M., & Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, 7(1), 15-32.
- Béjoint, H. (1990). Monosemy and the Dictionary. In *BudaLEX'88 Proceedings of the 3rd International EURALEX Congress* (pp. 11-26). EURALEX.
- Béjoint, H. (2000). *Modern lexicography: An introduction*. Oxford University Press.
- Béjoint, H. (2010). *The lexicography of English: From origins to present*. Oxford University Press.
- Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, 93(3), 498-520.
- Boonmoh, A., Singhasiri, W., & Hull, J. (2006). Problems using electronic dictionaries to translate Thai written essays into English. *rEFLECTIONS*, 8, 8-21.

- Bréal, M. (1900). *Semantics: Studies in the science of meaning*. Heinemann.
- Brezina, V., & Gablasova, D. (2017a). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, 36(1), 1-22.
- Brezina, V., & Gablasova, D. (2017b). How to produce vocabulary lists? Issues of definition, selection and pedagogical aims. A response to Gabriele Stein. *Applied Linguistics*, 38(5), 764-767.
- Browne, C. (2014). A new general service list: The better mousetrap we've been looking for. *Vocabulary Learning and Instruction*, 3(1), 1-10.
- Brugman, C., & Lakoff, G. (1988). Cognitive topology and lexical networks. In S. Small, G. Cottrell & M. Tanenhaus (Eds.), *Lexical ambiguity resolution* (pp. 477-508). Morgan Kaufmann.
- Cantos, P., & Sanchez, A. (2001). Lexical constellations: What collocates fail to tell. *International Journal of Corpus Linguistics*, 6(2), 199-228.
- Cantos, P., Sanchez, A., & Almela, M. (2009). An attempt to formalize word sense disambiguation: Maximizing efficiency by minimizing computational cost. *RESLA*, 22, 77-88.
- Caramazza, A., & Grober, E. (1976). *Polysemy and the structure of the subjective lexicon* [Paper presentation]. Georgetown University roundtable on languages and linguistics. Semantics: Theory and application, USA.
- Carston, R. (2021). Polysemy: Pragmatics and sense conventions. *Mind & Language*, 36(1), 108-133.
- Chabner, D. E. (2020). *The language of medicine*. Elsevier US.
- Charters, E. (2003). The use of think-aloud methods in qualitative research: An introduction to think-aloud methods. *Brock Education Journal*, 12(2), 68-82.

- Chen, L., Liu, H., & Friedman, C. (2005). Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21(2), 248-256.
- Chen, P., Ding, W., Bowes, C., & Brown, D. (2009). A fully unsupervised word sense disambiguation method using dependency knowledge. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 28-36). Association for Computational Linguistics.
- Chen, Q., & Ge, G.-C. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes*, 26(4), 502-514.
- Cohen, A., Glasman, H., Rosenbaum-Cohen, P. R., Ferrara, J., & Fine, J. (1988). Reading English for specialized purposes: Discourse analysis and the use of student informants. In P. Carrell, J. Devine & D. E. Eskey (Eds.), *Interactive approaches to second language reading* (pp. 152-167). Cambridge University Press.
- Cohen, L., Manion, L., & Morrison, K. (2011). Observation. In L. Cohen, L. Manion & K. Morrison (Eds.), *Research methods in education* (pp. 456-475). Taylor & Francis Group.
- Cohn, T. (2003). Performance metrics for word sense disambiguation. In *Proceedings of the Australasian Language Technology Workshop* (pp. 86-93). ALTA.
- Copestake, A., & Briscoe, T. (1995). Semi-productive polysemy and sense extension. *Journal of Semantics*, 12(1), 15-67.
- Covington, M. A., Grosz, B. J., & Pereira, F. C. (1994). *Natural language processing for Prolog programmers*. Prentice Hall.
- Cowan, J. R. (1974). Lexical and syntactic research for the design of EFL reading materials. *TESOL Quarterly*, 8(4), 389-399.

- Cowie, A. P. (1988). Stable and creative aspects of vocabulary use. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 126-139). Longman.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213-238.
- Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.
- Croft, W., & Cruse, D. A. (2004). *Cognitive linguistics*. Cambridge University Press.
- Cruse, A. (1986). *Lexical semantics*. Cambridge University Press.
- Cruse, A. (1992). Monosemy vs. polysemy. Review of Ruhl (1989). *Linguistics*, 30, 577-599.
- Cruse, D. A. (2000). *Meaning in language: An introduction to semantics and pragmatics*. Oxford University Press.
- Csábi, S. (2002). Polysemous words, idioms and conceptual metaphors: Cognitive linguistics and lexicography. In *Proceedings of the 10th EURALEX International Congress* (pp. 249-254). EURALEX.
- Dagan, I., & Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4), 563-596.
- Dalpanagioti, T. (2018). Corpus-based cognitive lexicography: Insights into the meaning and use of the verb stagger. In *Proceedings of the 8th EURALEX International Congress: Lexicography in global contexts* (pp. 649-662). EURALEX.
- Dang, T. N. Y. (2019). Corpus-based word lists in second language vocabulary research, learning, and teaching. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 288-303). Routledge.
- Dash, N. S. (2012). Polysemy and homonymy: A conceptual labyrinth. In *Proceedings of IndoWordNet Workshop* (pp. 1-7). Indian Institute of Technology.
- Dominiek, S. (1998). What linguists can and can't tell you about the human mind: A reply to Croft. *Cognitive Linguistics*, 9, 361-478.

- Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, 28(3), 157-169.
- Edmonds, P. (2006). Disambiguation, lexical. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (pp. 607-623). Elsevier.
- Evans, V. (2004). *The structure of time: Language, meaning and temporal cognition*. John Benjamins.
- Evans, V. (2005). The meaning of time: Polysemy, the lexicon and conceptual structure. *Journal of Linguistics*, 41(1), 33-75.
- Evans, V., & Green, M. (2006). *Cognitive linguistics: An introduction*. Edinburgh University Press.
- Farrell, P. (1990). *Vocabulary in ESP: A lexical analysis of the English of electronics and a study of semi-technical vocabulary*. (CLCS Occasional Paper No. 25). Dublin, Ireland: Trinity College, Centre for Language and Communication Studies.
- Flaherty, M. G. (1999). *A watched pot: How we experience time*. New York University Press.
- Flowerdew, J. (1993). Concordancing as a tool in course design. *System*, 21(2), 231-244.
- Frantzen, D. (2003). Factors affecting how second language Spanish students derive meaning from context. *The Modern Language Journal*, 87(2), 168-199.
- Fraser, S. (2007). Providing ESP learners with the vocabulary they need: Corpora and the creation of specialized word lists. *Hiroshima Studies in Language and Language Education*, 10, 127-143.
- Fraser, S. (2009). Breaking down the divisions between general, academic and technical vocabulary: The establishment of a single, discipline-based word list for ESP learners. *Hiroshima Studies in Language and Language Education*, 12, 151-167.
- Fraser, S. (2012). Factors affecting the learnability of technical vocabulary: Finding from a specialized corpus. *Hiroshima Studies in Language and Language Education*, 15, 123-142.

- Frisson, S., & Pickering, M. J. (2001). Obtaining a figurative interpretation of a word: Support for underspecification. *Metaphor and Symbol, 16*(3-4), 149-171.
- Gale, W. A., Church, K. W., & Yarowsky, D. (1992a). Work on statistical methods for word sense disambiguation. In *Proceedings of the AAAI Fall Symposium on probabilistic approaches to natural language* (pp. 54-60). Association for the Advancement of Artificial Intelligence.
- Gale, W. A., Church, K. W., & Yarowsky, D. (1992b). One sense per discourse. In *Proceedings of the DARPA Speech and Natural Language Workshop* (pp. 233-237). DARPA.
- Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics, 28*(2), 241-265.
- Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305-327.
- Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research, 19*(6), 645-666.
- Geeraerts, D. (1990). The lexicographical treatment of prototypical polysemy. In S. Tsohatzidis (Ed.), *Meanings and prototypes: Studies in linguistic categorization* (pp. 195-210). Routledge.
- Geeraerts, D. (2001). The definitional practice of dictionaries and the cognitive semantic conception of polysemy. *Lexicographica, 17*, 6-21.
- Geeraerts, D. (2007). Lexicography. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford handbook of cognitive linguistics* (pp. 1160-1174). Oxford University Press.
- Geeraerts, D. (2006). *Words and other wonders: Papers on lexical and semantic topics*. De Gruyter.
- Geeraerts, D. (2010). *Theories of lexical semantics*. Oxford University Press.

- Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly*, 25(3), 375-406.
- Graham, A. (2008). *The effects of homography on computer-generated high frequency word lists* [Master's thesis, Brigham Young University].
<https://www.proquest.com/docview/2504799565?pq-origsite=gscholar&fromopenview=true>
- Grefenstette, G., & Hanks, P. (2023). Competing Views of Word Meaning: Word Embeddings and Word Senses. *International Journal of Lexicography*, 1-9.
- Gries, S. T. (2015). Polysemy. In E. Dąbrowska & D. S. Divjak (Eds.), *Handbook of cognitive linguistics* (pp. 472–490). De Gruyter.
- Hausmann, F. J., & Wiegand, H. E. (1989). Component parts and structures of general monolingual dictionaries: A survey. In F. J. Hausmann, O. Reichmann, H. E. Wiegand & L. Zgusta (Eds.), *Wörterbücher: Ein internationales Handbuch zur Lexikographie* (pp. 328-360). De Gruyter.
- Hanks, P. (1990). Evidence and intuition in lexicography. In J. Tomaszczyk & B. Lewandowska-Tomaszczyk (Eds.), *Meaning and lexicography* (pp. 31-42). John Benjamins.
- Hanks, P. (1994). Linguistic norms and pragmatic exploitations or, why lexicographers need prototype theory, and vice versa. In F. Kiefer, G. Kiss & J. Pajzs (Eds.), *Papers in computational lexicography: Complex '94* (pp. 89-113). Hungarian Academy of Sciences.
- Hanks, P. (2000). Do word meanings exist? *Computers and the Humanities*, 34(1/2), 205-215.
- Hanks, P. (2002). Mapping meaning onto use. In M. Corréard, *Lexicography and Natural Language Proceeding: A festschrift in honour of B. T. S. Atkins* (pp. 156-198). EURALEX.
- Hanks, P. (2008). The lexicographical legacy of John Sinclair. *International Journal of Lexicography*, 21(3), 219-229.

- Higgins, J. J. (1966). Hard facts. *ELT Journal*, 21(1), 55-60.
- Hoey, M. (2012). *Lexical priming: A new theory of words and language*. Routledge.
- Hsu, W. (2013). Bridging the vocabulary gap for EFL medical undergraduates: The establishment of a medical word list. *Language Teaching Research*, 17(4), 454-484.
- Huizhong, Y. (1986). A new technique for identifying scientific/technical terms and describing science texts: (An Interim Report). *Literary and Linguistic Computing*, 1(2), 93-103.
- Hunston, S. (2005). Corpus linguistics. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (pp. 234-248). Elsevier.
- Hutchins, J. (1999). Warren Weaver memorandum: 50th anniversary of machine translation. *MT News International*, (22), 5-6.
- Hyland, K., & Tse, P. (2007). Is there an “academic vocabulary”? *TESOL Quarterly*, 41(2), 235-253.
- Hyland, K., & Tse, P. (2009). Academic lexis and disciplinary practice: Corpus evidence for specificity. *International Journal of English Studies*, 9(2), 111-129.
- Ide, N., & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1), 1-40.
- Ide, N., & Wilks, Y. (2006). Making sense about sense. In E. Agirre & P. Edmonds (Eds.), *Word sense disambiguation: Algorithms and applications* (pp. 47-74). Springer.
- Jackson, H. (2013). *Words and their meaning*. Routledge.
- Janssen, A. J. M. (2003). Monosemy versus polysemy. In H. Cuyckens, R. Dirven & J. Taylor (Eds.), *Cognitive approaches to lexical semantics* (pp. 93-122). De Gruyter.
- Jorgensen, J. C. (1990). The psychological reality of word senses. *Journal of Psycholinguistic Research*, 19(3), 167-190.
- Kempson, R. M. (1977). *Semantic theory*. Cambridge University Press.

- Kilgarriff, A. (1992). Dictionary word sense distinctions: An enquiry into their nature. *Computers and the Humanities*, 26, 365-387.
- Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, 31(2), 91-113.
- Kilgarriff, A. (2007). Word senses. In E. Agirre & P. Edmonds (Eds.), *Word sense disambiguation: Algorithms and application* (pp. 29-46). Springer.
- Kilgarriff, A. (2013). Using corpora as data sources for dictionaries. In H. Jackson (Ed.), *The Bloomsbury companion to lexicography* (pp. 77-96). Bloomsbury.
- Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J. B., Khalil, S., Kokkinakis, S. J., Lew, R., Sharoff, S., Vadlapudi, R., & Volodina, E. (2014). Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48, 121-163.
- Kilgarriff, A., & Rosenzweig, J. (2000). Framework and results for English SENSEVAL. *Computers and the Humanities*, 34(1), 15-48.
- Klein, D. E., & Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language*, 45(2), 259-282.
- Klepousniotou, E. (2002). The processing of lexical ambiguity: Homonymy and polysemy in the mental lexicon. *Brain and Language*, 81(1-3), 205-223.
- Klepousniotou, E., Titone, D., & Romero, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(6), 1534-1543.
- Knowles, G., & Mohd Don, Z. (2004). The notion of a "lemma": Headwords, roots and lexical sets. *International Journal of Corpus Linguistics*, 9(1), 69-81.
- Kwong, O. Y. (2013). *New perspectives on computational and cognitive strategies for word sense disambiguation*. Springer.

- Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal about the mind*. University of Chicago.
- Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1), 15-30.
- Le, C. N. N., & Miller, J. (2023). A core meaning-based analysis of English semi-technical vocabulary in the medical field. *English for Specific Purposes*, 70, 252-266.
- Leech, G. (1974). *Semantics*. Penguin.
- Lehrer, A. (1974). *Semantic fields and lexical structure*. North-Holland.
- Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual Conference on systems documentation* (pp. 24-26). SIGDOC.
- Levin, B. (1991). Building a lexicon: The contribution of linguistics. *International Journal of Lexicography*, 4(3), 205-226.
- Levow, G. A. (1997). *Corpus-based techniques for word sense disambiguation* (Technical Report No. AIM-1637). MIT AI Lab.
- Lew, R. (2013). Identifying, ordering and defining senses. In H. Jackson (Ed.), *The Bloomsbury companion to lexicography* (pp. 284-303). Bloomsbury.
- Lexical Computing CZ s.r.o. 2014-2021. *Sketch Engine for Language Learning*. Sketch Engine. <https://skell.sketchengine.eu/#home?lang=en>
- Li, E. S.-L., & Pemberton, R. (1994). An investigation of students' knowledge of academic and subtechnical vocabulary. In J. Flowerdew & A. K. K. Tong (Eds.), *Entering text* (pp. 183-196). The Hong Kong University of Science and Technology.
- L'Homme, M.-C. (2020). Lexical semantics for terminology. In *Proceedings of the 9th EURALEX International Congress: Lexicography for inclusion* (pp. 415-426). EURALEX.

- Lüdeling, A., & Kytö, M. (2008). *Corpus linguistics: An international handbook*. De Gruyter.
- Lyons, J. (1968). *Introduction to theoretical linguistics* (Vol. 510). Cambridge University Press.
- Lyon, J. (1977). *Semantics*. Cambridge University Press.
- Mahpeykar, N., & Tyler, A. (2015). A principled cognitive linguistics account of English phrasal verbs with up and out. *Language and Cognition*, 7(1), 1-35.
- Màrquez, L., Escudero, G., Martínez, D., & Rigau, G. (2006). Supervised corpus-based methods for WSD. In E. Agirre & P. Edmonds (Eds.), *Word sense disambiguation: Algorithms and applications* (pp. 167-216). Springer.
- Martínez, D., & Agirre, E. (2000). One sense per collocation and genre/topic variations. In *Proceedings of the Joint SIGDAT Conference on empirical methods in natural language processing and very large corpora (EMNLP/VLC)* (pp. 207-215). SIGDAT.
- McCarthy, D. (2006). Relating WordNet senses for word sense disambiguation. In *Proceedings of the ACL Workshop on making sense of sense: Bringing psycholinguistics and computational linguistics together* (pp. 17-24). Association for Computational Linguistics.
- Mel'čuk, I. (1988). Semantic description of lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria. *International Journal of Lexicography*, 1(3), 165-188.
- Mihalcea, R. (2006). Knowledge-based methods for WSD. In E. Agirre & P. Edmonds (Eds.), *Word sense disambiguation: Algorithms and applications* (pp. 107-131). Springer.
- Mihalcea, R. (2007). Using Wikipedia for automatic word sense disambiguation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference* (pp. 196-203). Association for Computational Linguistics.

- Mihalcea, R., Chklovski, T., & Kilgarriff, A. (2004). The Senseval-3 English lexical sample task. In *Proceedings of SENSEVAL-3, the 3rd International Workshop on the evaluation of systems for the semantic analysis of text* (pp. 25-28). Association for Computational Linguistics.
- Miller, D., & Biber, D. (2015). Evaluating reliability in quantitative vocabulary studies: The influence of corpus design and composition. *International Journal of Corpus Linguistics*, 20(1), 30-53.
- Miller, G. A. (1986). Dictionaries in the Mind. *Language and Cognitive Processes*, 1(3), 171-185.
- Moerdijk, F. (2003). The codification of semantic information. In P. Sterkenburg (Ed.), *A practical guide to lexicography* (pp. 273-296). John Benjamins.
- Moon, R. (1987). The analysis of meaning. In J. Sinclair (Ed.), *Looking up: An account of the COBUILD project in Lexical Computing* (pp. 86-103). Collins.
- Moon, R. (2004). On specifying metaphor: An idea and its implementation. *International Journal of Lexicography*, 17(2), 195-222.
- Murphy, M. L. (2010). *Lexical meaning*. Cambridge University Press.
- Nation, P. (2013). *Learning vocabulary in another language*. Cambridge University Press.
- Nation, P. (2018). The BNC/COCA word family lists. <https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-lists>
- Nation, P., Coxhead, A., Chung, T., & Quero, B. (2016). Specialized word lists. In P. Nation (Ed), *Making and using word lists for language learning and testing* (pp. 145-151). John Benjamins.
- Nation, P., & Heatley, A. (2005). RANGE [computer software]. <https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs>

- Nation, P., & Parent, K. (2016). Homoforms and polysemes. In P. Nation (Ed), *Making and using word lists for language learning and testing* (pp. 41-53). John Benjamins.
- Nesi, H., & Haill, R. (2002). A study of dictionary use by international students at a British university. *International Journal of Lexicography*, 15(4), 277-305.
- Nerlich, B., Todd, Z., Herman, V., & Clarke, D. D. (Eds.). (2003). *Polysemy: Flexible patterns of meaning in mind and language* (Vol. 142). De Gruyter.
- Ostermann, C. (2015). *Cognitive lexicography: A new approach to lexicography making use of cognitive semantics*. De Gruyter.
- Palmer, F. R. (1995). *Semantics*. Cambridge University Press.
- Panman, O. (1982). Homonymy and polysemy. *Lingua*, 58(1-2), 105-136.
- Paquot, M. (2007). Towards a productively-oriented academic word list. In J. Walinski, K. Kredens & S. Gozdz-Roszkowski (Eds.), *Practical applications in language and computers* (pp. 127-140). Peter Lang.
- Parent, K. (2012). The most frequent English homonyms. *RELC Journal*, 43(1), 69-81.
- Pedersen, T. (2006). Unsupervised corpus-based methods for WSD. In E. Agirre & P. Edmonds (Eds.), *Word sense disambiguation: Algorithms and applications* (pp. 133-166). Springer.
- Pellicer-Sánchez, A. (2019). Learning single words vs. multiword items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 158-173). Routledge.
- Perez, M. (2013). *Identification and analysis of the specialized vocabulary of British Law Reports: A corpus-driven study of this legal genre at the core of common law legal systems* [Doctoral dissertation, University of Murcia].
https://www.tdx.cat/handle/10803/128621?fbclid=IwAR29vxL15QEpAcg19TKI6ExdP_9GJ92NunxxFLyTHcJYvz4aOKaKcyv2i10#page=1
- Peters, P., & Fernández, T. (2013). The lexical needs of ESP students in a professional field. *English for Specific Purposes*, 32(4), 236-247.

- Pustejovsky, J. (1995). *The generative lexicon*. MIT Press.
- Quero, B. & Coxhead, A. (2018). Using a corpus-based approach to select medical vocabulary for an ESP course: The case for high-frequency vocabulary. In Y. Kirkgöz & K. Dikilitaş (Eds.). *Key issues in English for specific purposes in higher education* (pp. 51-75). Springer.
- Ravin, Y., & Leacock, C. (2000). *Polysemy: Theoretical and computational approaches*. Oxford University Press.
- Resnik, P. (2006). WSD in NLP applications. In E. Agirre & P. Edmonds (Eds.), *Word sense disambiguation: Algorithms and applications* (pp. 299-338). Springer.
- Rizzo, C., & Sanchez, A. (2010). Building new meanings in technical English from the perspective of the lexical constellation model. *Ibérica*, 20, 107-126.
- Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, 4(3), 328-350.
- Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3), 192-233.
- Rosch, E. (1977). Human categorization. In N. Warren (Ed.), *Studies in cross-linguistic psychology* (pp. 1-49). Academic Press.
- Rosch, E. (2004). Principles of categorization. In B. Aarts, D. Denison, E. Keizer & G. Popova (Eds.), *Fuzzy grammar: A reader* (pp. 91-108). Oxford University Press.
- Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4), 573-605.
- Ruhl, C. (1989). *On monosemy: A study in linguistic semantics*. State University of New York Press.
- Ruhl, C. (2002). Data, comprehensiveness, monosemy. *Studies in Functional and Structural Linguistics*, 171-190.

- Rundell, M. (1988). Changing the rules: Why the monolingual learner's dictionary should move away from the native-speaker tradition. In M. Snell-Hornby (Ed.), *ZuriLEX 86 Proceedings* (pp. 127-137). A. Francke.
- Rundell, M. (2012). It works in practice but will it work in theory? In *Proceedings of the 15th EURALEX International Congress* (pp. 47-92). EURALEX.
- Saussure, F. (2011). *Course in general linguistics*. Columbia University Press.
- Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, 47(4), 484-503.
- Scholfield, P. (1999). Dictionary use in reception. *International Journal of Lexicography*, 12(1), 13-34.
- Schuemie, M. J., Kors, J. A., & Mons, B. (2005). Word sense disambiguation in the biomedical domain: An overview. *Journal of Computational Biology*, 12(5), 554-565.
- Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1), 97-123.
- Shaw, P. (1991). Science research students' composing processes. *English for Specific Purposes*, 10(3), 189-206.
- Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. Routledge.
- Skoufaki, S., & Petric, B. (2021). Exploring polysemy in the Academic Vocabulary List: A lexicographic approach. *Journal of English for Academic Purposes*, 54, 1-14.
- Stock, P. F. (2008). Polysemy. In T. Fontenelle (Ed.), *Practical lexicography: A reader* (pp. 153-160). Oxford University Press.
- Stubbs, M. (2002). *Words and phrases: Corpus studies of lexical semantics*. John Wiley & Sons.

- Thompson, P., & Alzeer, S. N. (2019). A survey of issues, practices and views related to corpus-based word lists for English language teaching and learning. *International Journal of Applied Linguistics and English Literature*, 8(6), 43-53.
- Thurston, J., & Candlin, C. N. (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes*, 17(3), 267-280.
- Todd, R. W. (2017). An opaque engineering word list: Which words should a teacher focus on? *English for Specific Purposes*, 45, 31-39.
- Tyler, A., & Evans, V. (2001). Reconsidering prepositional polysemy networks: The case of over. *Language*, 77(4), 724–765
- Tyler, A., & Evans, V. (2003a). *The semantics of English prepositions: Spatial scenes, embodied meaning, and cognition*. Cambridge University Press.
- Tyler, A., & Evans, V. (2003b). Reconsidering prepositional polysemy networks: The case of over. *Trends in Linguistics Studies and Monographs*, 142, 95-160.
- Tyler, A., & Evans, V. (2004). Rethinking English ‘prepositions of movement’: The case of to and through. *Belgian Journal of Linguistics*, 18(1), 247-270.
- Ullmann, S. (1962). *Semantics: An introduction to the science of meaning*. Basil Blackwell.
- Van der Eijk, P., Alejandro, O., & Florenza, M. (1995). Lexical semantics and lexicographic sense distinction. *International Journal of Lexicography*, 8(1), 1-27.
- Van der Gucht, F., Willems, K., & De Cuypere, L. (2007). The iconicity of embodied meaning. Polysemy of spatial prepositions in the cognitive framework. *Language Sciences*, 29(6), 733-754.
- Van der Meer, G. (1997). Four English learner’s dictionaries and their treatment of figurative meanings. *English Studies*, 78, 556-571.
- Van der Meer, G. (1999). Metaphors and dictionaries: The morass of meaning or how to get two ideas for one. *International Journal of Lexicography*, 12(3), 195-208.

- Van der Meer, G. (2004). On defining: Polysemy, core meanings and ‘great simplicity’. In G. Williams & S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 807-815). EURALEX.
- Véronis, J. (2001). *Sense tagging: Does it make sense?* [Paper presentation]. Corpus Linguistics 2001 Conference, UK.
- Vicente, A., & Falkum, I. L. (2017). Polysemy. *Oxford research encyclopedia of linguistics*. <https://doi.org/10.1093/acrefore/9780199384655.013.325>.
- Vidhu Bhala, R. & Abirami, S. (2014). Trends in word sense disambiguation. *Artificial Intelligence Review*, 42(2), 159-171.
- Walter, E. (2010). Using corpora to write dictionaries. In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 428-443). Routledge.
- Wang, J., Liang, S. L., & Ge, G. C. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, 27(4), 442-458.
- Wang, K. W., & Nation, P. (2004). Word meaning in academic English: Homography in the academic word list. *Applied Linguistics*, 25(3), 291-314.
- Watson-Todd, R. (2017). An opaque engineering word list: Which words should a teacher focus on? *English for Specific Purposes*, 45, 31-39.
- Weaver, W. (1955). Translation. In A. D. Booth & W. N. Locke (Eds.), *Machine translation of languages: Fourteen essays* (pp. 15-23). MIT Press.
- West, M. (1953). *A general service list of English words*. Longman.
- Widdowson, H. (2003). *Defining issues in English language teaching*. Oxford University Press.
- Wilks, Y., Fass, D., Guo, C., McDonald, J. E., Plate, T., & Slator, B. M. (1990). Providing machine tractable dictionary tools. *Machine Translation*, 5, 99-154.

- Winkler, B. (2001). English learners' dictionaries on CD-ROM as reference and language learning tools. *ReCALL*, 13(2), 191-205.
- Yamamoto, Y. (2014). Multidimensional vocabulary acquisition through deliberate vocabulary list learning. *System*, 42, 232-243.
- Yarowsky, D. (1993). One sense per collocation. In *Proceedings of the ARPA Human Language Technology Workshop* (pp. 266-271). ARPA.
- Yarowsky, D. (1995). Unsupervised word sense disambiguation rivalling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics* (pp. 189-196). Association for Computational Linguistics.
- Yu, A., & Trainin, G. (2022). A meta-analysis examining technology-assisted L2 vocabulary learning. *ReCALL*, 34(2), 235-252.
- Zgusta, L. (1992). The Czech-Chinese dictionary and the theory of lexicography. *International Journal of Lexicography*, 5(2), 85-128.
- Zhou, Q., & Meng, Y. (2019). Combination of semantic relatedness with supervised method for word sense disambiguation. In *2019 International Conference on Asian Language Processing* (pp. 142-147). Institute of Electrical and Electronics Engineers.

APPENDICES

APPENDIX 1. Core meanings of 302 potentially confusing semi-technical medical words organized by frequency in Hsu's (2013) Medical Word List

<p>Diagnosis (Frequency: 23,342)</p> <p><i>Core meaning: Identification</i></p>	<p>Dominant (Frequency: 1,618)</p> <p><i>Core meaning: Commanding</i></p>
<p>Tumour (Frequency: 23,232)</p> <p><i>Core meaning: Swelling</i></p>	<p>Resolve (Noun) (Frequency: 1,606)</p> <p><i>Core meaning: Determination to do something</i></p> <p>Resolve (Verb)</p> <p><i>Core meaning 1: To transform</i></p> <p><i>Core meaning 2: To break into separate parts</i></p> <p><i>Core meaning 3: To bring to resolution</i></p>
<p>Renal (Frequency: 23,129)</p> <p><i>Core meaning: (Relating to) the kidneys</i></p>	<p>Eliminate (Frequency: 1,605)</p> <p><i>Core meaning: To remove</i></p>
<p>Syndrome (Frequency: 18,037)</p> <p><i>Core meaning: Combined symptoms or behaviours</i></p>	<p>Transcript (Frequency: 1,587)</p> <p><i>Core meaning: A written copy</i></p>
<p>Liver (Frequency: 16,579)</p> <p><i>Core meaning: (Relating to) the liver</i></p>	<p>Premature (Frequency: 1,579)</p> <p><i>Core meaning: Too early</i></p>
<p>Transplant (Frequency: 15,434)</p> <p><i>Core meaning: To move/A subject or process of moving something or someone from one place to another</i></p>	<p>Lobe (Frequency: 1,571)</p> <p><i>Core meaning: A division or part of a larger structure</i></p>
<p>Lesion (Frequency: 14,033)</p> <p><i>Core meaning: Relating to injury or disease</i></p>	<p>Median (Frequency: 1,569)</p> <p><i>Core meaning: Relating to the midpoint</i></p>

<p>Acute (Frequency: 13,801)</p> <p><i>Core meaning 1: Extreme</i></p> <p><i>Core meaning 2: Sharp</i></p>	<p>Susceptible (Frequency: 1,565)</p> <p><i>Core meaning: Capable of being affected by something</i></p>
<p>Chronic (Frequency: 12,465)</p> <p><i>Core meaning 1: Long lasting</i></p> <p><i>Core meaning 2: Bad</i></p>	<p>Interstitial (Frequency: 1,560)</p> <p><i>Core meaning: In-between</i></p>
<p>Primary (Frequency: 12,165)</p> <p><i>Core meaning: First and original</i></p>	<p>Placenta (Frequency: 1,552)</p> <p><i>Core meaning: The part in a plant/ human where the ovules/ foetus are attached</i></p>
<p>Disorder (Frequency: 11,877)</p> <p><i>Core meaning 1: (To put) out of order</i></p> <p><i>Core meaning 2: To give a contrary order</i></p>	<p>Hereditary (Frequency: 1,550)</p> <p><i>Core meaning: Characteristics or properties passed from a generation to the successive one</i></p>
<p>Artery (Frequency: 11,666)</p> <p><i>Core meaning: A channel</i></p>	<p>Displace (Frequency: 1,530)</p> <p><i>Core meaning: To shift from its original place</i></p>
<p>Surgical (Frequency: 10,491)</p> <p><i>Core meaning: Relating to surgery</i></p>	<p>Reflux (Frequency: 1,529)</p> <p><i>Core meaning: Relating to flowing back</i></p>
<p>Gene (Frequency: 10,425)</p> <p><i>Core meaning: Relating to heredity</i></p>	<p>Basal (Frequency: 1,526)</p> <p><i>Core meaning: Pertaining to the base</i></p>
<p>Hypertension (Frequency: 8,483)</p> <p><i>Core meaning: Extreme tension or pressure</i></p>	<p>Shunt (Frequency: 1,518)</p> <p><i>Core meaning: Relating to a diversion/ To divert</i></p>
<p>Anterior (Frequency: 8,415)</p> <p><i>Core meaning: Prior or in front of physically</i></p>	<p>Facilitate (Frequency: 1,499)</p> <p><i>Core meaning: To make easier</i></p>

<p>Posterior (Frequency: 8,080)</p> <p><i>Core meaning: Following or behind physically</i></p>	<p>Squamous (Frequency: 1,499)</p> <p><i>Core meaning: Scaly</i></p>
<p>Lateral (Frequency: 8,051)</p> <p><i>Core meaning: Relating to the side</i></p>	<p>Polyp (Frequency: 1,494)</p> <p><i>Core meaning 1: A mass arising from a surface</i></p> <p><i>Core meaning 2: An aquatic invertebrate</i></p>
<p>Recur (Frequency: 7,607)</p> <p><i>Core meaning: To occur again</i></p>	<p>Sarcoma (Frequency: 1,493)</p> <p><i>Core meaning 1: A malignant tumour</i></p> <p><i>Core meaning 2: Fleshy mass</i></p>
<p>Inflame (Frequency: 7,554)</p> <p><i>Core meaning: To make very hot</i></p>	<p>Anxiety (Frequency: 1,490)</p> <p><i>Core meaning: Worry</i></p>
<p>Biopsy (Frequency: 7,487)</p> <p><i>Core meaning: To remove/ Removal of a sample from a living creature</i></p>	<p>Swell (Frequency: 1,489)</p> <p><i>Core meaning: Rising/ To increase</i></p>
<p>Pulmonary (Frequency: 7,301)</p> <p><i>Core meaning: Relating to the lungs</i></p>	<p>Neural (Frequency: 1,483)</p> <p><i>Core meaning: Relating to nerves</i></p>
<p>Pathology (Frequency: 6,949)</p> <p><i>Core meaning: Relating to disease</i></p>	<p>Contraceptive (Frequency: 1,482)</p> <p><i>Core meaning: Prevent pregnancy</i></p>
<p>Mutate (Frequency: 6,780)</p> <p><i>Core meaning: To change</i></p>	<p>Protocol (Frequency: 1,482)</p> <p><i>Core meaning: A record</i></p>
<p>Hepatic (Frequency: 6,660)</p> <p><i>Core meaning: Relating to the liver</i></p>	<p>Incontinent (Frequency: 1,474)</p> <p><i>Core meaning: Unable to hold back</i></p>
<p>Malign (Frequency: 6,556)</p> <p><i>Core meaning: Harmful</i></p>	<p>Invasion (Frequency: 1,469)</p> <p><i>Core meaning: Hostile entrance</i></p>

<p>Abdomen (Frequency: 6,507)</p> <p><i>Core meaning: Belly</i></p>	<p>Compartment (Frequency: 1,458)</p> <p><i>Core meaning: A section</i></p>
<p>Defect (Frequency: 6,496)</p> <p><i>Core meaning 1: Relating to a deficiency</i></p> <p><i>Core meaning 2: To leave</i></p>	<p>Latter (Frequency: 1,457)</p> <p><i>Core meaning: Later</i></p>
<p>Prostate (Frequency: 6,304)</p> <p><i>Core meaning: Gland relating to seminal fluid</i></p>	<p>Biochemical (Frequency: 1,456)</p> <p><i>Core meaning: Relating to biochemical processes</i></p>
<p>Graft (Noun) (Frequency: 6,298)</p> <p><i>Core meaning 1: Transplant</i></p> <p><i>Core meaning 2: (Relating to) the spade</i></p> <p>Graft (Verb)</p> <p><i>Core meaning: To transplant</i></p>	<p>Dislocate (Frequency: 1,448)</p> <p><i>Core meaning: To put out of position</i></p>
<p>Bladder (Frequency: 6,283)</p> <p><i>Core meaning: A bag</i></p>	<p>Capillary (Frequency: 1,446)</p> <p><i>Core meaning: Hair-like</i></p>
<p>Node (Frequency: 5,893)</p> <p><i>Core meaning: A protuberance</i></p>	<p>Atrophy (Frequency: 1,446)</p> <p><i>Core meaning: Wasting</i></p>
<p>Cardiac (Frequency: 5,883)</p> <p><i>Core meaning 1: Pertaining to the heart</i></p> <p><i>Core meaning 2: Part of the stomach</i></p>	<p>Biology (Frequency: 1,443)</p> <p><i>Core meaning: Connected to living organisms</i></p>
<p>Vascular (Frequency: 5,811)</p> <p><i>Core meaning: Tubelike</i></p>	<p>Contraction (Frequency: 1,443)</p> <p><i>Core meaning 1: Act of acquiring</i></p> <p><i>Core meaning 2: Shrinking</i></p>
<p>Receptor (Frequency: 5,642)</p>	<p>Vomit (Frequency: 1,439)</p>

<i>Core meaning: A constituent of a cell that responds to a stimulus</i>	<i>Core meaning: The act, cause or product of ejecting contents/ To spout up</i>
Mechanism (Frequency: 5,553) <i>Core meaning: A set of processes or parts working together</i>	Pigment (Frequency: 1,432) <i>Core meaning: Colour</i>
Radiate (Frequency: 5,470) <i>Core meaning: To diverge from a central point</i>	Unilateral (Frequency: 1,428) <i>Core meaning: Relating to one side</i>
Fracture (Frequency: 5,399) <i>Core meaning: Broken or the act of breaking</i>	Axis (Frequency: 1,419) <i>Core meaning: Central line</i>
Vein (Frequency: 5,113) <i>Core meaning: A channel</i>	Resonance (Frequency: 1,413) <i>Core meaning: A sympathetic response</i>
Review (Noun) (Frequency: 5,111) <i>Core meaning: Looking over</i> Review (Verb) <i>Core meaning: To appraise</i>	Cortical (Frequency: 1,408) <i>Core meaning: Belonging to the external part</i>
Genetic (Frequency: 5,031) <i>Core meaning: Relating to origins</i>	Relapse (Frequency: 1,407) <i>Core meaning: (To) return to an undesirable state</i>
Plasma (Frequency: 5,002) <i>Core meaning 1: Liquid in which blood cells are suspended</i> <i>Core meaning 2: Ionised gas</i>	Cortex (Frequency: 1,405) <i>Core meaning: Outer part</i>
Tract (Frequency: 4,998) <i>Core meaning 1: A pamphlet</i> <i>Core meaning 2: An expanse of land</i>	Fistula (Frequency: 1,394) <i>Core meaning: A pipe</i>

<p>Epithelium (Frequency: 4,976)</p> <p><i>Core meaning: Outer layer of tissue</i></p>	<p>Transverse (Frequency: 1,387)</p> <p><i>Core meaning: (Something) lying across</i></p>
<p>Fetal (Frequency: 4,961)</p> <p><i>Core meaning: Relating to an unborn creature</i></p>	<p>Predispose (Frequency: 1,383)</p> <p><i>Core meaning 1: To make susceptible</i></p> <p><i>Core meaning 2: To give in advance</i></p>
<p>Spine (Frequency: 4,743)</p> <p><i>Core meaning 1: Sharp-pointed projection</i></p> <p><i>Core meaning 2: The backbone</i></p>	<p>Subcutaneous (Frequency: 1,382)</p> <p><i>Core meaning: Under the skin</i></p>
<p>Vessel (Frequency: 4,692)</p> <p><i>Core meaning: A container</i></p>	<p>Penetrate (Frequency: 1,371)</p> <p><i>Core meaning: To get through to</i></p>
<p>Secrete (Frequency: 4,562)</p> <p><i>Core meaning 1: To release</i></p> <p><i>Core meaning 2: To do something out of sight</i></p>	<p>Fragment (Frequency: 1,368)</p> <p><i>Core meaning: Broken part</i></p>
<p>Membrane (Frequency: 4,546)</p> <p><i>Core meaning: Covering layer</i></p>	<p>Modality (Frequency: 1,367)</p> <p><i>Core meaning: A way, method or manner</i></p>
<p>Medication (Frequency: 4,493)</p> <p><i>Core meaning: Medical treatment</i></p>	<p>Deposit (Frequency: 1,359)</p> <p><i>Core meaning: Placed somewhere safe</i></p>
<p>Undergo (Frequency: 4,439)</p> <p><i>Core meaning: To go through an experience</i></p>	<p>Probe (Frequency: 1,355)</p> <p><i>Core meaning: Relating to examining/ To examine</i></p>
<p>Portal (Frequency: 4,421)</p> <p><i>Core meaning: (Relating to) an entrance</i></p>	<p>Retard (Frequency: 1,347)</p> <p><i>Core meaning: Hold back progress</i></p>
<p>Systemic (Frequency: 4,299)</p> <p><i>Core meaning: Relating to a system</i></p>	<p>Delete (Frequency: 1,339)</p> <p><i>Core meaning: To remove</i></p>

<p>Medial (Frequency: 4,273)</p> <p><i>Core meaning: Middle</i></p>	<p>Frontal (Frequency: 1,333)</p> <p><i>Core meaning: Relating to the forepart</i></p>
<p>Ovary (Frequency: 4,092)</p> <p><i>Core meaning: Female reproductive organ</i></p>	<p>Profile (Frequency: 1,313)</p> <p><i>Core meaning: An outline/ To outline</i></p>
<p>Proximal (Frequency: 4,077)</p> <p><i>Core meaning: Close to</i></p>	<p>Transient (Frequency: 1,277)</p> <p><i>Core meaning: Temporary</i></p>
<p>Anatomy (Frequency: 4,037)</p> <p><i>Core meaning: Relating to body dissection</i></p>	<p>Spectrum (Frequency: 1,258)</p> <p><i>Core meaning 1: Insubstantial body</i></p> <p><i>Core meaning 2: Arrangement</i></p>
<p>Gland (Frequency: 3,859)</p> <p><i>Core meaning: An organ or group of cells that secrete or filter</i></p>	<p>Decline (Noun) (Frequency: 1,252)</p> <p><i>Core meaning: Weakening</i></p> <p>Decline (Verb)</p> <p><i>Core meaning 1: To go downhill</i></p> <p><i>Core meaning 2: To turn aside</i></p>
<p>Cyst (Frequency: 3,858)</p> <p><i>Core meaning: Cavity containing liquid</i></p>	<p>Perfusion (Frequency: 1,251)</p> <p><i>Core meaning: Flowing through</i></p>
<p>Activate (Frequency: 3,753)</p> <p><i>Core meaning: To put into motion</i></p>	<p>Vertical (Frequency: 1,246)</p> <p><i>Core meaning: Upright</i></p>
<p>Manifest (Noun) (Frequency: 3,718)</p> <p><i>Core meaning: A declaration</i></p> <p>Manifest (Adjective)</p> <p><i>Core meaning: Obvious</i></p>	<p>Differ (Frequency: 1,233)</p> <p><i>Core meaning: To be dissimilar</i></p>
<p>Fibre (Frequency: 3,714)</p> <p><i>Core meaning: Thread-like body</i></p>	<p>Recessive (Frequency: 1,216)</p> <p><i>Core meaning: Regressing</i></p>
<p>Administration (Frequency: 3,680)</p>	<p>Sedate (Frequency: 1,202)</p>

<i>Core meaning: An action or person or group of people carrying out or executing</i>	<i>Core meaning: Relating to being quiet</i>
Sinus (Frequency: 3,670) <i>Core meaning: A cavity</i>	Longitudinal (Frequency: 1,202) <i>Core meaning: Relating to length</i>
Suture (Frequency: 3,640) <i>Core meaning: Line of closure</i>	Skeletal (Frequency: 1,197) <i>Core meaning: Consisting of a framework</i>
Induce (Frequency: 3,633) <i>Core meaning 1: To bring about</i> <i>Core meaning 2: To infer</i>	Cosmetic (Frequency: 1,194) <i>Core meaning: Relating to beautifying</i>
Duct (Frequency: 3,595) <i>Core meaning: A channel</i>	Intern (Noun) (Frequency: 1,191) <i>Core meaning: A trainee</i> Intern (Verb) <i>Core meaning 1: To confine</i> <i>Core meaning 2: To word as a trainee</i>
Circulate (Frequency: 3,568) <i>Core meaning: To move around</i>	Phenotype (Frequency: 1,190) <i>Core meaning: Observable characteristic</i>
Trauma (Frequency: 5,552) <i>Core meaning: Injury</i>	Emerge (Frequency: 1,187) <i>Core meaning: To come into view</i>
Insert (Frequency: 3,508) <i>Core meaning: To put in/ Relating to putting in</i>	Alveolar (Frequency: 1,187) <i>Core meaning: Relating to a cell-like space</i>
Segment (Frequency: 3,504) <i>Core meaning: A division/ To divide</i>	Germ (Frequency: 1,183) <i>Core meaning: A source</i>
Venous (Frequency: 3,446) <i>Core meaning: Relating to veins</i>	Diuretic (Frequency: 1,182) <i>Core meaning: Promoting urination</i>

<p>Sequence (Frequency: 3,389)</p> <p><i>Core meaning: Successive order</i></p>	<p>Tolerate (Frequency: 1,181)</p> <p><i>Core meaning: To put up with</i></p>
<p>Toxic (Frequency: 3,383)</p> <p><i>Core meaning: (Relating to) a poison</i></p>	<p>Complement (Frequency: 1,179)</p> <p><i>Core meaning: Relating to completion</i></p>
<p>Superior (Frequency: 3,361)</p> <p><i>Core meaning: Higher</i></p>	<p>Prolapse (Frequency: 1,175)</p> <p><i>Core meaning: (Relating to) slip(ping) out of place</i></p>
<p>Necrosis (Frequency: 3,307)</p> <p><i>Core meaning: Death of tissues or cells</i></p>	<p>Precursor (Frequency: 1,174)</p> <p><i>Core meaning: Something that comes before another</i></p>
<p>Enhance (Frequency: 3,288)</p> <p><i>Core meaning: To increase</i></p>	<p>Refract (Frequency: 1,173)</p> <p><i>Core meaning: To deflect the course of light rays</i></p>
<p>Arch (Frequency: 3,253)</p> <p><i>Core meaning 1: To make/ Having a curved structure</i></p> <p><i>Core meaning 2: Cunning</i></p>	<p>Resolution (Frequency: 1,165)</p> <p><i>Core meaning 1: Conversion</i></p> <p><i>Core meaning 2: Coming to a solution</i></p>
<p>Isolate (Frequency: 3,241)</p> <p><i>Core meaning: To separate</i></p>	<p>Acuity (Frequency: 1,165)</p> <p><i>Core meaning: Sharpness</i></p>
<p>Stimulate (Frequency: 3,225)</p> <p><i>Core meaning: To stir to action</i></p>	<p>Anal (Frequency: 1,163)</p> <p><i>Core meaning: Relating to the anus</i></p>
<p>Intravenous (Frequency: 3,198)</p> <p><i>Core meaning: Within a vein</i></p>	<p>Cuff (Frequency: 1,162)</p> <p><i>Core meaning 1: Something round the wrist</i></p> <p><i>Core meaning 2: Relating to a blow with the fist</i></p>

<p>Specimen (Frequency: 3,188)</p> <p><i>Core meaning: An example</i></p>	<p>Regress (Frequency: 1,161)</p> <p><i>Core meaning: To revert</i></p>
<p>Prior (Frequency: 3,167)</p> <p><i>Core meaning: Before</i></p>	<p>Papillary (Frequency: 1,161)</p> <p><i>Core meaning: Relating to a small fleshy projection</i></p>
<p>Respirator (Frequency: 3,107)</p> <p><i>Core meaning: Something that helps with breathing</i></p>	<p>Yield (Noun) (Frequency: 1,145)</p> <p><i>Core meaning: Amount produced</i></p> <p>Yield (Verb)</p> <p><i>Core meaning: To give</i></p>
<p>Haemorrhage (Frequency: 3,092)</p> <p><i>Core meaning: Relating to draining away</i></p>	<p>Formula (Frequency: 1,141)</p> <p><i>Core meaning: A set form of something</i></p>
<p>Intervene (Frequency: 3,071)</p> <p><i>Core meaning: To come between</i></p>	<p>Fungus (Frequency: 1,140)</p> <p><i>Core meaning: Mushroom</i></p>
<p>Benign (Frequency: 3,017)</p> <p><i>Core meaning: Mild</i></p>	<p>Objective (Adjective) (Frequency: 1,16)</p> <p><i>Core meaning: Detached</i></p> <p>Objective (Noun)</p> <p><i>Core meaning 1: Something independent of the mind</i></p> <p><i>Core meaning 2: Target</i></p>
<p>Component (Frequency: 3,016)</p> <p><i>Core meaning: (Relating to) constituent parts</i></p>	<p>Migrate (Frequency: 1,131)</p> <p><i>Core meaning: To move to a new location</i></p>
<p>Underlie (Frequency: 2,958)</p> <p><i>Core meaning: To be underneath</i></p>	<p>Reflex (Frequency: 1,129)</p> <p><i>Core meaning 1: Reproduction of an original</i></p> <p><i>Core meaning 2: An automatic response</i></p>

Scar (Frequency: 2,933) <i>Core meaning: (Relating to) a wound</i>	Occlusion (Frequency: 1,124) <i>Core meaning: Closing</i>
Proliferate (Frequency: 2,902) <i>Core meaning: To generate in large quantities</i>	Stricture (Frequency: 1,121) <i>Core meaning 1: Narrowing</i> <i>Core meaning 2: Negative criticism</i>
Invasive (Frequency: 2,780) <i>Core meaning: Relating to attacking</i>	Evident (Frequency: 1,117) <i>Core meaning: Obvious</i>
Donor (Frequency: 2,766) <i>Core meaning: A giver</i>	Electrolyte (Frequency: 1,107) <i>Core meaning: Relating to ions</i>
Modify (Frequency: 2,761) <i>Core meaning: To make minor changes</i>	Prescribe (Frequency: 1,103) <i>Core meaning: To direct</i>
Anomaly (Frequency: 2,761) <i>Core meaning: Irregularity</i>	Entity (Frequency: 1,100) <i>Core meaning: Being</i>
Administer (Frequency: 2,749) <i>Core meaning: To execute a task</i>	Morphology (Frequency: 1,098) <i>Core meaning: Relating to form and structure</i>
Inferior (Frequency: 2,744) <i>Core meaning: Lower</i>	Encounter (Frequency: 1,097) <i>Core meaning: To meet</i>
Arise (Frequency: 2,705) <i>Core meaning: To move up</i>	Accomplish (Frequency: 1,096) <i>Core meaning: To complete</i>
Lens (Frequency: 2,702) <i>Core meaning: A curved surface that bends light rays</i>	Attribute (Frequency: 1,062) <i>Core meaning: (Relating to) ascribing</i>
Diffuse (Frequency: 2,694) <i>Core meaning: (Relating to) disperse</i>	Peel (Frequency: 1,056) <i>Core meaning 1: A baker's shovel</i>

	<i>Core meaning 2: (Relating to) the outer covering</i>
Optic (Frequency: 2,669) <i>Core meaning: Connected to vision</i>	Amplify (Frequency: 1,055) <i>Core meaning: To enlarge</i>
Adrenal (Frequency: 2,657) <i>Core meaning: Relating to an endocrine gland near the kidney</i>	Prenatal (Frequency: 1,054) <i>Core meaning: Before birth</i>
Superficial (Frequency: 2,656) <i>Core meaning: Lacking depth</i>	Sustain (Frequency: 1,042) <i>Core meaning: To maintain</i>
Acquire (Frequency: 2,648) <i>Core meaning: To obtain</i>	Compound (Frequency: 1,026) <i>Core meaning: Relating to bringing together</i>
Bilateral (Frequency: 2,630) <i>Core meaning: Involving both sides</i>	Lamina (Frequency: 1,025) <i>Core meaning: A thin layer</i>
Spontaneous (Frequency: 2,613) <i>Core meaning: Without stimulus</i>	Retract (Frequency: 1,024) <i>Core meaning: To draw back</i>
Fever (Frequency: 2,585) <i>Core meaning: Relating to burning</i>	Parallel (Frequency: 1,006) <i>Core meaning: Relating to side by side</i>
Prognosis (Frequency: 2,560) <i>Core meaning: Prediction</i>	Newborn (Frequency: 1,004) <i>Core meaning: (Relating to) being born recently</i>
Coronary (Frequency: 2,480) <i>Core meaning: Crown-like</i>	Oblique (Frequency: 1,002) <i>Core meaning: At an angle</i>
Implant (Frequency: 2,449) <i>Core meaning: Relating to inserting</i>	Smear (Frequency: 1,002) <i>Core meaning: To spread a thick substance</i>
Differential (Frequency: 2,384)	Translocate (Frequency: 1,002)

<i>Core meaning: Relating to distinguishing</i>	<i>Core meaning: To move from one place to another</i>
Rupture (Frequency: 2,341) <i>Core meaning: A break/ To break</i>	Matrix (Frequency: 1,001) <i>Core meaning: A support structure</i>
Nucleus (Frequency: 2,337) <i>Core meaning: Central</i>	Degenerate (Frequency: 988) <i>Core meaning: (To become) deficient in normal qualities</i>
Compress (Frequency: 2,321) <i>Core meaning: (Relating to) pressing together</i>	Consent (Frequency: 985) <i>Core meaning: (Relating to) agreeing</i>
Moderate (Frequency: 2,305) <i>Core meaning 1: Medium</i> <i>Core meaning 2: Relating to managing a discussion</i>	Retrospect (Frequency: 985) <i>Core meaning: Looking back</i>
Radical (Frequency: 2,302) <i>Core meaning 1: Relating to a root</i> <i>Core meaning 2: Progressive</i>	Stool (Frequency: 984) <i>Core meaning 1: A wooden seat</i> <i>Core meaning 2: Discharged faecal matter</i>
Adverse (Frequency: 2,283) <i>Core meaning: Unfavourable</i>	Polar (Frequency: 975) <i>Core meaning: Relating to pole(s)</i>
Organism (Frequency: 2,277) <i>Core meaning: A living structure</i>	Aspirate (Frequency: 972) <i>Core meaning: Marked with a breath</i>
Deform (Frequency: 2,245) <i>Core meaning: To mar</i>	Palsy (Frequency: 970) <i>Core meaning: (Relating to) paralysis</i>
Pituitary (Frequency: 2,241) <i>Core meaning: Relating to the pituitary gland</i>	Constitute (Frequency: 964) <i>Core meaning: To establish</i>

<p>Classification (Frequency: 2,238)</p> <p><i>Core meaning: Arrangement</i></p>	<p>Uptake (Frequency: 962)</p> <p><i>Core meaning: Absorption</i></p>
<p>Tubular (Frequency: 2,222)</p> <p><i>Core meaning: Tube-related</i></p>	<p>Blunt (Frequency: 952)</p> <p><i>Core meaning: Dull</i></p>
<p>Colon (Frequency: 2,216)</p> <p><i>Core meaning 1: Part of the large intestine</i></p> <p><i>Core meaning 2: A punctuation mark</i></p>	<p>Disseminate (Frequency: 952)</p> <p><i>Core meaning: To spread</i></p>
<p>Suppress (Frequency: 2,190)</p> <p><i>Core meaning: To keep down</i></p>	<p>Effusion (Frequency: 947)</p> <p><i>Core meaning: Spill</i></p>
<p>Absorb (Frequency: 2,165)</p> <p><i>Core meaning: To take in as part of something larger</i></p>	<p>Traction (Frequency: 942)</p> <p><i>Core meaning: Pulling</i></p>
<p>Cellular (Frequency: 2,136)</p> <p><i>Core meaning: Relating to cells</i></p>	<p>Hybrid (Frequency: 937)</p> <p><i>Core meaning: Cross-breeding</i></p>
<p>Synthesis (Frequency: 2,128)</p> <p><i>Core meaning: Putting together</i></p>	<p>Digital (Frequency: 933)</p> <p><i>Core meaning 1: Relating to numbers</i></p> <p><i>Core meaning 2: Relating to fingers</i></p>
<p>Orbit (Frequency: 2,126)</p> <p><i>Core meaning 1: The eye socket</i></p> <p><i>Core meaning 2: An elliptical course</i></p>	<p>Explore (Frequency: 930)</p> <p><i>Core meaning: To discover</i></p>
<p>Fixate (Frequency: 2,121)</p> <p><i>Core meaning: To stabilize</i></p>	<p>Mimic (Frequency: 930)</p> <p><i>Core meaning: Relating to imitating</i></p>
<p>Excise (Frequency: 2,106)</p> <p><i>Core meaning: Relating to removing</i></p>	<p>Saline (Frequency: 928)</p> <p><i>Core meaning: Relating to salt</i></p>
<p>Nutrition (Frequency: 2,058)</p>	<p>Antagonist (Frequency: 927)</p>

<i>Core meaning: Nourishment</i>	<i>Core meaning: Relating to opposing</i>
Ventilate (Frequency: 2,053) <i>Core meaning 1: To supply air</i> <i>Core meaning 2: To express a view</i>	Muscular (Frequency: 927) <i>Core meaning: Relating to muscles</i>
Curve (Frequency: 2,030) <i>Core meaning: Relating to bending</i>	Sheath (Frequency: 924) <i>Core meaning: A covering</i>
Adjacent (Frequency: 2,029) <i>Core meaning: Close to</i>	Crypt (Frequency: 915) <i>Core meaning: Recess</i>
Residue (Frequency: 2,008) <i>Core meaning 1: Remainder</i> <i>Core meaning 2: Small molecule in a polymer</i>	Traumatic (Frequency: 914) <i>Core meaning: Relating to an injury</i>
Valve (Frequency: 2,000) <i>Core meaning: Relating to the control of flow</i>	Diaphragm (Frequency: 912) <i>Core meaning: Partition</i>
Allograft (Frequency: 1,905) <i>Core meaning: Transplant</i>	Forceps (Frequency: 906) <i>Core meaning: Pincers</i>
Compose (Frequency: 1,897) <i>Core meaning: To put together</i>	Locus (Frequency: 903) <i>Core meaning: A place</i>
Limb (Frequency: 1,888) <i>Core meaning: Relating a body's appendages</i>	Plexus (Frequency: 902) <i>Core meaning: A network</i>
Preserve (Frequency: 1,858) <i>Core meaning: Keeping from harm/ damage</i>	Outline (Frequency: 896) <i>Core meaning: (To draw) the contour of something</i>

<p>Infarct (Frequency: 1,829)</p> <p><i>Core meaning: Relating to obstruction</i></p>	<p>Balloon (Frequency: 887)</p> <p><i>Core meaning: Inflated ball</i></p>
<p>Perforate (Frequency: 1,807)</p> <p><i>Core meaning: To pierce</i></p>	<p>Bundle (Frequency: 887)</p> <p><i>Core meaning 1: Relating to objects tied together</i></p> <p><i>Core meaning 2: To shove away or into</i></p>
<p>Minimise (Frequency: 1,793)</p> <p><i>Core meaning: To reduce</i></p>	<p>Concomitant (Frequency: 884)</p> <p><i>Core meaning: Relating to accompanying</i></p>
<p>Pulse (Frequency: 1,763)</p> <p><i>Core meaning: Relating to short bursts of movement</i></p>	<p>Retain (Frequency: 883)</p> <p><i>Core meaning: To hold back</i></p>
<p>Cataract (Frequency: 1,744)</p> <p><i>Core meaning 1: Waterfall</i></p> <p><i>Core meaning 2: Lens impairment</i></p>	<p>Stem (Frequency: 883)</p> <p><i>Core meaning 1: Relating to a support structure</i></p> <p><i>Core meaning 2: To stop moving</i></p>
<p>Mediate (Frequency: 1,743)</p> <p><i>Core meaning: To make less extreme</i></p>	<p>Regenerate (Frequency: 882)</p> <p><i>Core meaning: To form again</i></p>
<p>Transfuse (Frequency: 1,679)</p> <p><i>Core meaning: To transfer liquid</i></p>	<p>Abort (Frequency: 881)</p> <p><i>Core meaning: To end prematurely</i></p>
<p>Conduct (Frequency: 1,651)</p> <p><i>Core meaning: Relating to directing</i></p>	<p>Resuscitate (Frequency: 880)</p> <p><i>Core meaning: To revive</i></p>
<p>Discharge (Frequency: 1,632)</p> <p><i>Core meaning: (To) release</i></p>	<p>Replicate (Frequency: 878)</p> <p><i>Core meaning: To copy</i></p>
<p>Radial (Frequency: 1,629)</p> <p><i>Core meaning: Diverging from a central point</i></p>	<p>Rib (Frequency: 864)</p> <p><i>Core meaning: Relating to a long curved piece of bone or other substance</i></p>

Fascia (Frequency: 1,622) <i>Core meaning: Band-like object</i>	Vesicle (Frequency: 863) <i>Core meaning: A small sac</i>
---	---

APPENDIX 2. Qualitative and quantitative analysis results of 40 sampled words

Single-meaning words

- 1 ***Acute***
Core meaning 1: (adj) (Of disease) sudden and severe (2)
Core meaning 2: (adj) Relating to an angle of less than 90 degrees (1)
- 2 ***Cardiac***
Core meaning: (adj) Relating to the heart (2)
- 3 ***Cataract***
Core meaning: (n) A medical condition that stops the eye's lens being transparent, making it difficult to see (2)
- 4 ***Chronic***
Core meaning: (adj) (Of disease) long-lasting (2)
- 5 ***Colon***
Core meaning 1: (n) A punctuation mark that introduces something
Core meaning 2: (n) The biggest part of the large intestine (2)
- 6 ***Disorder***
Core meaning: (n) An illness (2)
- 7 ***Induce***
Core meaning: (v) To make something happen (2)
- 8 ***Intern***
Core meaning 1: (v) To keep someone in a place especially for political reasons
Core meaning 2: (v) To work as a supervised trainee (2)
- 9 ***Liver***
Core meaning: (n) An organ of the human (or animal) body (2)
- 10 ***Palsy***
Core meaning: (n) Paralysis that can be accompanied with shaking (2)

11 **Secrete**
Core meaning 1: (v) To make and release a liquid (2)
Core meaning 2: (v) To hide something out of sight

12 **Stool**
Core meaning 1: (n) A wooden seat (2)
Core meaning 2: (n) Solid waste from the body (2)

13 **Tumour**
Core meaning: (n) A lump caused by disease (2)

Multi-meaning words with single core meaning

14 **Absorb**
Core meaning: To take in
(v) To take something in and make it part of something else (2)
(v) To take in something and make it part of what someone knows

15 **Benign**
Core meaning: Mild
(adj) (Of disease) not harmful (2)
(adj) (Of weather) pleasant

16 **Compound**
Core meaning: Relating to bringing together
(v) To add to and make worse (2)
(v) To combine (2)
 (n) A mixture (2)
 (adj) Made up of more than one thing (2)

17 **Conduct**
Core meaning: Relating to doing
(v) To do (2)

(n) Behaviour (2)

18 ***Circulate***

Core meaning: To move around

(v) (Of the blood) to flow round the body (2)

(v) (Of money/papers/information) to move around

(v) (Of viruses/cells/water/air) to move around (2)

19 ***Degenerate***

Core meaning: Deficient in normal qualities

(v) To get worse (2)

(adj) With low moral standards (2)

(n) Someone who behaves badly (2)

20 ***Fascia***

Core meaning: Band-like object

(n) A thin sheet of tissue (2)

21 ***Inferior***

Core meaning: Lower

(adj) Below (2)

(adj) Less important (1)

(n) Someone who is less important

22 ***Lobe***

Core meaning: A rounded division or region, or part of a larger structure

(n) A rounded subdivision (2)

(n) Of the brain, liver, lung, prostate, pituitary (2)

(n) (The soft lower part) of the external ear (2)

(n) A region

(n) A distinctively shaped region surrounding a star in a binary system (The Roche lobe)

(n) The region in a galaxy that emits strong radio energy

(n) Part of a flower

(n) The part of a camwheel that sticks out

(n) The stronger part of a radiation pattern around an aerial

23 ***Migrate***

Core meaning: To move to a new location

(v) (Of cells or organs) to move to another part of the body (2)

(v) (Of animals) to travel to another place to find food or mate

24 ***Parallel***

Core meaning: Relating to side by side

(adj) Occurring at the same time (2)

(adj) Similar (2)

(v) To match (2)

(n) A similarity

(adj) The same distance from other lines all the way along (2)

(v) To be the same distance from something (2)

(adv) In the same direction and at the same distance (2)

(n) Lines that are the same distance from each other all the way along

25 ***Predispose***

Core meaning: To make something likely to happen

(v) To make (a person or animal) likely to have a particular illness (2)

(v) To make someone tend to do something

26 ***Primary***

Core meaning: First

(adj) Culture of cells from the tissue where a disease started (3)

(adj) Found in the tissue or organ where it started (3)

(n) A neoplasm found in the tissue or organ where a disease started (3)

(adj) Important (2)

(n) An American election in which party members vote for who will represent the party in later elections

(adj) (Of education) elementary (1)

(adj) Relating to earliest symptoms of diseases (3)

(adj) Not linked to a previous illness (3)

27 ***Prior***

Core meaning: Before

(adj) Earlier (2)

(adv) Happening before (2)

28 ***Radiate***

Core meaning: Spreading (from a central point)

(v) To go out from a central point (2)

(v) To send out rays (of light) (2)

(v) To display a feeling (2)

29 ***Sedate***

Core meaning: Relating to being calm

(v) To give someone a drug to make them sleepy (2)

(adj) Calm and quiet

30 ***Shunt***

Core meaning: Relating to a diversion

(n) Something that lessens the current in the main circuit

(n) The channel for blood to flow (2)

(v) To make blood flow through a different channel (2)

(v) To move a train onto a different track

(v) To push aside

Multi-meaning words with more than one core meaning

31 ***Arch***

Core meaning 1: (To make) a curved shape

(n) The curved part under the foot (2)

(n) A structure with a curved top (2)

(v) To make into or provide a curve (2)

(v) To go over (2)

Core meaning 2: (adj) Cheeky

32 ***Diffuse***

Core meaning 1: (v) To make something weaker

Core meaning 2: Widespread

(v) To (make something) spread (2)

(adj) Spread out (2)

(adj) (Of disease) in more than one place (2)

33 ***Defect***

Core meaning 1: A lack

(n) An imperfection

(n) Something wrong with part of the body (2)

(n) Something that is not perfect

Core meaning 2: (v) To leave (and join the other side)

34 ***Moderate***

Core meaning 1: Not excessive

(adj) (Of intensity, quality or person) modest (2)

(v) To make or become less intense (2)

(n) Someone who does not express extreme ideas

Core meaning 2: (v) To manage a discussion or group (2)

35 ***Orbit***

Core meaning 1: (n) The eye socket (3)

Core meaning 2: Elliptical course

(n) The path something in space follows round something bigger (2)

(v) To follow a path in space round something bigger (2)

36 ***Peel***

Core meaning 1: (n) A pole with a flat part at one end for removing bread from an oven

Core meaning 2: Outer layer

(n) A face treatment that makes the surface of the skin smoother (2)

(n) The outer layer of a fruit (2)

(v) To remove the outer layer of something (2)

37 ***Radical***

Core meaning 1: Relating to a root

(adj) (Of treatment) working against the root of a disease or tumour, etc. (2)

Core meaning 2: (adj) Non-traditional

38 ***Reflex***

Core meaning 1: (n) A copy of an original

Core meaning 2: A response

(n) An automatic response to a stimulus (2)

(adj) Done as an automatic response (2)

39 ***Resolve***

Core meaning 1: (n) Determination to do something

Core meaning 2: To bring to an end

(v) To solve a problem (2)

(v) To end a disease (2)

40

Stem

Core meaning 1: Relating to a supporting structure

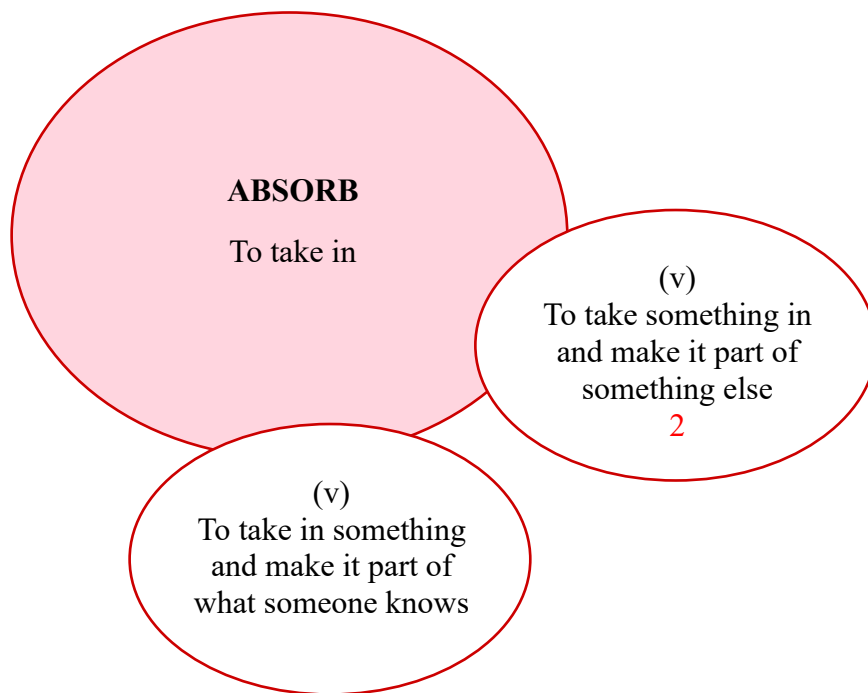
(n) The main part of a supporting structure (2)

(v) To have its origin in (2)

Core meaning 2: (v) To stop something flowing

(ACRONYM) STEM = Science Technology Engineering and Mathematics field/
education

APPENDIX 3. A pilot version of SemiMed (40 Lexical Constellations)



ACUTE (adj)

(Of disease) sudden and
severe

2

ACUTE (adj)

Relating to an angle of less than
90 degrees

1

ARCH (adj)
Cheeky

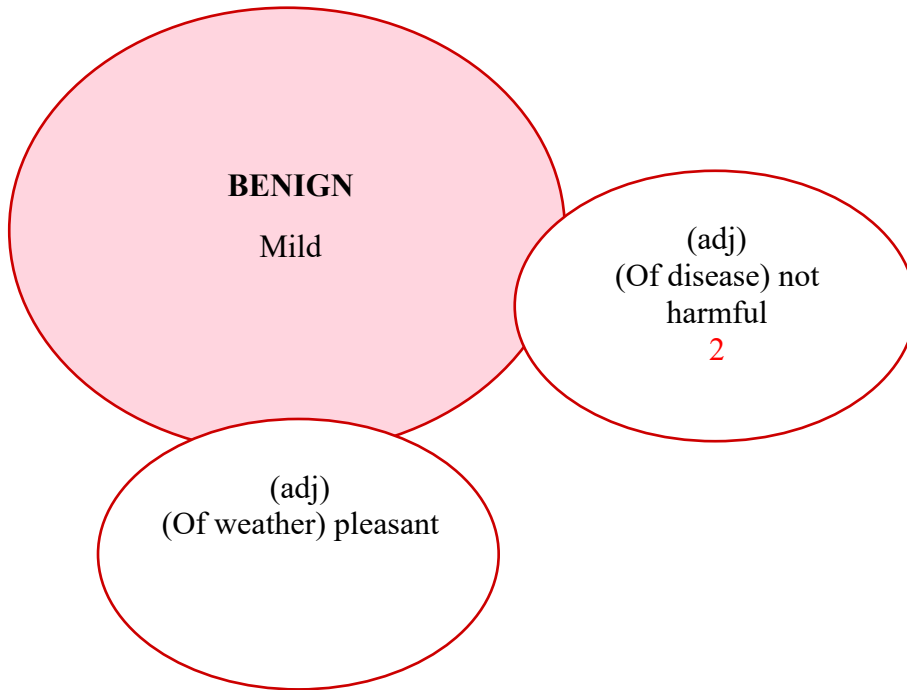
ARCH
(To make) a curved shape

(n)
A structure with a
curved top
2

(n)
The curved part
under the foot
2

(v)
To go over
2

(v)
To make into
or provide a
curve
2



CARDIAC (adj)

Relating to the heart

2

CATARACT (n)

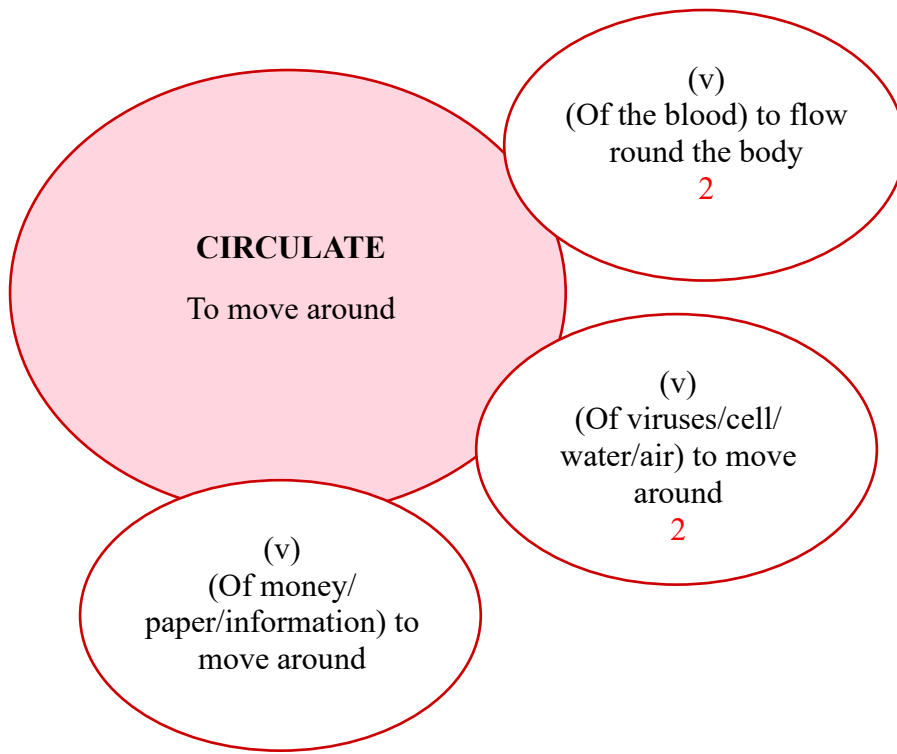
A medical condition that stops the eye's lens being transparent, making it difficult to see

2

CHRONIC (adj)

(Of disease) long-lasting

2



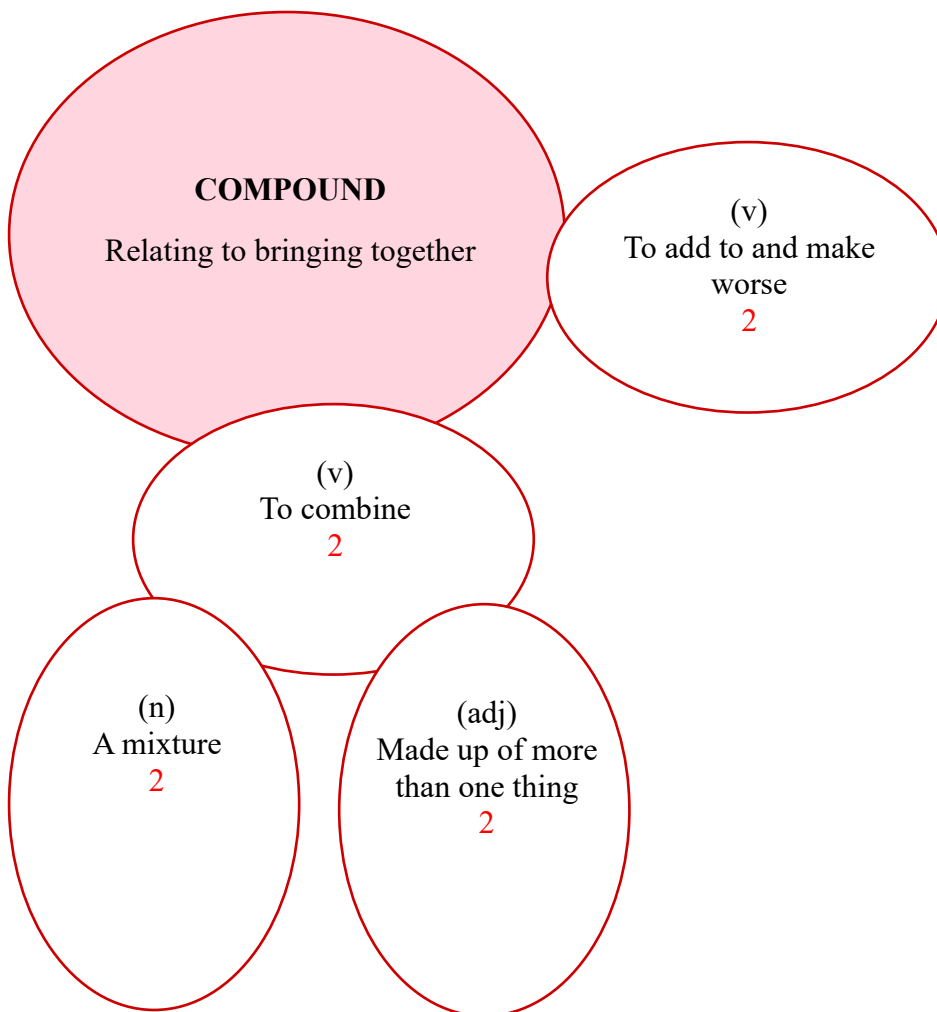
COLON (n)

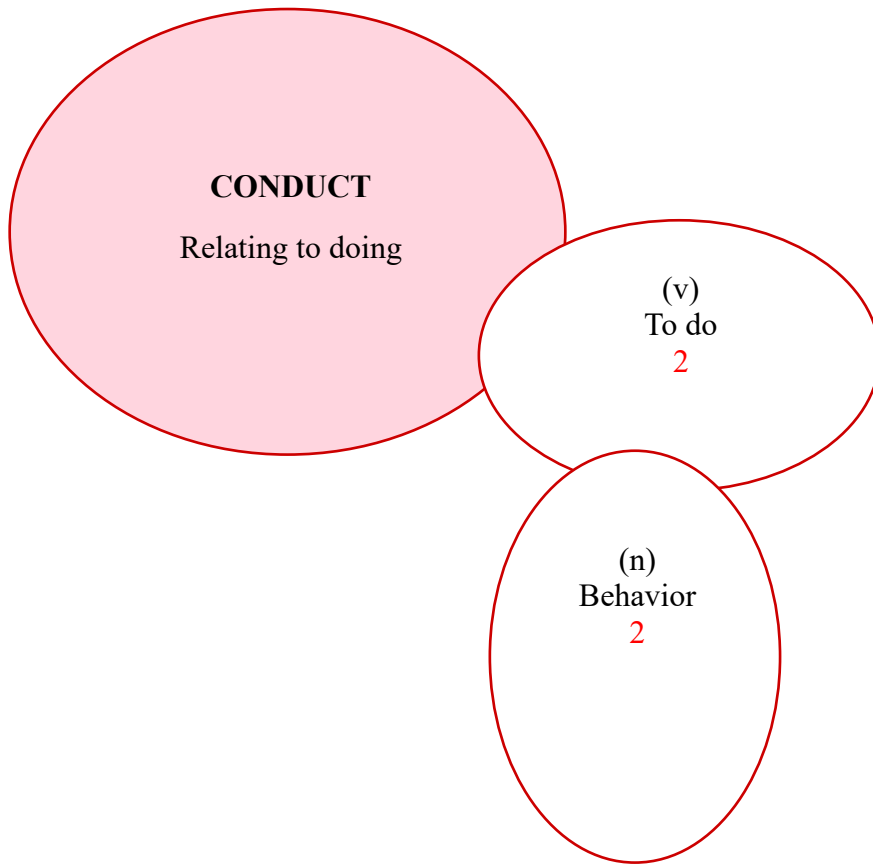
A punctuation mark that
introduces something

COLON (n)

The biggest part of the large
intestine

2





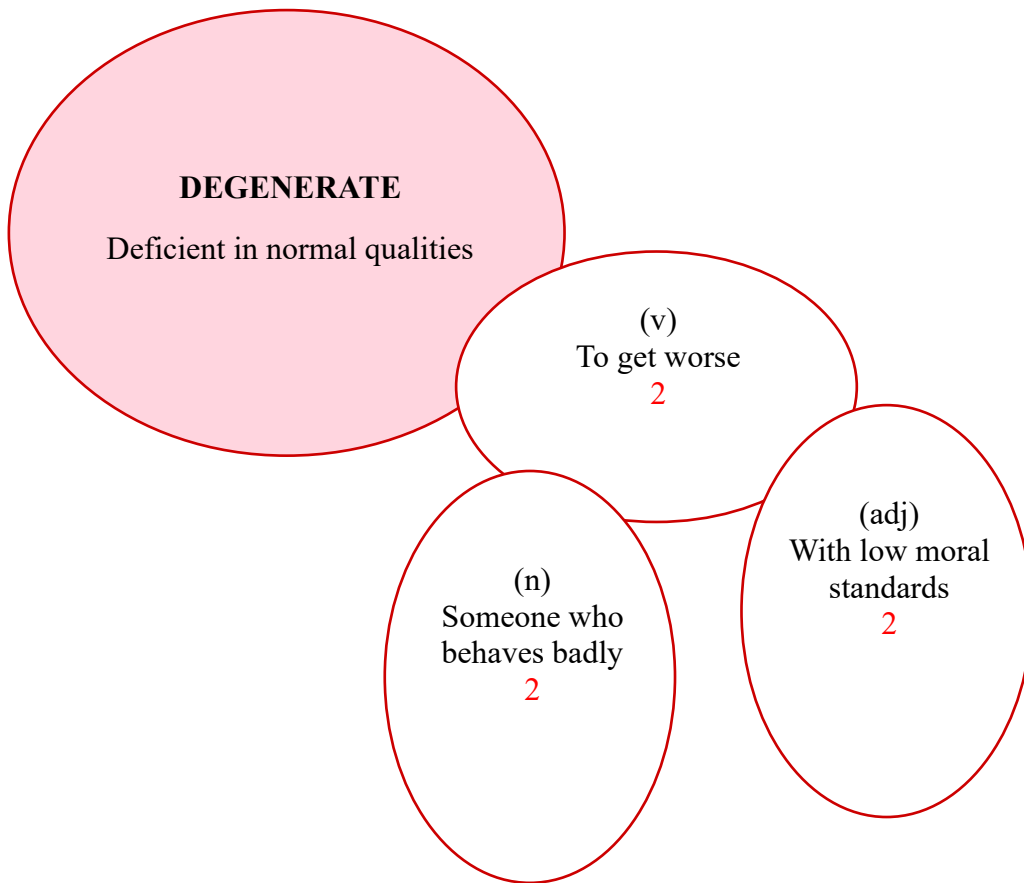
DEFECT (v)
To leave (and join the other side)

DEFECT
A lack

(n)
An imperfection

(n)
Something that
is not perfect

(n)
Something
wrong with part
of the body
2



DIFFUSE (v)
To make something weaker

DIFFUSE
Widespread

(adj)
Spread out
2

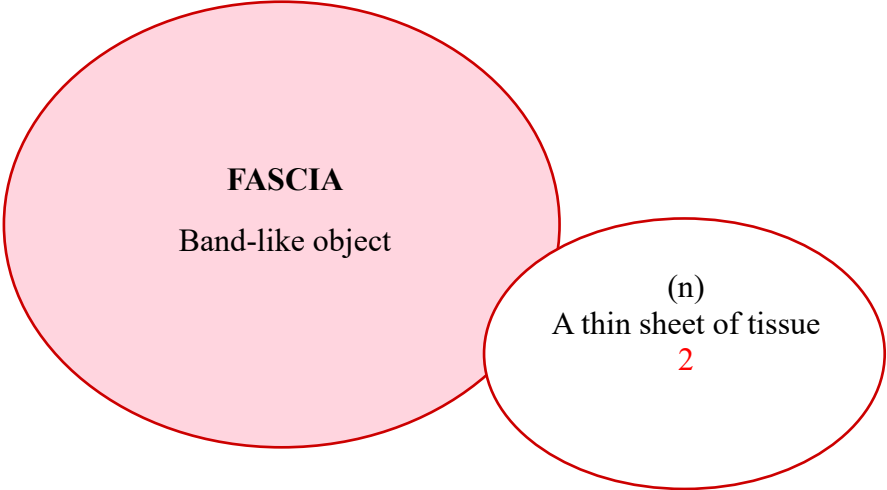
(v)
To (make something)
spread
2

(adj)
(Of disease) in
more than one
place
2

DISORDER (n)

An illness

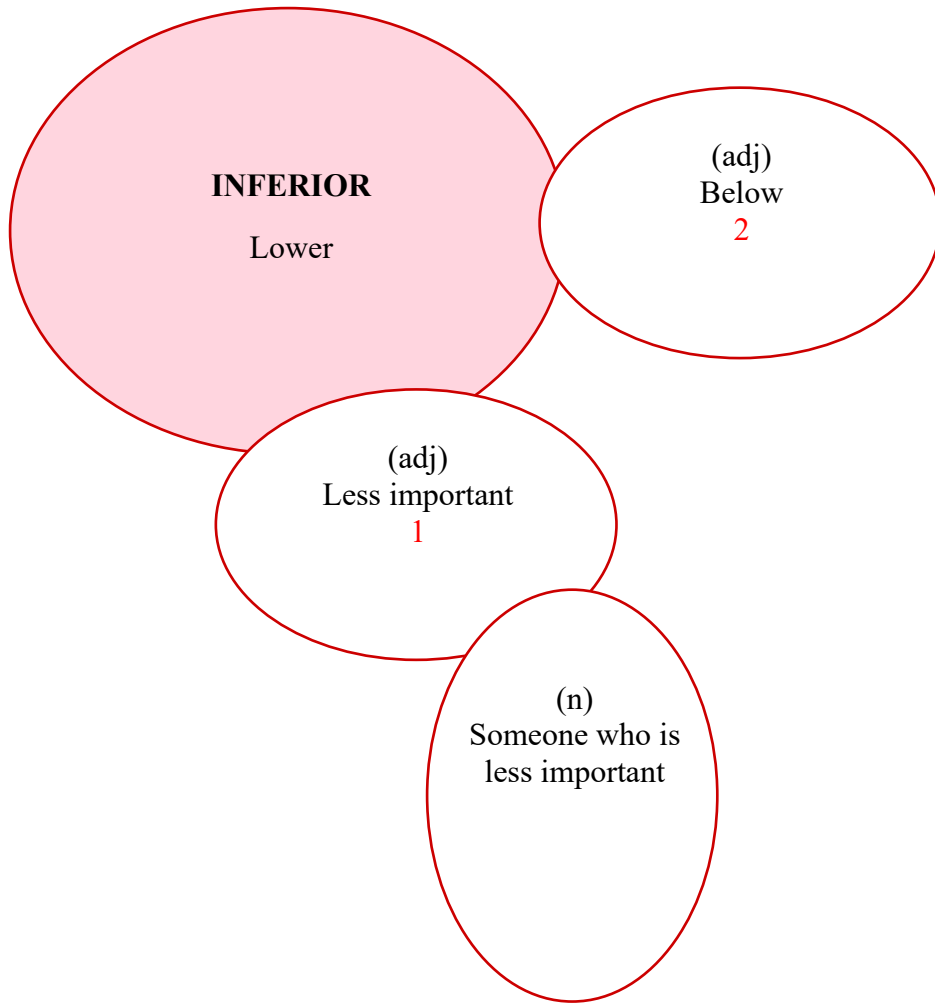
2



INDUCE (v)

To make something happen

2



INTERN (v)

To keep someone in a place, especially for political reasons

INTERN (v)

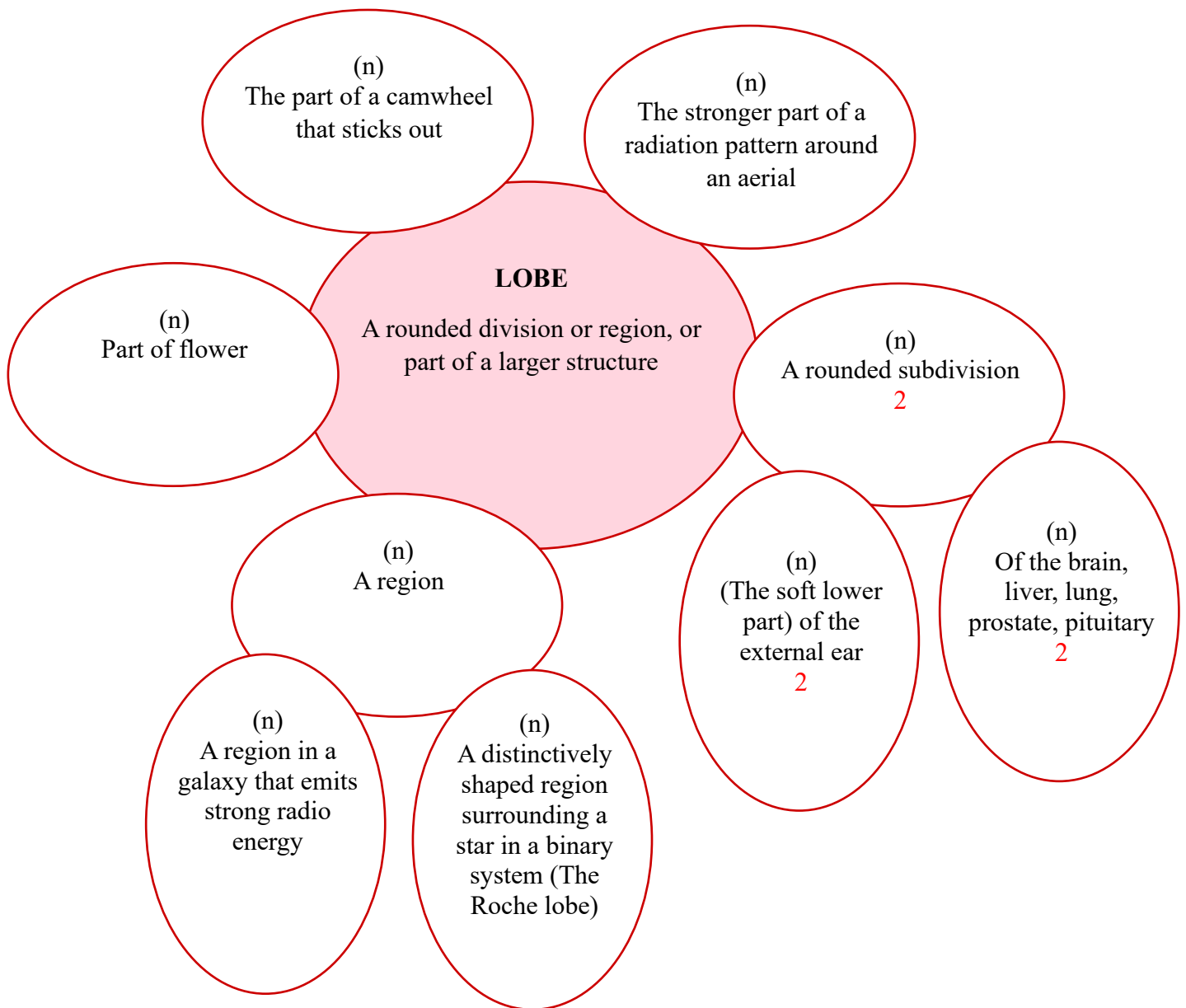
To work as a supervised trainee

2

LIVER (n)

An organ of the human (or
animal) body

2



MIGRATE

To move to a new location

(v)

(Of cells or organs) to
move to another part of
the body

2

(v)

(Of animals) to travel to
another place to find
food or mate

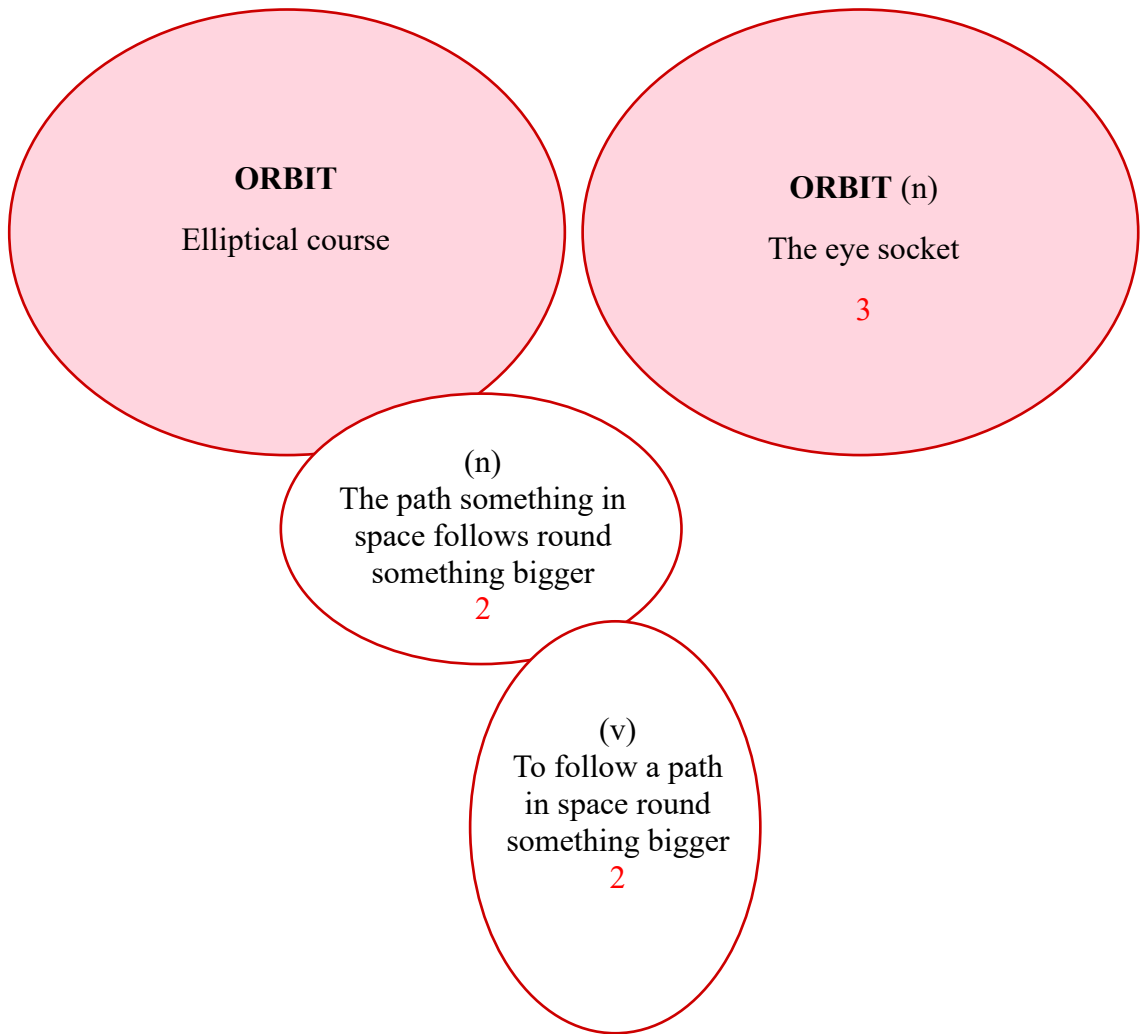
MODERATE (v)
To manage a discussion or group
2

MODERATE
Not excessive

(adj)
(Of intensity, quality, or person) modest
2

(n)
Someone who does not express extreme ideas

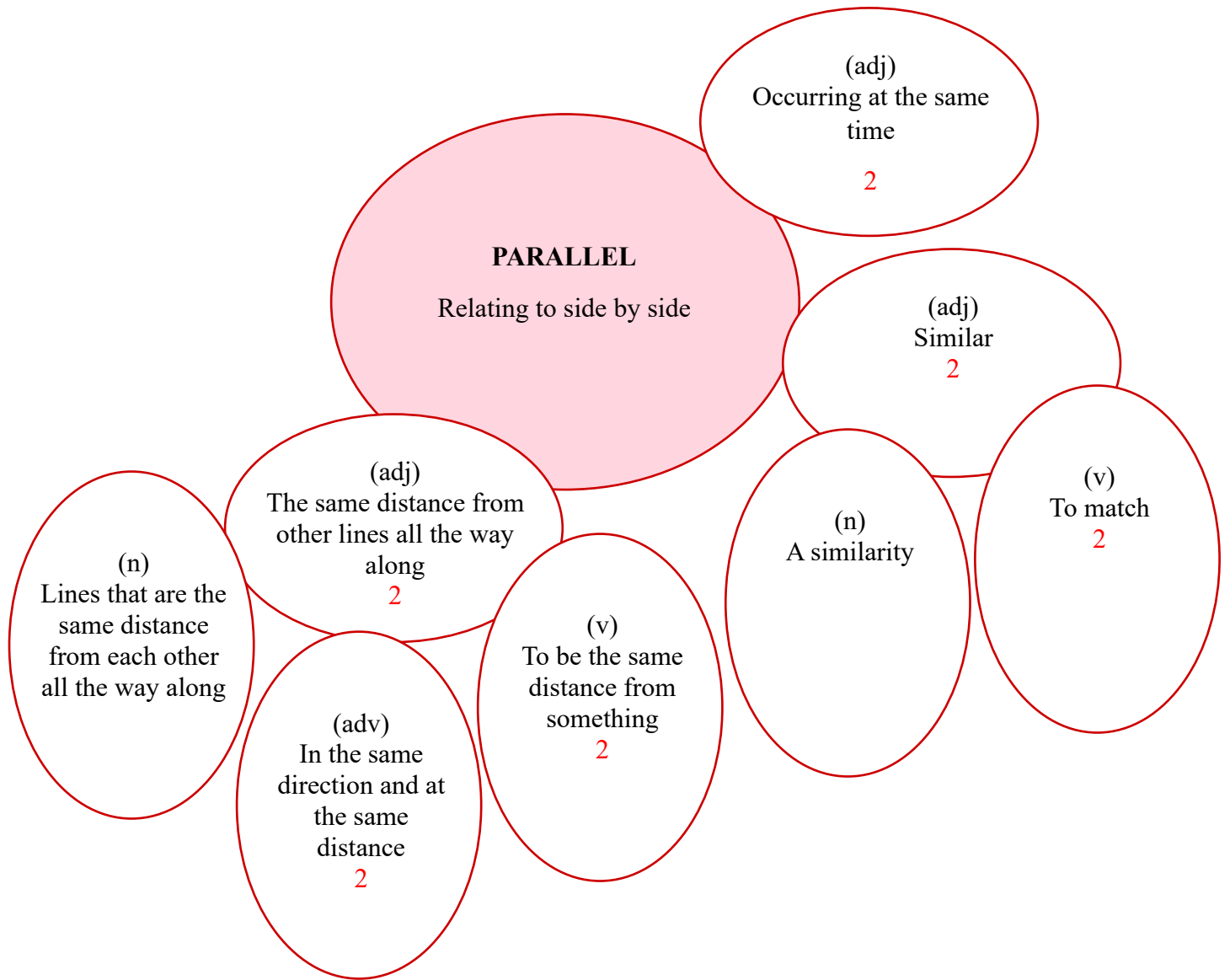
(v)
To make or become less intense
2



PALSY (n)

Paralysis that can be
accompanied with shaking

2



PEEL (n)

A pole with a flat part at one end
for removing bread from an oven

PEEL

Outer layer

(n)

The face treatment that
makes the surface of the
skin smoother

2

(n)

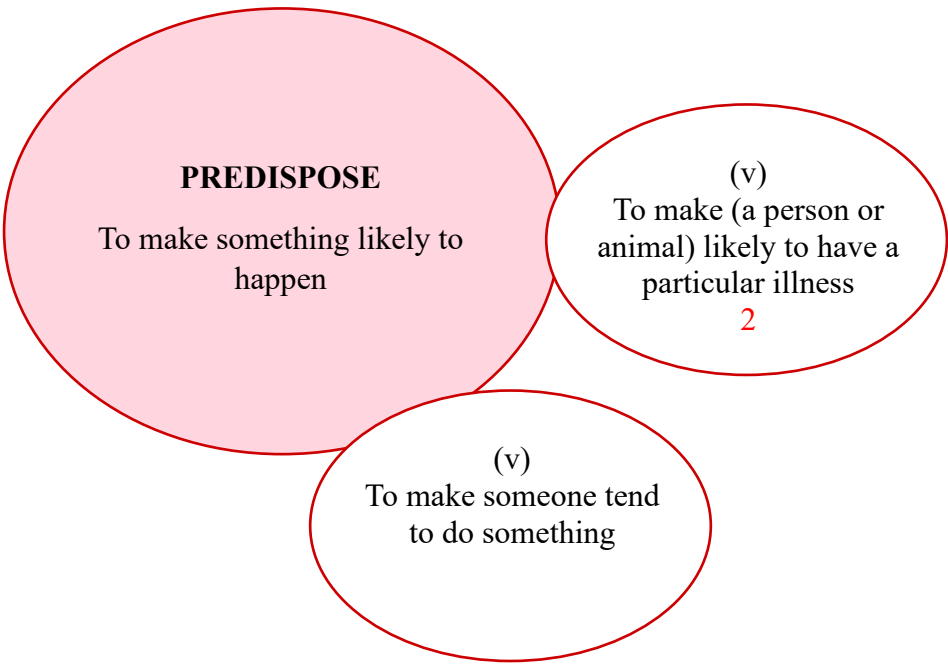
The outer layer of a
fruit

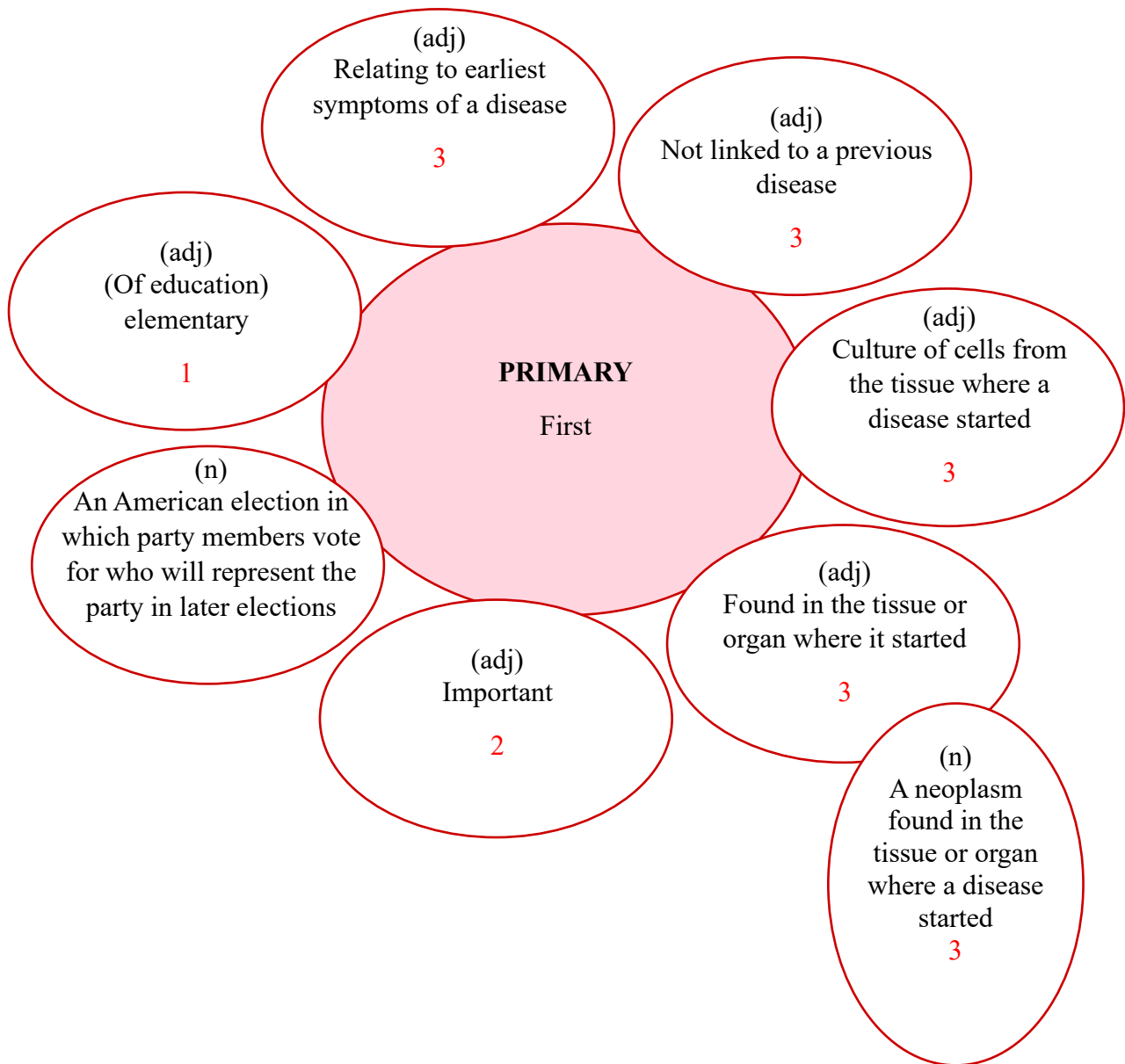
2

(v)

To remove the
outer layer of
something

2





PRIOR
Before

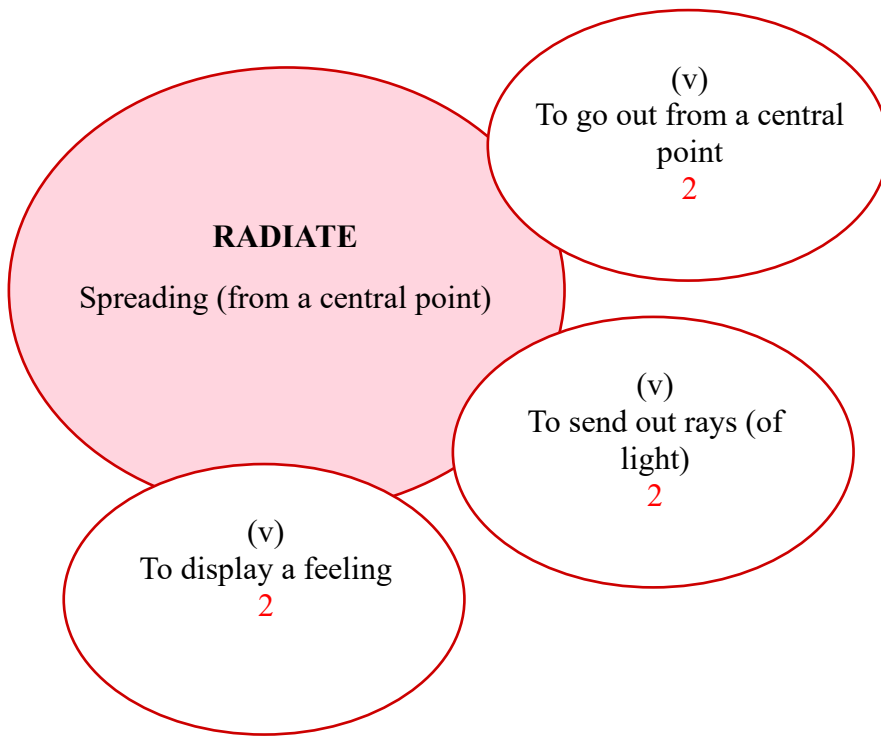
(adj)
Earlier
2

(adv)
Happening
before
2

RADICAL (adj)
Nontraditional

RADICAL
Relating to a root

(adj)
(Of treatment) working
against the root of a
disease or tumour, etc.
2



REFLEX (n)
A copy of an original

REFLEX
A response

(n)
An automatic response
to a stimulus
2

(adj)
Done as an
automatic
response
2

RESOLVE (n)
Determination to do something

RESOLVE
To bring to an end

(v)
To end a disease
2

(v)
To solve a problem
2

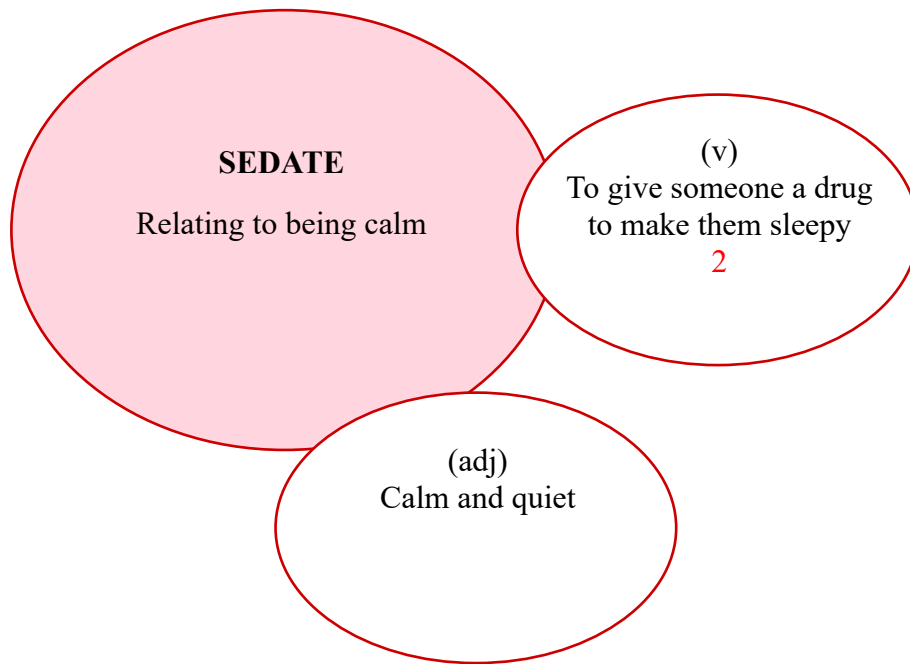
SECRETE (v)

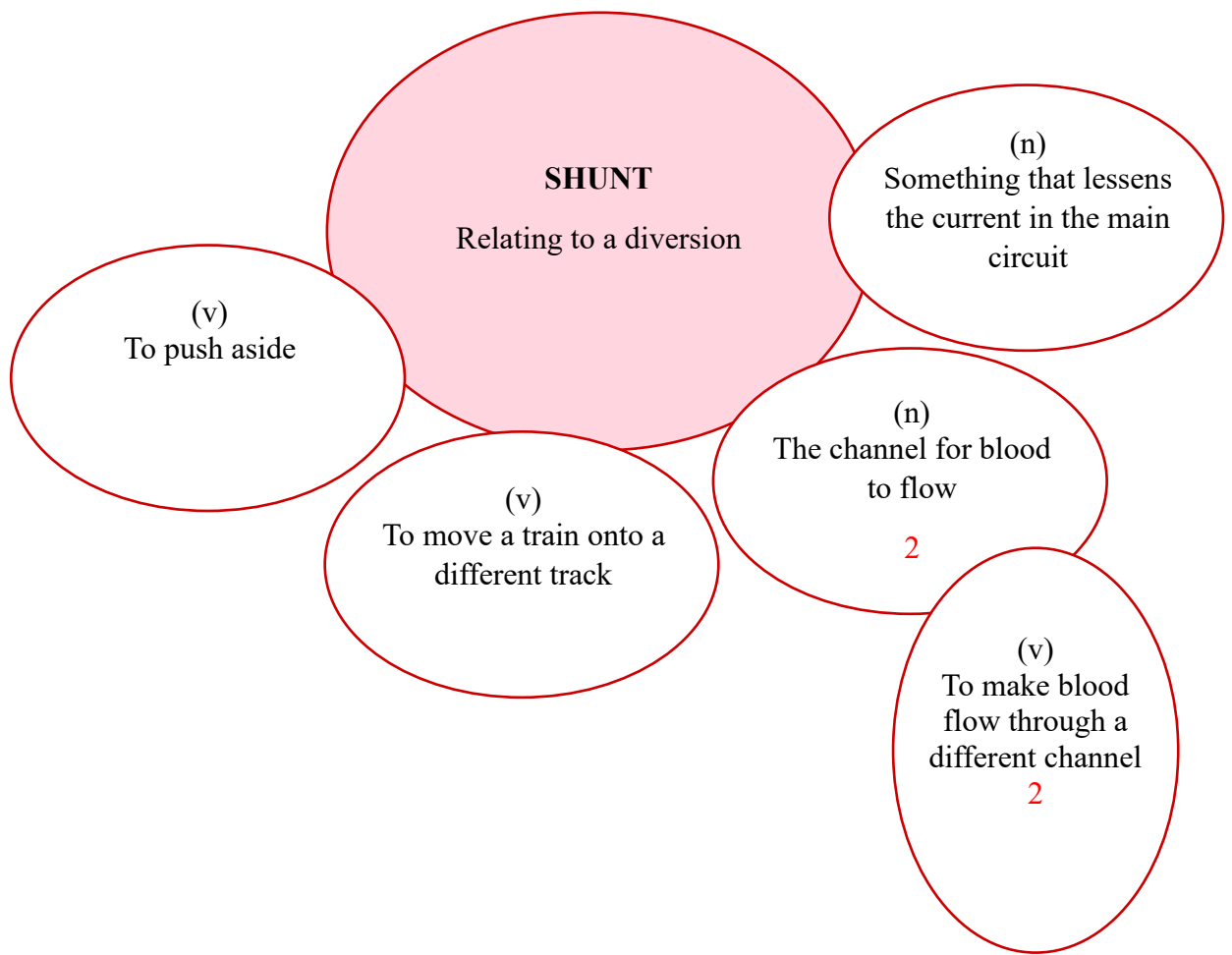
To hide something out of sight

SECRETE (v)

To make and release a liquid

2





STEM (v)
To stop something flowing

STEM
Relating to a
supporting structure

(n)
The main part of a
supporting structure
2

(ACRONYM) **STEM** =
Science Technology
Engineering and
Mathematics field/education

(v)
To have its origin in
2

STOOL (n)

A wooden seat

2

STOOL (n)

Solid waste from the body

2

TUMOUR (n)

A lump caused by disease

2

APPENDIX 4. Ethics approval (The University of Adelaide)



RESEARCH SERVICES
OFFICE OF RESEARCH ETHICS, COMPLIANCE
AND INTEGRITY
THE UNIVERSITY OF ADELAIDE

LEVEL 4, RUNDLE MALL PLAZA
50 RUNDLE MALL
ADELAIDE SA 5000 AUSTRALIA

TELEPHONE +61 8 8313 5137
FACSIMILE +61 8 8313 3700
EMAIL hrec@adelaide.edu.au

CRICOS Provider Number 00123M

Our reference 35448

11 January 2022

Dr Julia Miller
School of Education

Dear Dr Miller

ETHICS APPROVAL No: H-2022-004
PROJECT TITLE: Piloting SemiMed—a mini semantic visualization dictionary of English semi-technical medical vocabulary for upper intermediate learners of medical English

The ethics application for the above project has been reviewed by the Low Risk Human Research Ethics Review Group (Faculty of Arts and Faculty of the Professions) and is deemed to meet the requirements of the *National Statement on Ethical Conduct in Human Research 2007 (Updated 2018)* involving no more than low risk for research participants.

You are authorised to commence your research on: 11/01/2022
The ethics expiry date for this project is: 31/01/2025

NAMED INVESTIGATORS:

Chief Investigator: Dr Julia Miller
Student - Postgraduate
Doctorate by Research (PhD): Miss Chinh Ngan Nguyen Le
Associate Investigator: Dr Stephen John Kelly

CONDITIONS OF APPROVAL: Thank you for your response to the matter raised. The revised application provided 23.12.2021 has been approved.

Ethics approval is granted for three years and is subject to satisfactory annual reporting. The form titled Annual Report on Project Status is to be used when reporting annual progress and project completion and can be downloaded at <http://www.adelaide.edu.au/research-services/oreci/human/reporting/>. Prior to expiry, ethics approval may be extended for a further period.

Participants in the study are to be given a copy of the information sheet and the signed consent form to retain. It is also a condition of approval that you immediately report anything which might warrant review of ethical approval including:

- serious or unexpected adverse effects on participants,
- previously unforeseen events which might affect continued ethical acceptability of the project,
- proposed changes to the protocol or project investigators; and
- the project is discontinued before the expected date of completion.

Yours sincerely,

Ms Amy Lehmann
Secretary

The University of Adelaide

APPENDIX 5. Ethics approval (A University of Medicine and Pharmacy in Vietnam – name deleted for anonymity)



Approval number: H2022/015


February 25th, 2022

Ms. Nguyen Le Ngan Chinh,



Subject: Approval of the study “Piloting SemiMed – a mini semantic visualization dictionary of English semi-technical medical vocabulary for upper intermediate learners of medical English”.

Dear Ms. Nguyen Le Ngan Chinh,

The Institutional Ethics Committee of  has reviewed and approved the following study:

Project title: “Piloting SemiMed – a mini semantic visualization dictionary of English semi-technical medical vocabulary for upper intermediate learners of medical English”

Investigator:


Ms. Nguyen Le Ngan Chinh., Lecturer, Hue University of Foreign Languages, Vietnam; PhD candidate, The University of Adelaide, Australia

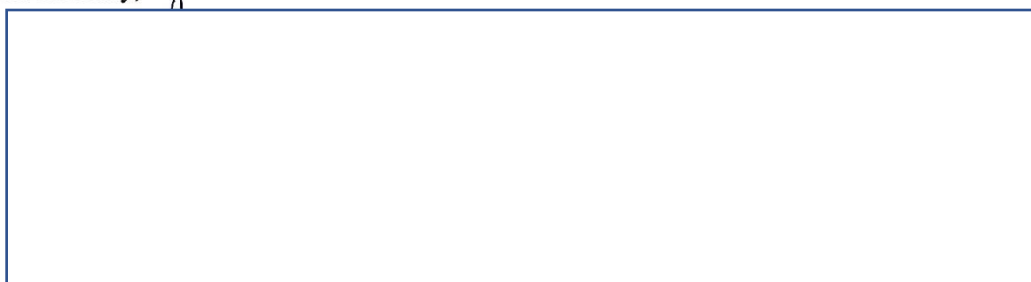
This project is approved for the research period from **March to December 2022.**

It is your responsibility to ensure that all people associated with the study are made aware of what has been actually approved.

Please note that the following conditions apply to your approval. Failure to abide by these conditions may result in suspension or discontinuation of approval and/or disciplinary action.

- a. Limit of Approval: Approval is limited strictly to the study as submitted in your application
- b. All procedures within this study must follow what have been submitted in your ethics application.
- c. Approval is for the above mentioned period. Research must be renewed (if needed) until it is complete.

Yours sincerely, 



APPENDIX 6. Invitation email to rector (A University of Medicine and Pharmacy in Vietnam)

This invitation is being sent out on behalf of the researcher and your personal details have not been provided to the researcher.

Dear [Name]

I am a former lecturer of English for Medical Purposes at X University of Foreign Languages and am now studying for a PhD at the University of Adelaide, Australia.

I am writing to ask your permission for me to invite students of English for Medical Purposes via the English Club to take part in a one hour online trial of a new vocabulary learning resource. Each participant will receive VND 250,000. The trial will allow the students to practise their medical vocabulary and will help me in the development of a resource to assist students of English for Medical Purposes. Their names, and the name of the university, will remain anonymous in my PhD thesis and any related publications.

There is more information in the attached information sheet.

I look forward to hearing from you.

Kind regards

Chinh Ngan NGUYEN LE chinngan.nguyenle@adelaide.edu.au.

APPENDIX 7. Invitation email to participants (A University of Medicine and Pharmacy in Vietnam)

This invitation is being sent out on behalf of the researcher and your personal details have not been provided to the researcher.

Dear students

I am a former lecturer of English for Medical Purposes at X University of Foreign Languages and am now studying for a PhD at the University of Adelaide, Australia.

I am looking for students learning English for Medical Purposes to take part in a one hour online trial of a new vocabulary learning resource. Each participant will receive VND 250,000. Your name, and the name of the university, will remain anonymous in my PhD and any publications related to my study.

There is more information in the attached information sheet.

If you would like be involved in the study, please contact me at chinhngan.nguyenle@adelaide.edu.au.

Kind regards

Chinh Ngan NGUYEN LE (PhD student at the University of Adelaide, Australia)

APPENDIX 8. Participant information sheet

PROJECT TITLE: Piloting SemiMed—A mini semantic visualization dictionary of semi-technical medical vocabulary: A response to semantic deficiencies in a medicine-related wordlist

HUMAN RESEARCH ETHICS COMMITTEE APPROVAL NUMBER: H-2022-004

PRINCIPAL INVESTIGATOR: Dr Julia Miller

STUDENT RESEARCHER: Ms Chinh Ngan Nguyen Le

STUDENT'S DEGREE: Doctor of Philosophy

Dear Participant,

You are invited to participate in the research project described below.

What is the project about?

This research project is about piloting SemiMed—a semantic visualization dictionary of 302 semi-technical medical words. A sample of SemiMed, which includes 30 *visualized semantic diagrams* of piloted semi-technical medical words, will be introduced to 30 EFL (English as a Foreign Language) medical students through custom-written medical scenarios. The students will be encouraged to consult the provided lexical resources to use the piloted words as appropriately as they can in the scenarios. Follow-up focus groups will be conducted to get opinions from the students on their usage of the SemiMed sample.

Who is undertaking the project?

This project is being conducted by Chinh Ngan Nguyen Le and funded by the A.S. Hornby Dictionary Research Awards. This research will form the basis for the degree of Doctor of Philosophy at the University of Adelaide under the supervision of Dr Julia Miller and Dr Stephen Kelly.

Why am I being invited to participate?

You are being invited as you are an EFL student who majors in medical fields and possesses a required level of English, upper-intermediate and above (equivalent to IELTS 5.5 and above).

What am I being invited to do?

You are being invited to join a 60-minute Zoom meeting and get involved in the following activities:

- Induction (15-20 minutes): Receive instruction on how to use SemiMed, especially how to interpret information presented in *visualized semantic diagrams* and be informed of dictionaries you are requested to use
- Role-play (10-15 minutes): Act out medical scenarios and be observed by the student researcher
- Focus group (30 minutes): Give verbal feedback on the usefulness and practicality of SemiMed

How much time will my involvement in the project take?

The Zoom meeting will take up to 60 minutes of your time, at a time convenient to you.

Are there any risks associated with participating in this project?

There are no foreseeable risks in this research other than the time (around an hour) involved in induction, role-play and focus group.

If you are uncomfortable at any time during the role-play or focus group, it will be discontinued and only recommenced if you wish. If you do not want your role-play or focus group data to be used, it will be destroyed with your consent.

You will be referred to by a pseudonym.

You will be advised to talk to the Department of Student Affairs if you feel any distress during the role-play or focus group.

What are the potential benefits of the research project?

The project will provide participants with an opportunity to access the newly developed lexical resource (SemiMed), sharpen their communicative skills through the medical role-play and discuss their experience of using SemiMed. Each participant will receive VND 250,000 as an incentive for his/her participation.

Can I withdraw from the project?

Participation in this project is completely voluntary. If you agree to participate, you can withdraw from the study at any time during the research and up to the time of submission of the thesis. In this case, your information will not be included in the research.

What will happen to my information?

Confidentiality and privacy:

- All participants will be referred to by a pseudonym in any published research findings. While all efforts will be made to remove any information that might identify you, as the sample of participants is small, and only one university is involved, complete anonymity cannot be guaranteed. However, the utmost care will be taken to ensure that no personally identifying details are revealed. The confidentiality and privacy of all participants will be upheld and their views and opinions will not be publicly accessible in a personally identifiable manner.

Storage:

- The data, including audio or video recordings, will be securely stored on the University of Adelaide servers.
- The principal investigator will keep the records for five years from the date of any publication or public interest.

Publishing:

- Results may be made accessible to the public in the form of a book chapter, journal article, conference presentation, report to the funding body, and Ph.D. thesis, but participants will not be identifiable.

Sharing:

- Participants will be offered a chance to see the transcribed focus group notes within eight weeks of the Zoom meeting, and to make changes, or withdraw data if necessary, within a week of receiving the notes.
- Your de-identified data may be used for future research purposes by any researcher in any field.

Your information will only be used as described in this participant information sheet and it will only be disclosed according to the consent provided, except as required by law.

Who do I contact if I have questions about the project?

If you have questions or inquiries regarding the project, you should contact the student researcher or principal investigator:

Name	Phone number	Email
Ms Chinh Ngan Nguyen Le, Student Researcher	+61 404 754 279	chinhngan.nguyenle@adelaide.edu.au
Dr Julia Miller, Principal investigator	+61 8 8313 4983	julia.miller@adelaide.edu.au

What if I have a complaint or any concerns?

The study has been approved by the Human Research Ethics Committee at the University of Adelaide (approval number H-2022-004). This research project will be conducted according to the *NHMRC National Statement on Ethical Conduct in Human Research 2007 (Updated 2018)*.

If you have questions or problems associated with the practical aspects of your participation in the project, or wish to raise a concern or complaint about the project, then you should consult the Principal Investigator. If you wish to speak with an independent person regarding concerns or a complaint, the University's policy on research involving human participants, or your rights as a participant, please contact the Human Research Ethics Committee's Secretariat on:

Phone: +61 8 8313 6028

Email: hrec@adelaide.edu.au

Post: Level 3, Rundle Mall Plaza, 50 Rundle Mall, ADELAIDE SA 5000

Any complaint or concern will be treated in confidence and fully investigated. You will be informed of the outcome.

If I want to participate, what do I do?

If you would like to participate in this research project, please email Ms. Chinh Ngan Nguyen Le (chinhngan.nguyenle@adelaide.edu.au). She will provide you with a consent form to be signed and returned to her. You will be given a copy of the consent form and this information sheet for your personal documentation.

Yours sincerely,

Ms Chinh Ngan Nguyen Le

Dr Julia Miller

APPENDIX 9. Consent form

1. I have read the attached Information Sheet and agree to take part in the following research project:

Title:	Piloting SemiMed—A mini semantic visualization dictionary of semi-technical medical vocabulary: A response to semantic deficiencies in a medicine-related wordlist
Ethics Approval Number:	H-2022-004

2. I have had the project, so far as it affects me, and the potential risks and burdens fully explained to my satisfaction by the research worker. I have had the opportunity to ask any questions I may have about the project and my participation. My consent is given freely.
3. Although I understand the purpose of the research project, it has also been explained that my involvement may not be of any benefit to me.
4. I agree to participate in the Zoom meeting outlined in the participant information sheet.
5. I agree to be:
- Audio recorded Yes No
- Video recorded Yes No
6. I understand that I am free to withdraw from the project at any time and that this will not affect my study at the X University of Medicine and Pharmacy, now or in the future.
7. I have been informed that the information gained in the project may be published in a book/journal article/conference presentations/report/thesis.
8. I have been informed that in the published materials I will not be identified and my personal results will not be divulged.
9. I hereby provide ‘unspecified’ consent for the use of my data in any future research:
- Yes No

10. I understand my information will only be disclosed according to the consent provided, except where disclosure is required by law.
11. I am aware that I should keep a copy of this Consent Form, when completed, and the attached Information Sheet.

Participant to complete:

Name: _____ Signature: _____

Date: _____

APPENDIX 10. Medical scenarios

Medical Scenario 1: Bowel

Specialist: We'll have to do tests, but I think you may have a **tumour**¹ of the colon.

Patient (scared): What does that mean?

Nurse: It means . . .

Specialist: What colour are your **stools**²?

Patient: My what?

Nurse: When you go to the toilet. Your . . . What colour is it?

Patient: Black.

Specialist: It might be nothing, but we need to do an operation. You may have a tumour. It might be **benign**³. That means . . . Or we may have to do a **radical**⁴ operation. That means . . .

Patient: You think I've got cancer?

Nurse: Maybe. But we don't know until we do the operation.

Patient: Oh. What does my **colon**⁵ do?

Nurse: It **absorbs**⁶ water and moves the waste along so it can be passed out.

Specialist: But don't worry. There is a very good chance of removing any cancer if we find it quickly.

Medical Scenario 2: Eye

Specialist: Hello, X. Thank you for coming today. Do you know why you're here?

Patient: Yes. I can't see properly.

Specialist: Can you look at me? Now look at the nurse. I want to see your eye **reflexes**¹.

Patient: What do you mean?

Nurse: That means . . .

Specialist: And in your case you have a **cataract**².

Patient: What's that?

Nurse: It means . . .

Patient: But I wear glasses.

Specialist: Yes. You are short sighted. But a cataract is a **chronic**³ eye condition.

Patient: You mean it's bad?

Nurse: Maybe. It means it . . . But we can **resolve**⁴ the problem with a small operation.

Patient: Why have I got a cataract?

Specialist: There could be lots of reasons. You also have diabetes, so that **predisposes**⁵ you to cataracts too.

Patient: Predisposes?

Specialist: [Explains] . . . Your eyes are also **secreting**⁶ more mucus than is normal.

Patient: Secret?

Nurse: [Explains] . . .

Specialist: But we can treat that with eye drops. We'll arrange your cataract surgery next time we see you. It's not urgent.

Medical Scenario 3: Heart

Patient: What's wrong with me?

Specialist: You have **acute**¹ coronary syndrome.

Patient: Cute? Like kittens?

Nurse: No, 'acute'. That means . . .

Specialist: Have you had any **prior**² **cardiac**³ problems?

Patient: Uh?

Specialist: [Explains] . . .

Nurse: Your heart is what makes your blood **circulate**⁴.

Patient: Circle?

Nurse: [Explains] . . .

Specialist: We thought maybe there was a **shunt**⁵ between the right and left sides of your heart. In other words, . . . But now we know that's not the case.

Patient: So I didn't have a heart attack?

Nurse: No. You had pain **radiating**⁶ into your arms.

Specialist: [Explains] . . . But we are sure it's acute coronary syndrome. Make an appointment and I'll see you again next week.

Medical Scenario 4: Liver

Specialist: You have a problem with one of the **lobes**¹ in your **liver**².

Patient: Lobes? Like ear lobes?

Nurse: No. Your liver is divided . . .

Patient: What is my liver, anyway? What does it do?

Nurse: Your liver . . .

Specialist: You have a **disorder**³ of the lobe. That means . . . There is a risk of infection. It's only **moderate**⁴.

Patient: Moderate?

Nurse [Explains] . . .

Specialist: We'll need to cut through the **fascia**⁵.

Patient (alarmed): Cut my face?

Specialist: No. 'Fascia' refers to . . . But don't worry. We'll **sedate**⁶ you before the operation.

Nurse: [Explains] . . . You won't know anything about it till you wake up.

Patient (still a bit worried): OK.

Specialist: Don't worry. I've done this operation hundreds of times. You'll be fine.

Medical Scenario 5: Pregnancy

Nurse: Hello X. Please sit down. Do you know why you're here?

Pregnant patient: No.

Specialist: We would like to undertake some screening tests just to make sure your baby is growing well and doesn't have any **defects**¹. A defect is . . . For example, a problem with the heart.

Nurse: We **conduct**² tests like this on all pregnant women. Mostly it's fine, but sometimes, towards the end of the pregnancy if the baby is not growing enough, we might have to **induce**³ labour.

Pregnant patient: Induce?

Nurse: That means . . .

Specialist: Sometimes the problem is **compounded**⁴ by a range of factors.

Pregnant patient: Arrange the factors?

Nurse: No. 'Compound' means . . . Many things could happen. For instance, some cells can **migrate**⁵ from the baby to the mother and cause problems for the mother.

Pregnant patient: I'm not emigrating!

Specialist: No. It means . . . Some cells like carbon dioxide **diffuse**⁶ from the baby to the mother which is normal. It allows the CO₂ produced by the baby to cross over into the mother so she can breathe it out.

Nurse: Usually everything is fine. We'll make an appointment for a blood test and scan for you next week.

Pregnant patient: Thank you.

* [. . .] indicates the participant has to improvise.

APPENDIX 11. Focus group questions

The topic today is about your experience in using SemiMed to act out provided medical scenarios.

1. What course(s) are you studying?
2. What dictionaries do you normally use to help you understand medical words?
3. Have you had any problems finding information in those dictionaries?
4. What are your general feelings about SemiMed?
5. What positive experiences did you have in using SemiMed to help you understand and use pilot words in medical scenarios?
6. Do you think SemiMed would work so well in other situations?
7. What are SemiMed's key strengths?
 - a. *Which SemiMed's aspect(s) do you find beneficial?*
 - b. *What might influence and motivate you to choose SemiMed over other conventional dictionaries?*
8. What are SemiMed's key weaknesses?
 - a. *What specific concerns did you face when using SemiMed?*
 - b. *If you could choose a feature of SemiMed that needs to be improved, what would you choose and why?*
9. If you could add any feature to SemiMed, what would it be? Why?

* Questions in *italics* are prompt questions and will only be used if necessary.