



Design of a hybrid MWE tool

Research Proposal

Gabriel Avbenake
MSc Student
Applied Computing
1902228@student.uwtsd.ac.uk
University of Wales Trinity St David

Table of Content

Table of Content	1
Table of Figures.....	1
Introduction	2
Background to the Proposed Research.....	2
Classification of MWEs.....	2
What is Multiword Expression (MWE).....	3
Rationale for the Research.....	4
Aims and Objectives.....	4
Expected Outcomes	5
Research Methods	5
Philosophy.....	5
Approach.....	6
Strategy	7
Data Collection & Data Analysis Methods	7
Data Evaluation	8
Risks and Limitations.....	8
Development Methods	9
Software Development Methodology	9
Resources & Planning	11
Resources Required	11
Gantt Chart	11
References	16

Table of Figures

Figure 1: Classification of Multiword Expressions	2
Figure 2: Research Onion Diagram (based upon Saunders et al's diagram, 2007)	6
Figure 3: The different phases in a V-Model of the SDLC (TutorialsPoint, n.d.)	10
Figure 4: Task Sheet View for the Proposed Project.....	12
Figure 5: Gantt Chart for the Proposed Project, View 1	13
Figure 6: Gantt Chart for the Proposed Project, View 2	14
Figure 7: Gantt Chart for the Proposed Project, View 3	15
Figure 8: Resource Sheet of Proposed Project	15

Introduction

Background to the Proposed Research

The field of Natural Language Processing (NLP) has developed rapidly over the recent decades. NLP is a branch of computer science that enables computers to extract meaning from unstructured text (Dickson, 2020) (Manning & Schütze, 1999). NLP is important because it fosters communication, knowledge management and discovery (Eisenstein, 2019).

Machine translation (MT), information retrieval (IR) especially web search, Optical Character Recognition (OCR), digital assistants, Spell checking and grammar corrections, Morphological and Syntactic analysis (e.g. POS taggers and Parsers), other computational semantics (e.g. question-answering, automatic summarisation, text simplification), foreign language learning, amongst others, are some of the applications of NLP (Jurafsky & Martin, 2008). These have greatly impacted our lives in this information age era.

Multiword expressions (MWEs) are part and parcel of what NLP has to deal with. MWEs are replete in our everyday speech and writings. Idioms such as he kicked the bucket, Shoot the breeze, rack one's brains etc are some examples of MWEs. Apart from idiomatic expressions, many other word combinations present some peculiarities which allow us to see them as MWEs (Ramisch, 2015). MWEs can occur in a wide range of lexical and syntactic configurations (e.g. nominal, verbal and conjunctions, etc).

Classification of MWEs

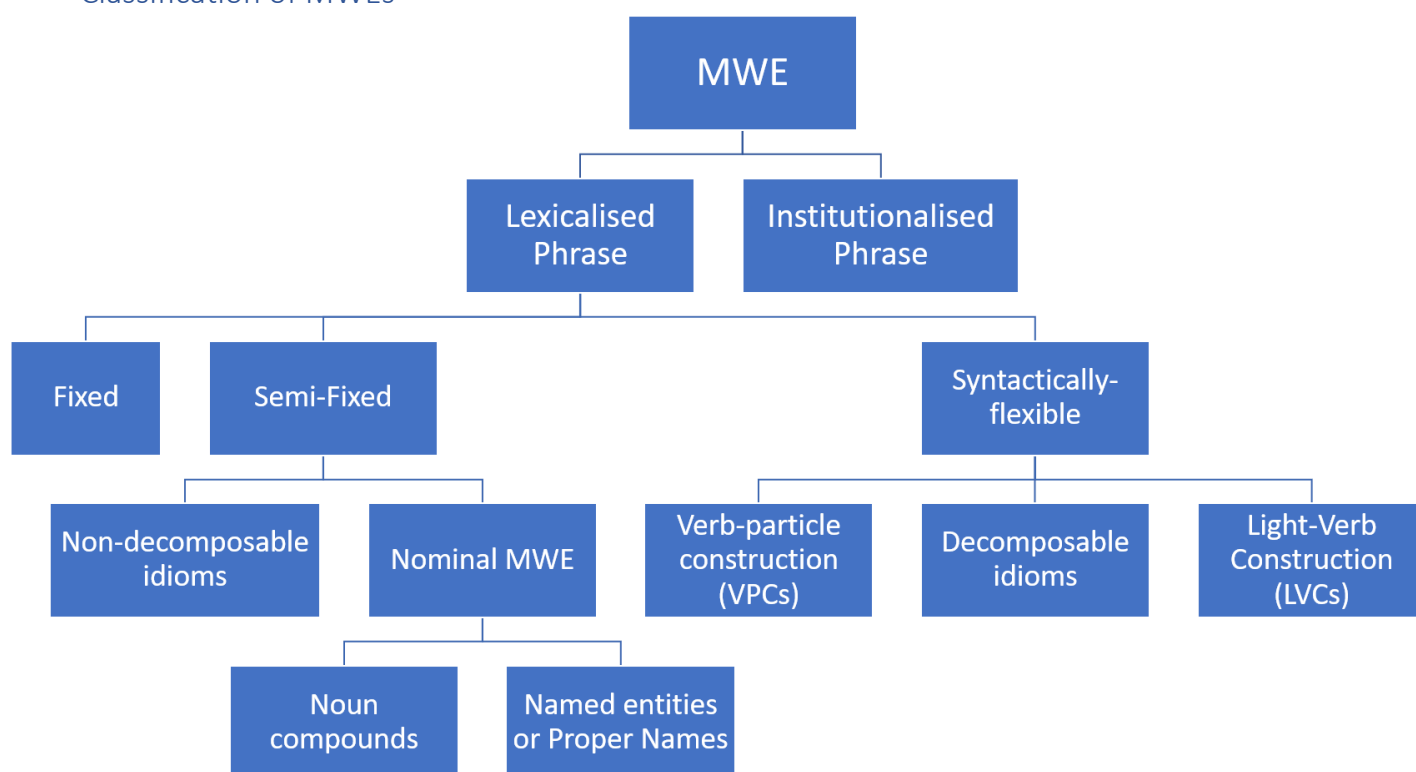


Figure 1: Classification of Multiword Expressions

Fig 1 is an adaptation from Baldwin and Kim diagram (Baldwin & Kim, 2010). This gives an idea of the depth and breadth of MWEs. This is not the only classification of MWEs out there. For example, Gantar et al has a different version of classification (Gantar, et al., 2019). The version in Fig 1 above

was chosen because it would assist, from the programming point of view, to identify MWEs in corpora. The version chosen focuses more on the structure of the MWES which could help in identifying potential MWEs from corpora.

Lexicalised Phrase/Expressions are MWEs that have been explicitly encoded in the lexicons, while Institutionalised Phrase/Expressions are MWEs that have not been. Lexicalised Expressions are composed of Fixed, Semi-fixed and Syntactically-flexible MWEs. The fixed MWEs show some degree of fixedness in terms of structure (Baldwin & Kim, 2010). They do not undergo any lexical modifications. They include prepositional phrases like on air, on top of, in addition to, by the way, etc. Further examples include by and large, in short, ad hoc, etc. There are also conjunctions like (as well as, such as) and prepositional phrases (in front of, next to, with regards to). You could not change any word in it, you cannot say: by and larger OR by and very large.

Semi-Fixed expressions are lexically-variable MWEs. One could not alter their word order and composition, but they could undergo some degree of lexical variation (Baldwin & Kim, 2010). E.g. Kick/kicked/kicks the bucket but the bucket was kicked by him is not the same thing. Semi-Fixed MWEs consists of Non-decomposable idioms (also known as Verb-noun Idiomatic combination – VNIC, in terms of their structure) and nominal MWEs. Nominal MWEs are made up of nominal compounds (NC) (such as loan shark, white trash, attorney general, part of speech, open secret, connecting flights, French kiss) or Named entities (such as San Francisco 49ers).

Syntactically flexible expressions are MWEs which undergo syntactic variations (Baldwin & Kim, 2010). They consist of Verb-Particle Construction (VPCs), Verb-noun constructions and Decomposable idioms (VNICs). Verb-Particle Construction (VPCs) or Phrasal Verbs consists of a verb plus one or more particles. E.g. write up, find out, make out, come across. Verb-noun constructions or Light-Verb Constructions (LVCs) consists of support verbs and nouns or prepositional phrases. E.g. give confidence, make a living, give a lecture, come into bloom. Decomposable idioms (VNICs) are distinguishable from non-decomposable idioms in that semantic decomposition is possible here, whereas it is not possible in non-decomposable idiom. In other words, it is possible to associate components of the decomposable VNIC with distinct elements of the idiom's interpretation. Examples of decomposable idioms are rack one's brains, spill the beans, pull strings, Let the cat out of the bag, save your breath, cut me a break, buy me some time, etc. On the other hand, examples of non-decomposable idioms are kick the bucket, shoot the breeze, get the hang of it, etc.

What is Multiword Expression (MWE)

To be clear, not all word combinations are MWEs, e.g. Dark Chocolate is not an MWE. Though MWEs are quite easily understood by the native speakers of a language, it is hard to identify what features distinguish MWEs from free word combinations (Gelbukh & Kolesnikova, 2013). What constitute MWEs and what does not is controversial. It is paramount that the features of MWEs are determined and numerated, if their identification, extraction and treatment in NLP applications is going to be possible. A universal agreement on the features of MWEs is difficult partly because of the multifaceted nature and types of MWEs. Some opine that translation into another language is one way of detecting MWEs.

In literature, there are two useful definitions, which shall be the working definition of this research:

- Multiword expressions (MWEs) are lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity (Baldwin & Kim, 2010).

- MWEs are word combination whose meaning might not be fully determined from the meaning of their components (Atkins & Rundell, 2008).

These two attempts would guide this research. They would be used to guide our choice of MWEs.

Rationale for the Research

Building NLP applications that are capable of dealing with MWEs is very important because (1) MWEs are ubiquitous in everyday language (Biber, et al., 1999) (Gantar, et al., 2019) and (2) Ignore MWEs in NLP tasks generate ungrammatical, ambiguous or unnatural output (Ramisch, 2015) leading to probable information loss. This is particularly true in the area of Optical character recognition (OCR), Morphological and syntactic analysis (POS taggers and Parsers), Information Retrieval, Foreign language learning, Spell checking and grammar corrections, Machine translation (MT) and other computational semantics (e.g. Automatic summarisation, text simplification, question-answering etc).

Despite their prevalence, MWEs have often been neglected in the development of NLP applications. MWEs identification has been the focus instead of the integration of automatically acquired or manually compiled MWE resources into applications (Kordoni et al 2011, 2013, 2014). Identification and integration of MWEs are not easy problems to solve because of the pervasive and complex nature of MWEs in language (Gelbukh & Kolesnikova, 2013), especially the diversity of their semantic, lexical, statistical and syntactical properties.

Since 2003, an annual workshop on MWEs has been organised by SIGLEX-MWE section or endorsed by SIGLEX (SIGLEX-MWE, n.d.). SIGLEX is a Special Interest Group of the Association for Computational Linguistics (ACL) that focuses on Lexicon (Association for Computational Linguistics (ACL), 2020). SIGLEX provides an umbrella for research interests on lexical issues ranging from lexicography and the use of online dictionaries to computational lexical semantics (SIGLEX, 2016). SIGLEX-MWE is a section of SIGLEX that is dedicated to promoting scientific activity on MWEs (SIGLEX-MWE, n.d.).

The 2020 SIGLEX-MWE Workshop, tagged MWE-LEX 2020, will be a Joint Workshop on Multiword Expressions and Electronic Lexicons (SIGLEX-MWE, 2020). The organizers opine that appropriately built electronic MWE lexicons is important for NLP. However, building such MWE e-lexicons raises lots of issues such as the lemmatization and standardised representation of morphological, syntactic and semantic properties of MWEs. The workshop is proposed to pave the way towards a better understanding of these issues, and to foster convergence and scientific innovation. They hope for a large standardised multilingual, possibly interconnected, NLP-oriented MWE lexicons. They submit that the development of such lexicons is challenging and calls for tools which would leverage, on the one hand, MWEs encoded in pre-existing NLP-unaware lexicons and, on the other hand, automatic MWE discovery in large non-annotated corpora (SIGLEX-MWE, 2020). This research is an attempt in this regard.

Aims and Objectives

The aim of this research is to design a semi-supervised system that can tag MWEs in text and be connected to a lexicon simultaneously. In other words, to design a hybrid system that can identify

MWEs, link them to corresponding entries in lexica and augment existing resources with possible new expressions.

The objectives are:

- To design a tool which will automatically discover MWEs in large non-annotated corpora to enrich MWE lists from traditional human-readable lexicons.
- To evaluate how useful the tool is, in terms of accuracy.
- To analyse the possibility of easy integration of the enriched lexicon into NLP applications.

Expected Outcomes

At the end of the project, it is expected that there would be a tool that will produce an enriched MWE lexicon, with minimal human effort, from various corpora.

The enriched MWE would be in a format ready to be plugged into the pipeline of various NLP applications.

This would improve NLP applications to generate less ambiguous, natural and grammatical output to prevent information loss and boost confidence in NLP output.

This would ultimately foster communication, knowledge management and discovery.

Research Methods

Philosophy

The research onion model, presented by Saunders, Lewis and Thornhill in their book titled Research Methods for Business Students, has become the standard for presenting research methods. The model is illustrated in Fig 2, obtained from <https://onion.derby.ac.uk/> (University of Derby, n.d.). Following that model, the philosophical stance from which this research would be conducted is Positivism.

This research would be based on valid scientific knowledge and methods. It is going to be characterized by the design and testing of a tool derived from the existing theory of knowledge. The research questions/objectives would be centred on the design and testing of the MWE tool. These questions/objectives would be derived from existing theory of scientific knowledge, using scientific methods. As such, the questions/objectives can be tested and allows explanations that are measurable against accepted knowledge. The result would be quantifiable, lending themselves to statistical analysis (University of Derby, n.d.).

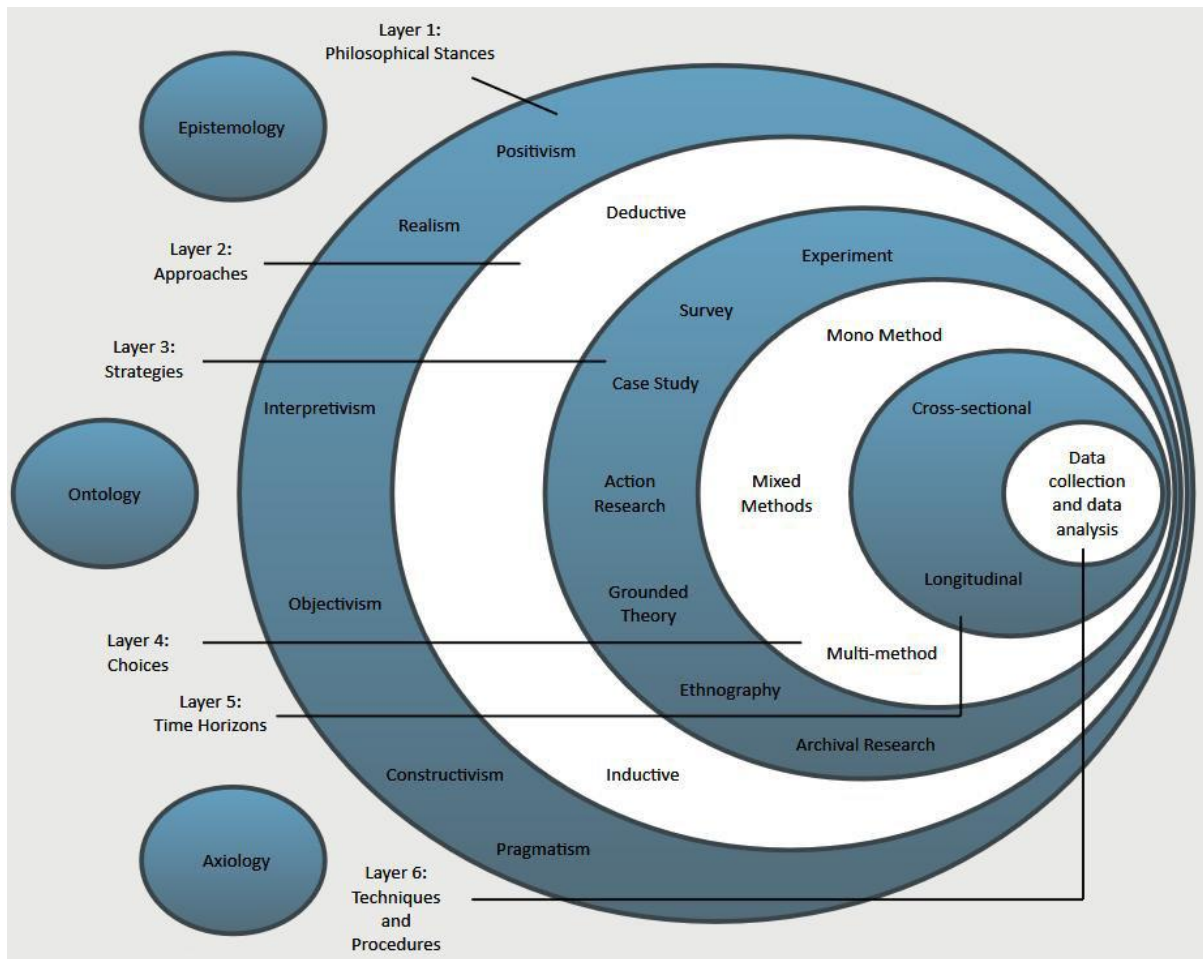


Figure 2: Research Onion Diagram (based upon Saunders et al's diagram, 2007)

Approach

Still following the research onion model, on the second layer, the deductive approach would be used in this research. This is particularly suited to the positivistic philosophical stance (Snieder & Larner, 2009). Following the deductive approach, the research would begin with development of hypothesis based on pre-existing theory/theories from literature review as well as formulating the approach to test it (Silverman, 2013). Specifically, looking at literature, the requirements of the tool to be designed would have to be formulated, as well as how the designed tool would be evaluated.

The aim in the deductive approach is to find the answers to the research questions/statement that are set at the start of the research using pre-existing knowledge and theories in literature. In this research, after the tool is designed guided by research questions and findings, the tool would be used to gather data (MWEs) in corpora and its results analysed and evaluated, using accepted statistical methods to help confirm or reject or improve the tool that is the result of the research questions in the first place. Following the deductive approach, the research would be characterized by the development from general to particular: the general theory or knowledge base is first established. Then, the specific knowledge gained from the research process, in this case the designed tool, is then tested through data collection and evaluation (Kothari, 2004). In the end, the findings derived would either help confirm or reject or improve the designed tool or research questions or hypothesis.

Strategy

The research strategy describes how the researcher intends to carry out the work (Saunders, et al., 2007). It is concerned with how a researcher plans to collect data for the research and analyse them. Available strategies include Experiment, Survey, Case study, Action research, Grounded theory, Ethnography or Archival research. The strategy is chosen based on the data required for the research and the purpose of the study. Of course, more than one strategy could be chosen if justifiable.

Action research is the strategy to be adopted in this study because it is primarily designed to deal with a specific problem in a specific situation. MWEs is a specific problem in the class of lexical units that NLP has to deal with. The process of action research begins with setting up of a clear objective. Thereafter, the problem is fully diagnosed and subsequently, a list of solutions is prepared and presented as recommendations to solve the issue (University of Derby, n.d.). This research begins with the clear objective of extracting MWEs in corpora to augment existing MWE lexicons and present the new MWE e-lexicon in a format that is NLP-aware and ready. This problem is fully diagnosed; from pre-existing knowledge in literature, the following sub-questions, amongst others, are researched:

- 1) what are the features of MWEs that could distinguish them in corpora,
- 2) how best could these MWEs be extracted using these features,
- 3) which pre-existing tools could best assist in doing these in terms of programming languages, lexicons, corpora and other existing lexical tools,
- 4) what form should be designed tool or solution take
- 5) what format should the resultant MWE lexicon take to be NLP application ready and aware.
- 6) how best can the designed tool or solution be evaluated to check its efficiency and accuracy

At the end of the study, rather than a list of solutions, a tool would be designed and recommended as solution to solve the issue or the clear set objective.

In the evaluation of the designed tool, some kind of Survey strategy could be used to ask two or more human annotators to label relevant and irrelevant results in the extracted MWEs from corpora. This would, then, be used to calculate the precision and recall rate.

The fourth layer of the research onion, also known as research choice, helps the researcher ensure whether it is ideal to combine both quantitative and qualitative methodology or to use only one methodology. In this study, the mixed method would be used. Both quantitative and qualitative research methods would be used in the process of the study, data collection and analysis. Using this method, the limitations of each individual method can be mitigated.

At the fifth layer of the Onion, the time horizon choices, the cross-sectional method would be used given the time constraint of this study. The duration of data collection and research would be confined to a short period of time.

Data Collection & Data Analysis Methods

As earlier stated, both quantitative and qualitative methods will be used to collect data with the data being analysed using the appropriate quantitative or qualitative technique.

To identify the requirements for the MWE tool and answer other questions which has been raised above such as (1) what are the features of MWEs that could distinguish them in corpora, (2) how

best could these MWEs be extracted using these features, etc, a review of literature would be carried out to get scholars consensus where they exists. Where no such consensus exists, current scholarly works would be used to back decisions of one choice over another. All the knowledge gained would help in fulfilling all the requirements in Software development process to eventually design the MWE tool. Some of the findings could mean implementing some statistical linguistic measures in the MWE tools such as Word Frequency Distributions, N-Grams, Lexical Association Measures, etc (Ramisch, 2015).

After the MWE tool is developed, it would be used to extract MWEs (data) from corpora. The corpora would have to be representative of the target language if the results are to be credible. Certainly, the meaning of representative depends on the context (e.g. application, domain, genre, sublanguage) (Ramisch, 2015). In this study, the focus would not be multilingual due to time constraint. Our target language would only be the English language and our focus would only be written corpora like the 100-million words British National Corpus and a collection of news from the British Broadcasting Corporation (BBC). The British National Corpus was chosen because it is free, vast and contains words of text texts from a wide range of genres (e.g. spoken, fiction, magazines, newspapers, and academic) (Oxford University Press, n.d.). The BBC News website was chosen because it contains news from various aspects such as politics, sports, fashion, health, business, entertainment and art, science and technology, amongst many others. This would ensure that the data collection (MWE extraction) would not be limited to just one specific knowledge domain or sublanguage. More so, since language is also evolving, the use of a news website would ensure the MWE extraction is also current.

Data Evaluation

The extracted MWE lists extracted by the MWE tool would be given to human annotators to label relevant and irrelevant results about them. This would, then, be used to calculate the precision and recall rate. An online annotation tool might be used that would enable the annotators to label results online and easily.

Certainly, there are issues with this method of data evaluation. Generally, MWE extraction tool evaluation is an issue, requiring a research in its own stead (Ramisch, 2015). Ramisch opined that instead of an intrinsic evaluation of MWE tools (evaluating it by itself), an extrinsic evaluation would be used that evaluates it by judging its usefulness or impact in improving an existing NLP application (Ramisch, 2015). The pros and cons of the two methods lies beyond the scope of this project. Thus, the human annotation method would be used, while noting its limitations.

Risks and Limitations

There are several issues that have been identified and these will need to be resolved before and during the project.

On the risk side, the following have been identified:

1. Time Constraint
2. Poor time management
3. Workload constraints
4. Not meeting all the requirements due to unrealistic goals/plans
5. Unavailability of Linguistic Experts / Human annotators

The use of project planning tools, such as a Gantt chart with PERT analysis should help mitigate the risks surrounding time management and the setting of unrealistic goals. With these tools, the critical paths of the project would be identified and prioritized. MoSCoW prioritisation would help here as well. The goal is to draw up a more realistic timeline. On the issue of workload constraints, the researcher would just have to be mindful of what is due in the coming months. As regards the availability of human annotators, gentle reminders would be used.

The only legal or ethical issues that have been identified currently would be that the researcher should use only open-source or free software or tools for the project. If proprietary or commercial software or tools are necessary, the appropriate permissions and licences must be obtained.

On the limitations for the project, the most significant would be not having all the skills for the project. Though the research has appreciable programming skill in Python and other programming languages, context knowledge might not be adequate in the area of lexicography and computer linguistics. The researcher is not an expert in NLP or Linguistics. Certainly, efforts would be made to gain some knowledge and skill through self-learning. Nonetheless, this would impact on the timescale. Therefore, the best option to this problem is to start off with a simple system and as the build progresses make the system more complex.

Another limitation would be at the evaluation phases. It would have been wonderful to have as many qualified annotators as possible to take place in order to get consensus where there are disagreements in their labelling. This would have compensated for human error and biases. However, given time constraint and availability of willing experts, annotators might be limited to two or three.

Development Methods

Software Development Methodology

Since a software is going to be developed, the development method would be one of the many software development methodology.

Software Development Life Cycle (SDLC) offers a systematic process for building and delivering software applications (Roebuck, 2011). Generally, an SDLC has the following stages: Planning and requirement analysis, Designing project architecture, Development and programming, Testing, and Deployment (Wysocki, 2010). Having explored the various popular SDLC models such as Agile SDLC, DevOps SDLC, Iterative SDLC, Lean SDLC, Spiral SDLC and Waterfall SDLC as well as the Rapid Application Development (RAD) approach, the V SDLC/Model was chosen because of its Verification and Validation functionality.

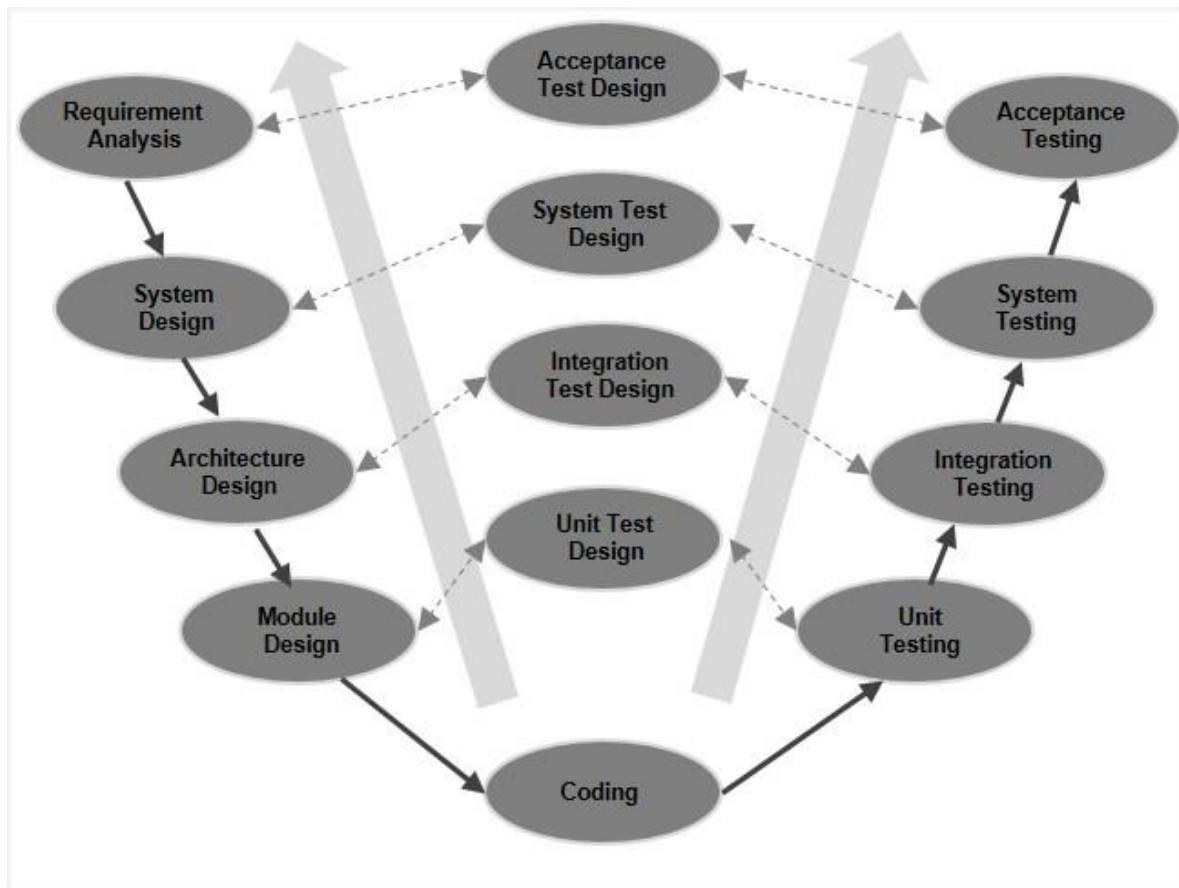


Figure 3: The different phases in a V-Model of the SDLC (TutorialsPoint, n.d.)

The V-shaped SDLC Model is based on the Waterfall model in the sense that its development process is like flowing water, where the development team moves step by step through different SDLC stages. However, unlike the Waterfall model where the testing is done at the end of the project, V Model introduces testing at each and every stage of the development. Every stage ends with testing before the team moves on to the next. Thus, this model is also called Validation-Verification (Font Jr., 2012). Its Verification and Validation functionality helps to check whether all requirements have been fulfilled before proceeding to the next stage.

Since the objectives of this research is clear, the Agile and other iterative options might not be ideal. Also, since user interactions is not key in this project, the RAD approach might not be ideal as well. What is needed is a model that would ensure each step of the development model is correct and validated, to guarantee the end product would be valid and reliable. What is need is a tool what would efficiently and effectively extract MWEs from corpora with minimal human effort and with the highest feasible Precision rate and Recall rate. To achieve this, the important component is the research on the how. Thus, what is needed is a SDLC that emphasizes the SDLC process and validates them. Though, of course, time constraint is always going to be a problem in this research. The MoSCoW prioritisation of scope (into musts, shoulds, coulds and will not haves) would help to adjust the project deliverables to meet the time constraint (Agile Business Consortium, 2019).

Resources & Planning

Resources Required

In order to successfully complete the project, the following resources will be required:

- PC/Laptop – to be used for the building of the system
- Python IDLE (preferably 3.7 version) with various downloadable software packages (such as nltk, matplotlib, pandas, bs4, etc)
- NLP tools such as UCS toolkit, mwetoolkit, Text-NSP, StringNet, POS taggers and Parsers, RASP etc
- Lexicons / Lexical Databases such as WordNet, Cambridge/Acquilex Lexical Database System, BabelNet etc
- General-purpose Corpora such as the 100-million words British National Corpus and the collection of news from the BBC Website, etc
- NLP and Computer Linguistics books and journals that refer to past and current research on NLP and Computer Linguistics
- Books and Journals on Lexicography and Linguistics.
- At least 2 linguistic experts who would act as human annotators

Gantt Chart

In order to plan and manage each step of the project, a Gantt chart has been constructed (see figures 4-8). As shown on Fig 4, the task sheet, the plan has been divided into four distinct sections: Requirements & Specification, Development, Implementation and Finalization of Dissertation. Each task has been listed in their right order under the appropriate section before assigning dependencies to appropriate tasks. Finally, realistic completion dates have been assigned to each task.

Resources needed for the project (see Fig. 8) have been considered and entered into the Resource sheet and assigned to each individual task. The Gantt Chart is broken down here into three views (see Fig 5-7) to make a reasonable view possible.

TASK SHEET		 Task Mode ▾	Task Name ▾	Duration ▾	Start ▾	Finish ▾	Predecessors ▾	Resource Names ▾
	1		▸ Requirements and Specification	50 days?	Fri 01/05/20	Thu 09/07/20		
	2		Carry out Extensive Literature Review	30 days	Fri 01/05/20	Thu 11/06/20		Research for Lit. Review
	3		Undertake Intensive Self Learning on NLP	30 days	Fri 01/05/20	Thu 11/06/20		Books and Journals on Lexicography and Linguistics
	4		Write up Introduction including Aims and Objectives	2 days	Fri 12/06/20	Mon 15/06/20	2	Laptop[1]
	5		Write up Literature Review	5 days?	Tue 16/06/20	Mon 22/06/20	4	Laptop[1]
	6		Write up Methodology	2 days	Tue 23/06/20	Wed 24/06/20	5	Laptop[1]
	7		Write up Requirements & System Specification	5 days?	Tue 23/06/20	Mon 29/06/20	5	Laptop[1]
	8		▸ Development	41 days?	Tue 30/06/20	Tue 25/08/20		
	9		Design Software	12 days	Tue 30/06/20	Wed 15/07/20	7	Laptop[1]
	10		Build Software	24 days?	Thu 16/07/20	Tue 18/08/20	9	Python IDLE[1],NLP tools [1]
	11		Test Software	5 days?	Wed 19/08/20	Tue 25/08/20	10	General-purpose Corpora[1]
	12		▸ Implementation	21 days?	Wed 26/08/20	Wed 23/09/20		
	13		Use Software to Extract MWE from Corpora	2 days?	Wed 26/08/20	Thu 27/08/20	11	General-purpose Corpora[1]
	14		Human Annotation Phase	14 days?	Fri 28/08/20	Wed 16/09/20	13	Human Annotators
	15		Analysis and Evaluation of Result	5 days?	Thu 17/09/20	Wed 23/09/20	14	Laptop[1]
	16		▸ Finalization of Dissertation	5 days?	Thu 24/09/20	Wed 30/09/20		
	17		Write up Presentation and Analysis of Data	2 days	Thu 24/09/20	Fri 25/09/20	15	Laptop[1]
	18		Write up Discussion of Result	1 day	Mon 28/09/20	Mon 28/09/20	17	Laptop[1]
	19		Write up the Conclusion	1 day?	Tue 29/09/20	Tue 29/09/20	18	Laptop[1]
	20		Write up the rest of the Dissertation (Title Page, Acknowledgement, Abstract, Table of Contents, List of Figures and Tables, List of Abbreviations, Glossary, Index, Reference List, Appendices)	1 day?	Wed 30/09/20	Wed 30/09/20	19	Laptop[1]

Figure 4: Task Sheet View for the Proposed Project

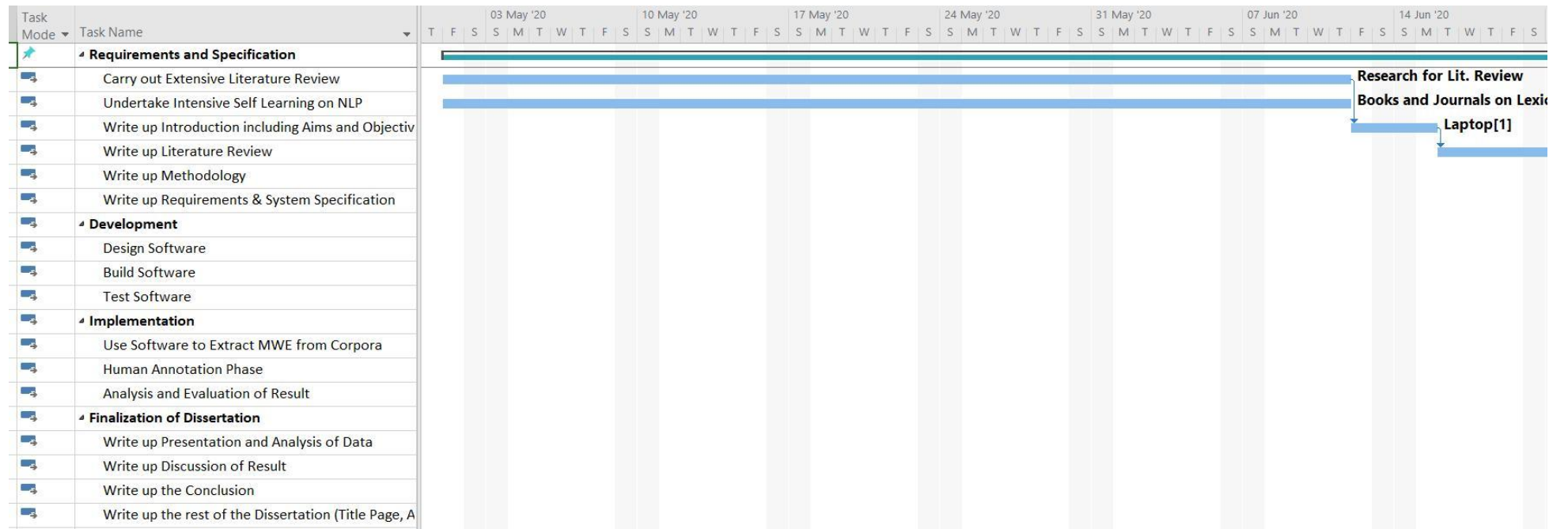


Figure 5: Gantt Chart for the Proposed Project, View 1

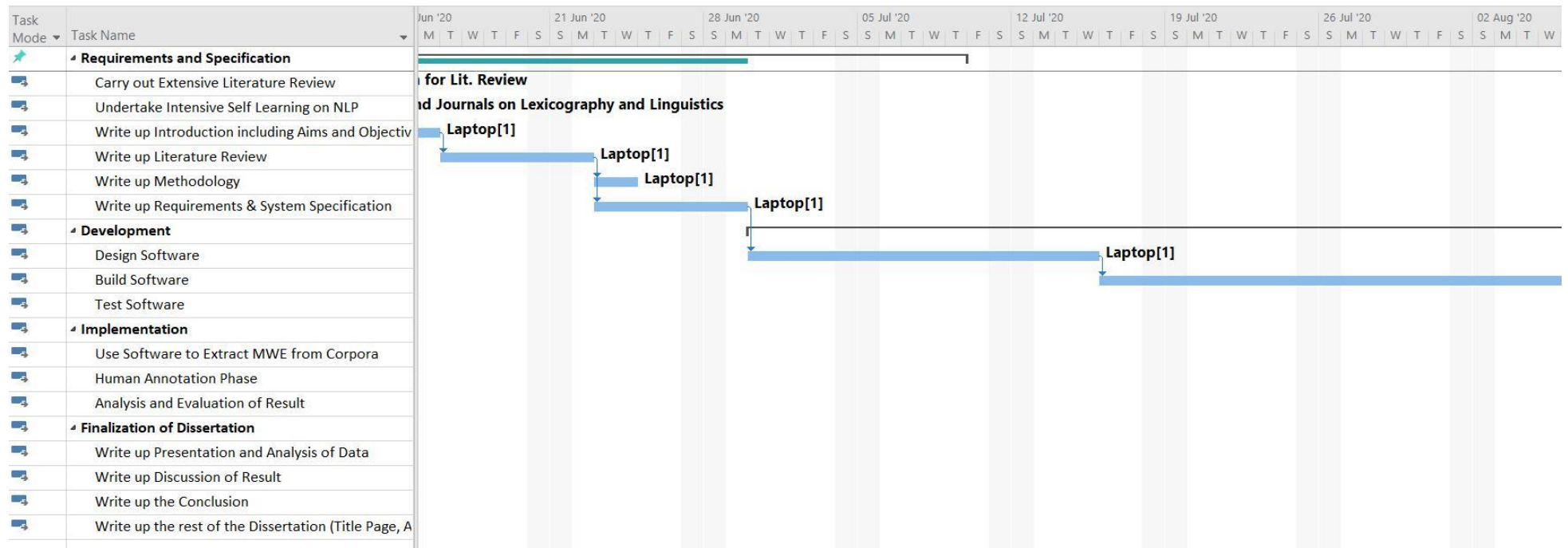


Figure 6: Gantt Chart for the Proposed Project, View 2

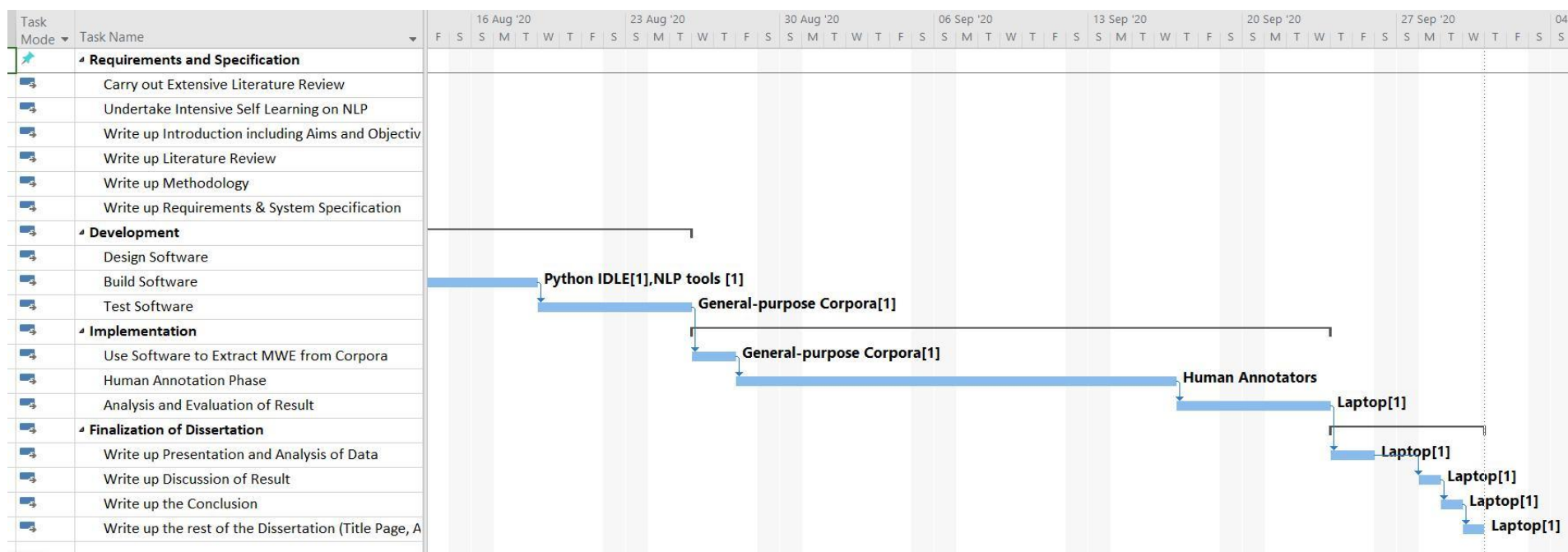


Figure 7: Gantt Chart for the Proposed Project, View 3

	i	Resource Name	Type	Material	Initials	Group	Max.	Std. Rate	Ovt. Rate	Cost/Use	Accrue	Base
1		Human Annotators	Work		H		100%	£0.00/hr	£0.00/hr	£0.00	Prorated	Standard
2		Laptop	Material		L			£0.00		£0.00	Prorated	
3		Python IDLE	Material		P			£0.00		£0.00	Prorated	
4		NLP tools	Material		N			£0.00		£0.00	Prorated	
5		Lexical Databases	Material		LD			£0.00		£0.00	Prorated	
6		General-purpose Corpora	Material		G			£0.00		£0.00	Prorated	
7		Research for Lit. Review	Cost		R						Prorated	
8		Books and Journals on Lexicography and Linguistics	Cost		B						Prorated	

Figure 8: Resource Sheet of Proposed Project

References

- Oxford University Press, n.d. *British National Corpus (BNC)*. [Online]
Available at: <https://www.english-corpora.org/bnc/>
[Accessed 25 March 2020].
- Agile Business Consortium, 2019. *Chapter 10: MoSCoW Prioritisation*. [Online]
Available at: https://www.agilebusiness.org/page/ProjectFramework_10_MoSCoWPrioritisation
[Accessed 25 March 2020].
- Association for Computational Linguistics (ACL), 2020. *SIG Compliance*. [Online]
Available at:
[https://www.aclweb.org/adminwiki/index.php?title=SIG_Compliance#SIGLEX: SIG on the lexicon](https://www.aclweb.org/adminwiki/index.php?title=SIG_Compliance#SIGLEX:SIG_on_the_lexicon)
[Accessed 29 February 2020].
- Atkins, B. T. S. & Rundell, M., 2008. *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.
- Baldwin, T. & Kim, S. N., 2010. Multiword Expressions. In: N. Indurkha & F. J. Damerau, eds. *Handbook of Natural Language Processing*. Second Edition ed. London: CRC Press, p. 267–292.
- Biber, D. et al., 1999. *Longman Grammar of Spoken and Written English*. 1st Edition ed. Harlow: Pearson Education.
- Dickson, B., 2020. What is Natural Language Processing (NLP)? *PC Magazine*, Feb. pp. 122-127.
- Eisenstein, J., 2019. *Introduction to Natural Language Processing*. Cambridge M.A.: The MIT Press.
- Font Jr., V. M., 2012. *The Ultimate Guide to the SDLC*. North Carolina: A FontLife Publication.
- Gantar, P., Colman, L., Escartin, C. P. & Alonso, H. M., 2019. Multiword Expressions: Between Lexicography and NLP. *International Journal of Lexicography*, 32(2), pp. 138 - 162.
- Gelbukh, A. & Kolesnikova, O., 2013. Multiword Expressions in NLP: General Survey and a Special Case of Verb-Noun Constructions. In: S. Bandyopadhyay, S. K. Naskar & A. Ekbal, eds. *Emerging Applications of Natural Language Processing: Concepts and New Research*. Hershey: IGI Global, pp. 1 - 21.
- Jurafsky, D. & Martin, J. H., 2008. *SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd ed. ed. New Jersey: Pearson Prentice Hall.
- Kordoni, V., Ramisch, C. & Villavicencio, A. eds., 2011. *Proceedings of the ACL workshop on multiword expressions: from parsing and generation to the real world (MWE 2011)*. Portland, Association for Computational Linguistics, p. 144.
- Kordoni, V., Ramisch, C. & Villavicencio, A. eds., 2013. *Proceedings of the 9th workshop on multiword expressions (MWE 2013)*. Atlanta, Association for Computational Linguistics, p. 144.
- Kordoni, V. et al. eds., 2014. *Proceedings of the 10th workshop on multiword expressions (MWE 2014)*. Gothenburg, Association for Computational Linguistics, p. 133.

- Kothari, C. R., 2004. *Research methodology: methods and techniques*.. New Delhi: New Age International.
- Manning, C. & Schütze, H., 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.
- Ramisch, C., 2015. *Multiword Expressions Acquisition: A Generic and Open Framework*. New York: Springer.
- Roebuck, K., 2011. *Systems Development Life Cycle (Sdlc) : High-Impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors*. Samford: Tebbo.
- Saunders, M., Lewis, P. & Thornhill, A., 2007. *Research Methods for Business Students*. 6th ed. ed. London: Pearson.
- SIGLEX, 2016. *SIGLEX: ACL Special Interest Group*. [Online]
Available at: <http://alt.qcri.org/siglex/>
[Accessed 29 February 2020].
- SIGLEX-MWE, 2020. *Joint Workshop on Multiword Expressions and Electronic Lexicons (MWE-LEX 2020)*. [Online]
Available at: http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_02_MWE-LEX_2020_lb_COLING_rb
[Accessed 29 February 2020].
- SIGLEX-MWE, n.d. *SIGLEX-MWE*. [Online]
Available at: <http://multiword.sourceforge.net/PHITE.php?sitesig=MWE>
[Accessed 29 February 2020].
- SIGLEX-MWE, n.d. *SIGLEX-MWE Workshops*. [Online]
Available at: <http://multiword.sourceforge.net/PHITE.php?sitesig=CONF>
[Accessed 29 February 2020].
- Silverman, D., 2013. *Doing Qualitative Research: A practical handbook*.. London: Sage.
- Snieder, R. & Lerner, K., 2009. *The Art of Being a Scientist: A Guide for Graduate Students and their Mentors*.. Cambridge: Cambridge University Press.
- TutorialsPoint, n.d. *SDLC - V-Model*. [Online]
Available at: https://www.tutorialspoint.com/sdlc/sdlc_v_model.htm
[Accessed 25 March 2020].
- University of Derby, n.d. *Research Onion Diagram*. [Online]
Available at: <https://onion.derby.ac.uk/onion.pdf>
[Accessed 20 March 2020].
- University of Derby, n.d. *Research Onion Diagram (based upon Saunders et al's diagram, 2009)*. [Online]
Available at: <https://onion.derby.ac.uk/>
[Accessed 15 March 2020].
- Wysocki, R. K., 2010. *Effective Software Project Management*. Indianapolis: John Wiley & Sons.