
AUTOMATIC GENERATION OF FACTUAL QUESTIONS FROM VIDEO DOCUMENTARIES

YVONNE SKALBAN

A thesis submitted in partial fulfilment of the requirements of the University of
Wolverhampton for the degree of Doctor of Philosophy

October 2013

This work or any part thereof has not previously been presented in any form to the University or to any other body whether for the purposes of assessment, publication or for any other purpose (unless otherwise indicated). Save for any express acknowledgments, references and/or bibliographies cited in the work, I confirm that the intellectual content of the work is the result of my own efforts and of no other person.

The right of Yvonne Skalban to be identified as author of this work is asserted in accordance with ss.77 and 78 of the Copyright, Designs and Patents Act 1988. At this date copyright is owned by the author.

Signature

Date

This thesis is dedicated to the memory of my grandfather, Siegfried Krix.

“Never give up, never give in”

*Sensei Cyril Cummins, 8th Dan
Birmingham & Halesowen Shotokan Karate Clubs*

ABSTRACT

Questioning sessions are an essential part of teachers' daily instructional activities. Questions are used to assess students' knowledge and comprehension and to promote learning. The manual creation of such learning material is a laborious and time-consuming task. Research in Natural Language Processing (NLP) has shown that Question Generation (QG) systems can be used to efficiently create high-quality learning materials to support teachers in their work and students in their learning process. A number of successful QG applications for education and training have been developed, but these focus mainly on supporting reading materials. However, digital technology is always evolving; there is an ever-growing amount of multimedia content available, and more and more delivery methods for audio-visual content are emerging and easily accessible. At the same time, research provides empirical evidence that multimedia use in the classroom has beneficial effects on student learning. Thus, there is a need to investigate whether QG systems can be used to assist teachers in creating assessment materials from these different types of media that are being employed in classrooms.

This thesis serves to explore how NLP tools and techniques can be harnessed to generate questions from non-traditional learning materials, in particular videos. A QG framework which allows the generation of factual questions from video documentaries has been developed and a number of evaluations to analyse the quality of the produced questions have been performed.

The developed framework uses several readily available NLP tools to generate questions from the subtitles accompanying a video documentary. The reason for choosing video

documentaries is two-fold: firstly, they are frequently used by teachers and secondly, their factual nature lends itself well to question generation, as will be explained within the thesis. The questions generated by the framework can be used as a quick way of testing students' comprehension of what they have learned from the documentary. As part of this research project, the characteristics of documentary videos and their subtitles were analysed and the methodology has been adapted to be able to exploit these characteristics.

An evaluation of the system output by domain experts showed promising results but also revealed that generating even shallow questions is a task which is far from trivial. To this end, the evaluation and subsequent error analysis contribute to the literature by highlighting the challenges QG from documentary videos can face.

In a user study, it was investigated whether questions generated automatically by the system developed as part of this thesis and a state-of-the-art system can successfully be used to assist multimedia-based learning. Using a novel evaluation methodology, the feasibility of using a QG system's output as 'pre-questions' with different types of pre-questions (text-based and with images) used was examined. The psychometric parameters of the automatically generated questions by the two systems and of those generated manually were compared. The results indicate that the presence of pre-questions (preferably with images) improves the performance of test-takers and they highlight that the psychometric parameters of the questions generated by the system are comparable if not better than those of the state-of-the-art system.

In another experiment, the productivity of questions in terms of time taken to generate questions manually vs. time taken to post-edit system-generated questions was analysed. A

post-editing tool which allows for the tracking of several statistics such as edit distance measures, editing time, etc, was used. The quality of questions before and after post-editing was also analysed. Not only did the experiments provide quantitative data about automatically and manually generated questions, but qualitative data in the form of user feedback, which provides an insight into how users perceived the quality of questions, was also gathered.

ACKNOWLEDGEMENTS

Completing this thesis was far more difficult than I could have ever anticipated. I would not have been able to finish it without the support of a large number of people, who helped and supported me in many ways. I wish I could thank each and every single person by name, but then this section would be a thesis in its own right. Apologies to anyone I did not mention, your help did not go unnoticed.

First of all, I would like to thank my Director of Studies, Professor Ruslan Mitkov, for introducing me to the area of NLP, giving me the opportunity to complete a PhD degree under his tutelage and for his continuous support throughout my studies.

I am grateful to have had a wonderful supervisory team, consisting of Dr. Lucia Specia and Dr. Le An Ha. Their expertise, continued guidance and advice were invaluable. I am especially indebted to Dr Le An Ha, who played a crucial part in me actually submitting this thesis, as he never stopped believing in me when I almost gave up. His endless support gave me the confidence to believe in myself and my research, which is probably the most significant lesson I am taking away from this experience. Cãm on, An!

I would like to thank my former fiancé Richard Payne, not only for the help with this thesis, including the countless hours of programming and teaching me statistics, but also for the love and support he has given to me over the years. Most importantly, I would like to thank Richard for instilling many important values in me, such as drive, dedication and attention to detail, which have helped me get to where I am now.

I could not have completed this thesis without the love and support from my family; my parents Rotraud and Edward Skalban, my sisters Lisa and Kristin and my grandparents Irene Krix and Rolf Oblotzki. It has been a difficult time for me and you have done so many things to let me know that I am not alone and you never stopped believing in me. Thanks for all the survival parcels and kind gestures. I hope I have done you proud!

I am grateful to Austin Birks for the countless ways in which he has given me help and support, but most importantly for his friendship and many light-hearted hours together which have made the stressful end phase of this PhD a lot more bearable.

I would also like to thank Anna-Maria Clark, Christopher Dabrowa and Sophie Stoll for being amazing, understanding friends. A big thank you to all my friends at Birmingham & Halesowen Shotokan Karate Clubs and Sensei Cyril Cummins. Karate has kept me sane during the write up phase! I would like to express my gratitude to colleagues from the Research Group in Computational Linguistics, many of whom have become my friends especially Dr Miranda Chong, Alison Carminke, Natalia Ponomareva, Dr Iustina Ilisei, Dr Irina Temnikova. I would like to thank Catalina Hallett for her help with programming, Wilker Aziz for his excellent support when I used his post-editing tool PET and Emma Franklin for proofreading my thesis.

TABLE OF CONTENTS

CHAPTER 1: Introduction.....	1
1.1 A brief introduction to Computational Linguistics and Natural Language Processing.....	1
1.2 A brief introduction to Question Generation.....	4
1.3 Aims and contributions	5
1.4 Structure of this thesis	10
CHAPTER 2: Questions and Question Generation.....	13
2.1 The value of questions in education	13
2.1.1 Higher versus lower order questions	14
2.2 Question taxonomies	18
2.3 Question Generation in NLP.....	20
2.4 Summary	33
CHAPTER 3: Documentary Videos and Subtitles in Question Generation.....	35
3.1 Multimedia learning: videos in the classroom	35
3.2 Why use documentary videos?.....	40
3.3 Documentary genres.....	42
3.4 Using subtitles in Question Generation.....	43
3.4.1 Accessibility of subtitles	43
3.4.2 Format and structure of subtitles.....	44
3.4.3 Challenges posed by using subtitles.....	46
3.4.4 Comparing subtitles to other text types.....	48
3.5 Summary	56
CHAPTER 4: A framework for Question Generation from Documentary Videos.....	59
4.1 Overview	60
4.2 The GATE architecture for Natural Language Processing.....	63
4.3 Setting up subtitles for use with GATE.....	67

4.4	Linguistic pre-processing	67
4.4.1	Tokenisation	67
4.4.2	Sentence splitting	68
4.4.3	Part-Of-Speech Tagging	69
4.4.4	Syntactic Parsing	69
4.4.5	Named Entity Recognition and Gazetteer Lookup	70
4.4.6	Morphological analysis	71
4.4.7	Pronoun resolution and sentence simplification	72
4.5	Rule-based approach to question generation	75
4.5.1	Question rules and helper rules	76
4.6	Extracting images to accompany questions	85
4.7	Error Analysis 1: Human expert opinion	87
4.7.1	Methodology	87
4.7.2	Discussion	89
4.7.3	Conclusion	93
4.8	Error Analysis 2: Performance of transformational rules	96
4.8.1	Methodology	97
4.8.2	Error analysis	101
4.8.3	Conclusion	107
4.9	Summary	108
CHAPTER 5: Evaluation 1 - Prequestions and Psychometric parameters		111
5.1	Evaluation: QG as a three-step-process	111
5.1.1	Evaluation of the key concept identification task	112
5.1.2	Evaluation of the question type determination task	115
5.1.3	Evaluation of the question realisation task	115
5.1.4	QG shared evaluation tasks	118
5.2	Experiment 1	120
5.2.1	Background	121
5.2.2	Methodology	122
5.2.3	Results	128
5.2.4	Conclusion	135
5.3	Summary	136

CHAPTER 6: Evaluation 2 - Post-editing Versus manual generation	139
6.1 Experiment 2	139
6.1.1 Background and related work.....	139
6.1.2 Methodology	140
6.1.3 Results	145
6.1.4 Conclusion.....	159
6.2 Summary	161
CHAPTER 7: Conclusion and thesis review	162
7.1 Review of contributions	162
7.2 Future work	167
7.2.1 Larger experiments.....	167
7.2.2 Different presentation of questions	167
7.2.3 Images as distractor	168
7.2.4 Use of other NLP resources.....	168
Bibliography	169
Appendix A: Previously Published Work.....	178
Appendix B: Coh-Metrix Output for subtitles and Wikipedia articles for two genres	179
Appendix C: Questions given to the human expert evaluators	187
Appendix D: Evaluators' Scoring Sheets for experiment 1	210
Appendix E: Sample Questions and screenshots from Experiment 1.....	216
Appendix F: PET output (extract)	219

LIST OF TABLES

Table 1 Questions generated in Chen et al. (2009).....	31
Table 2 Coh Metrix indices for subtitles and Wikipedia articles.....	51
Table 3 Rules employed by the system - prefix QR denotes a question rule, HR denotes a helper rule	84
Table 4 Extracted screenshots for questions.....	86
Table 5 Examples of questions assessed by evaluators	89
Table 6 Error type classes and numbers of times errors were observed in the dataset.....	91
Table 7 Usable and unusable questions generated by the two systems broken down by question types (all question types	99
Table 8 Usable and unusable questions generated by the two systems broken down by question types (shared question types)	100
Table 9 Observed error types in the output of 2 Question Generation frameworks	102
Table 10 Questions with similar content generated by all three QG methods.....	125
Table 11 Pre-questions with screenshots extracted from the video	126
Table 12 Pre-question scenarios. All scenarios contained 9 identical post-questions, 3 generated by each method (WLV, CMU and Manual).....	127
Table 13 Breakdown of correct and incorrect answers per pre-question type	129
Table 14 Seconds taken to answer post-questions depending on pre-question type.....	130
Table 15 Discriminating powers for all three QG methods.....	131
Table 16 Qualitative feedback from test-takers	134
Table 17 Numbers of question post-edited and manually created.....	145
Table 18 Comparison between post-editing times and manual creation	146
Table 19 Ratings for ‘usability before post-editing’ assigned per documentary and system	148
Table 20 Ratings for ‘usability after post-editing’ assigned per documentary and system	150
Table 21 Average scores assigned by post-editors before and after post-editing, per system	153
Table 22 Difference in question scores before and after post-editing	154
Table 23 Sums and averages of edit operations and HTER values per post-editor.....	155

LIST OF FIGURES

Figure 1 Bloom’s Taxonomy.....	14
Figure 2 Bloom’s revised Taxonomy	15
Figure 3 Example of automatically generated questions as proposed by Sano et al. 2008.....	26
Figure 4 Screenshot of srt subtitle file.....	45
Figure 5 Conceptual design of the framework	61
Figure 6 Screenshot of post-editing environment	62
Figure 7 Screenshot of the GATE Developer GUI.....	66
Figure 8 Post-editing questions in PET Screenshot	143
Figure 9 Assessing post-editing effort and question quality in PET	144
Figure 10 Assessing post-editing effort and question quality in PET	144
Figure 11 Usability ratings before and after post-editing assigned by post-editor 1.....	151
Figure 12 Usability ratings before and after post-editing assigned by post-editor 2.....	151

LIST OF ABBREVIATIONS

ADJP	Adjective Phrase
ADVP	Adverb Phrase
API	Application Programming Interface
CL	Computational Linguistics
CMU	Carnegie Mellon University
DP	Discrimination Power
DT	Determiner
GATE	General Architecture for Text Engineering
GUI	Graphical User Interface
HOTS	Higher Order Thinking Skills
HTER	Human-targeted Translation Error Rate
IDE	Integrated Development Environment
IDF	Inverse Document Frequency
IE	Information Extraction
IR	Information Retrieval
JAPE	Java Annotations Pattern Engine
LHS	Left Hand Side
LOTS	Lower Order Thinking Skills
LSA	Latent Semantic Analysis
MCQ	Multiple Choice Question
MT	Machine Translation
NE	Named Entity

NER	Named Entity Recognition
NLG	Natural Language Generation
NLP	Natural Language Processing
NP	Noun phrase
POS	Part of Speech
PP	Prepositional phrase
QA	Question answering
QG	Question Generation
QG STEC	Question Generation Shared Task and Evaluation Challenge
RHS	Right Hand Side
WSD	Word Sense Disambiguation

CHAPTER 1: INTRODUCTION

In this Chapter, the research area of Natural Language Processing is introduced in Section 1.1 and its sub-discipline, Question Generation, in Section 1.2. In Section 1.3, the aims and original contributions of the research undertaken as part of this thesis are outlined.

1.1 A brief introduction to Computational Linguistics and Natural Language Processing

The terms Computational Linguistics (CL) and Natural Language Processing (NLP) describe research areas exploiting the benefits of computer science, linguistics, statistics and other fields to bridge the language gap between machines and humans; these areas utilise the processing power of computers and the linguistic expertise of humans to analyse, understand and generate natural language. Rather than replacing humans entirely, however, machines assist by performing laborious and time-consuming tasks. Computational Linguistics is normally viewed as the application of linguistic theories and computational techniques to problems of natural language processing (Hinrichs, 2005). One formal account of CL is given by the Association for Computational Linguistics (2005):

“Computational linguistics is the scientific study of language from a computational perspective. Computational linguists are interested in providing computational models of various kinds of linguistic phenomena. These models may be “knowledge-based” (“hand-crafted”) or “data-driven” (“statistical” or “empirical”).”

Researchers in these areas have a variety of techniques at their disposal from which they choose depending on the type of language analysis to be performed. For example, in some contexts a statistical approach might be more appropriate than a rule-based approach and vice versa. There are no restrictions with regards to language, mode or genre or whether a text is spoken or written, but typically, NLP is used to analyse text which is ‘naturally occurring’ rather than text which is artificially created for the purpose of the analysis (Liddy, 2003).

Human communication involves different linguistic levels. For example, in spoken exchanges, the phonetic level is concerned with physical properties of speech sounds, while syntactic processes make sure that sentences adhere to certain sentence formation rules governing a language. On the pragmatic level, the meaning of an utterance is regarded in context. For humans, these processes are natural and often take place subconsciously. While humans are thought to utilise all of these levels since each level conveys different types of meaning, NLP systems can make use of specific levels only or a combination of levels (ibid.). NLP can be regarded as a field related to Artificial Intelligence, as it aims to provide ‘human-like language processing’.

Research in CL and NLP started as early as the 1950’s. Back then, the main research focus was on machine translation, triggered by the famous memorandum by Warren Weaver (1949). The memorandum stipulated goals and methods for machine translation that broke away from simple word-by-word approaches. Weaver was a widely recognised expert in statistics and computing, but also had a large influence on major policy-makers in U.S. government agencies; for this reason, his publication essentially paved the way for machine translation research in the United States.

Major progress in CL and NLP has been achieved mostly in recent decades; this has been due to the advances in technology and widespread use of computers as well as the better understanding of the mechanisms of human language from several linguistic viewpoints and the availability of data and statistical methods to process it. The research areas are very broad and include a variety of sub-disciplines. NLP research has produced many successful practical applications, for example, machine translation systems which automatically translate from one language to another and question answering systems which search collections of textual data for the correct answer to a user's question in natural language. Many tasks in NLP can actually be achieved with near-human accuracy; Part-of-Speech (POS) taggers, for example, reach an accuracy of ~97%, while syntactic parsers reach an accuracy of up to 92%¹ for English. While some areas of NLP have been enjoying the attention of researchers for several decades, such as Information Extraction (IE), Information Retrieval (IR), Automatic Summarisation and Speech Recognition, new areas have been emerging, too. One of these new areas is Question Generation, which is the main research area this thesis is concerned with. CL and NLP are still very challenging and promising fields with a significant commercial interest for efficient and accurate resources to process human language. NLP tools and techniques have proven beneficial in the medical and bio-medical domain, for example, to process medical notes (Patrick, Wang and Budd, 2006) and to identify biological entities, such as gene names, in texts. In educational settings, NLP has been extremely popular, too. Research (Mitkov, Ha and Karamanis, 2006) has shown that systems for Question Generation can assist educators in the laborious task of creating assessment materials, while systems for automatic essay

¹ [http://aclweb.org/aclwiki/index.php?title=Parsing_\(State_of_the_art\)](http://aclweb.org/aclwiki/index.php?title=Parsing_(State_of_the_art))

scoring (Burstein and Attali, 2006) and systems for plagiarism detection (Chong and Specia, 2012) are supporting educators in their daily routine.

1.2 A brief introduction to Question Generation

Question generation, a sub-discipline of NLP and the focus of the research presented here, is concerned with the automatic generation of questions often from texts and other input sources. Automatically generated questions are useful in many contexts; questions can, for example, be generated from information repositories to serve as candidates for Frequently Asked Questions (FAQs). They may be used in medical settings by patients and doctors or in legal settings by solicitors (Rus, Graesser and Cai, 2008). Questions can take a variety of surface realisations, such as multiple choice questions (MCQs), cloze items (aka fill-in-the-blank questions) and concept completion questions (aka wh-questions, who, what, where, when) just to name a few. In educational contexts, MCQs are popular because they provide a form of quick formative assessment to teachers and students and can help to save time and resources (Mitkov, Ha and Karamanis, 2006). Automatically generated questions can also be used to promote and assess deeper learning by providing questions that human or computer tutors might ask and by suggesting questions that learners might ask themselves in their learning process, for example whilst reading (Rus, Graesser and Cai, 2008). Rus (n.d.) defines Question Generation as:

“[...] the task of automatically generating questions from various inputs such as raw text, database, or semantic representation. Question Generation is regarded as a discourse task involving the following four steps: (1) when to ask the

question, (2) what the question is about, i.e. content selection, (3) question type identification, and (4) question construction. Question Generation is an important component in dialogue systems, virtual environments, and learning technologies such as Intelligent Tutoring Systems, inquiry-based environments, and instructional games.”

As part of this thesis, a framework was developed which automatically generates factual questions from documentary videos. While videos are often used to deliver learning content in educational and training settings, the manual creation of materials to assess and support learners' comprehension of the subject matter depicted is not only very time-consuming, but also cost and labour-intensive, as in order to create high-quality assessment and support materials domain experts are required. The questions generated with this approach can, for example, be used by teachers to test students' comprehension of a video shown in class. The framework makes use of existing NLP resources and a rule-based approach to form questions from the subtitles accompanying a documentary; the exact methodology is described in Chapter 4 and an evaluation of the proposed approach is presented in Chapters 5 and 6.

1.3 Aims and contributions

Questions are an integral part of teachers' daily instructional activities; teachers spend between 35% and 50% of their instructional time conducting questioning sessions (Cotton, 2001). Questions are used to assess students' knowledge and comprehension and to promote learning. The manual creation of such learning material is a time-consuming task.

Research in Natural Language Processing (NLP) has shown that Question Generation (QG) systems can be used to efficiently create high-quality learning materials to support teachers in their work and students in their learning process (Mitkov, Ha and Karamanis, 2006).

A number of successful QG applications for education and training have been developed, but these focus mainly on supporting reading materials. However, digital technology is always evolving; there is an ever-growing amount of multimedia content available, and more and more delivery methods for audio-visual content are emerging and are easily accessible. At the same time, research provides empirical evidence that multimedia use in the classroom has beneficial effects on student learning. Thus, there is a need to investigate whether QG systems can be used to assist teachers in creating assessment materials from these different types of media that are being employed in classrooms.

The main aim of this thesis is to investigate how NLP tools and techniques can be harnessed to generate questions from multimedia learning materials, in particular videos, to be used in educational contexts and to support educators in the laborious and time-consuming task of generating assessment materials. As part of the research, several research questions will be answered:

1. How can Natural Language Processing tools and techniques be used when automatically generating questions from multimedia learning materials?
2. What are the characteristics of video documentaries and their subtitles and how do they affect the Question Generation process?
3. How can the effects of system-generated questions be evaluated in educational settings?

4. How do system-generated questions differ from those created by human experts?

In order to answer these research questions, this thesis is organised into two main parts.

Part 1 consists of Chapters 1 to 3 which present the background information for the research. These Chapters are used to describe the motivation for the research, explain common terminology, and provide a comprehensive review of existing approaches in Question Generation. **Part 2** consists of Chapters 4 to 6, which describe the proposed framework and experiments performed.

Chapter 2 partly answers research question 1, by providing a comprehensive review of existing approaches in Question Generation (this question is also partly answered in Chapter 4). Question Generation is still a very young research area, but it has proven to be beneficial in educational settings. The value of questions in educational settings and a number of issues related to this, for example, the incidence of question use in the classroom, how questions are processed by the brain and whether different types of questions ('higher versus lower order') affect the learning process in different ways, are examined.

An answer to research question 2 is provided in Chapter 3, by describing different genres of documentaries and explaining the benefits of using videos for teaching. In addition, the characteristics of documentary subtitles are discussed, their advantages and challenges for Question Generation highlighted and a qualitative analysis comparing subtitles to another text type is performed.

In order to meet the main aim, which is to investigate how NLP tools and techniques can be harnessed to generate questions from multimedia learning materials, in particular videos, a framework is proposed which uses NLP tools and techniques to generate factual questions from documentary videos. To the best of my knowledge, no such framework has been proposed yet. In **Chapter 4**, two error analyses are described which helped to identify error types in the automatically questions. Based on these error analyses, several improvements to the framework and its transformational rules were made. Research questions 3 and 4 are answered by the experiments described in **Chapters 5 and 6**. The framework has undergone several cycles of developments, evaluations and improvements.

The research undertaken as part of this project produced several original contributions: The **first original contribution** is a framework for Question Generation from video documentaries, which makes use of several existing NLP tools and techniques. The framework is described in detail in Chapter 4. The questions generated using this methodology can be used as a quick way of testing students' comprehension of what they have learned from the documentary. The reason for choosing video documentaries is two-fold; firstly, documentary videos are frequently used by teachers and secondly, their factual nature lends itself well to question generation, as will be explained in Chapter 3.

The framework uses several readily available NLP tools to generate questions from the subtitles accompanying a video documentary. Although several text-based QG systems have already been developed, these differ from the approach described in this thesis in that the type of text they process is, by nature, different from documentary subtitles. There are different types of documentaries and certain features of the documentary subtitles affect

the Question Generation process. Not all genres of documentary videos are suitable for factual Question Generation and some will yield a larger number of useful questions than others. Thus, the **second original contribution** is the analysis of the characteristics of documentary videos and their subtitles and the adaptation of the methodology to be able to exploit these characteristics. This analysis can be found in Chapter 3.

In a user study described in Chapter 5, a novel evaluation methodology is proposed, the **third original contribution**. The evaluation approach employed is a double-blind, randomised, controlled crossover study to investigate whether questions generated automatically by the system developed as part of this thesis (system WLW) and a state-of-the-art system (system CMU) can successfully be used to assist multimedia-based learning. The feasibility of using a QG system's output as pre-questions was examined; with different types of pre-questions used: text-based and with images. The psychometric parameters of the automatically generated questions by the two systems and of those generated manually were compared. Specifically, the effect such pre-questions have on test-takers' performances on a comprehension test about a scientific video documentary was investigated. The discrimination power of the questions generated automatically against that of questions generated manually was also compared. The results indicate that the presence of pre-questions (preferably with images) improves the performance of test-takers. They indicate that the psychometric parameters of the questions generated by system WLW are comparable to, if not better than those of the state-of-the-art system. In addition, the ability to extract images from the video is a feature that is unique to system.

Not only did the user study provide quantitative data about automatically and manually generated questions, but qualitative data in the form of user feedback, which provides an insight into how users perceived the quality of questions, was also gathered. The evaluation method employed is a novel and unique approach to investigate a large number of research questions in one experiment, whilst at the same time eliminating variables that could influence the results, such as cross-group-performance and cross-question-performance.

In another experiment, the **fourth original contribution**, the productivity of questions in terms of time taken to generate questions manually vs. time taken to post-edit system-generated questions was analysed. A post-editing tool which allows tracking several statistics such as post-editing time, and edit distance between the original and the post-edited question and others was used. The quality of questions before and after post-editing was also analysed. The experiments provide a unique insight into the nature of automatically generated questions by combining quantitative analyses with qualitative feedback from users and human expert evaluators.

1.4 Structure of this thesis

The remainder of this thesis is organised as follows: Chapter 2 describes the area of Question Generation and work related to the research presented. Chapter 3 discusses issues surrounding the use of documentary videos and subtitles for question generation. Chapter 4 serves to give a detailed account of the methodology of the QG system and describes two error analyses. In Chapter 5, evaluation in Question Generation is discussed. In Chapters 5

and 6, the results of two experiments to evaluate the QG framework are presented. Finally, Chapter 7 reviews this thesis, its aims and contributions and discusses future work.

CHAPTER 2: QUESTIONS AND QUESTION GENERATION

This Chapter serves to describe Question Generation in detail. In Section 2.1, the value and importance of questions in education are discussed. In recent decades, researchers have exhibited diverging opinions with regards to the learning effect of more complex questions ('higher order') versus simpler recall questions ('lower order'). A review of different research findings can be found in Section 2.1.1. In Section 2.2, different taxonomies and types of questions are described. Finally, in Section 2.3, related literature is reviewed and it is explored how questions can be generated automatically using NLP tools and techniques.

2.1 The value of questions in education

Questions are an integral part of teachers' instructional activities; questioning sessions make up between 35-50% of instructional time (Cotton, 2001). In the classroom, questions are used to assess students' knowledge and comprehension and to promote learning. Questions are a widely researched topic in education. Several papers have been dedicated to examining the incidence and types of questions that teachers ask (e.g. Guszak, 1967; Gall, 1970). Wright & Nuthall (1970) investigated the effects of different types of questions, while Gall and Rhody (1978) compared the effect of teacher questions with other instructional methods. Other research has been concerned with training teachers to use certain types of questions (e.g. Galassi, Gall, Dunning, & Banks, 1974); teaching

students how to answer questions (e.g. Raphael & Wonnacott, 1985) and teaching students to generate their own questions (e.g. Commeyras & Sumner, 1998).

2.1.1 Higher versus lower order questions

In the 1950s, a group of educational psychologists led by Benjamin Bloom, developed the ‘Taxonomy of Educational Objectives’ (Bloom, 1956), which is still of importance today as it has become a key tool in structuring and understanding the learning process. According to the taxonomy, there are three psychological domains of learning (ibid.):

- the **Cognitive** domain – processing information, knowledge and mental skills
- the **Affective** domain – attitudes and feelings
- the **Psychomotor** domain – manipulative, manual or physical skills

Bloom is best known for his work in the cognitive domain, which involves knowledge and the development of intellectual skills. It includes the recall or recognition of specific facts, procedural patterns, and concepts that serve to develop intellectual abilities and skills. There are six major categories of skills, each described using a gerund (see Figure 1). The categories are arranged in terms of degree of difficulty and complexity, with the Lower Order Thinking Skills (LOTS) at the bottom and the Higher Order Thinking Skills (HOTS) towards the top.

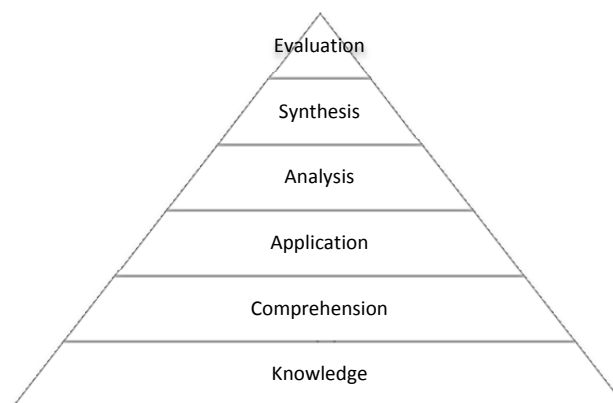


Figure 1 Bloom's Taxonomy

The main idea is that the lower ones must normally be mastered before the next ones can take place, meaning that one cannot apply new knowledge unless it is comprehend first, etc. This hierarchy can also be applied to the teacher questions. Teachers can ask lower order questions which recall facts and foster comprehension, or higher order questions which will foster critical thinking and prompt students to apply knowledge they have gained.

In 2001, a revised version of Bloom's taxonomy was published by a former student, Lorin Anderson with David Krathwohl (2001). In the revised version, the sequence within the taxonomy is altered and nouns are used for each category (see Figure 2). Anderson and Krathwohl considered creativity to be higher within the cognitive domain than evaluation.

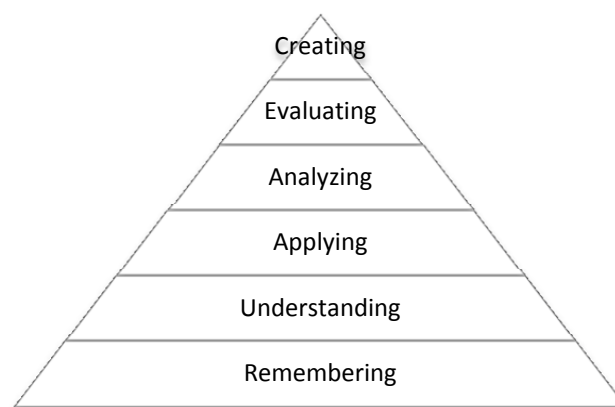


Figure 2 Bloom's revised Taxonomy

A related framework for teaching and reading has been proposed by Pearson and Johnson (1978). Their 'taxonomy of question-answer relations' takes into account whether the answer to a question can be deducted from the source text or whether additional knowledge is required to answer a question. Three different types of question-answer relations have been identified:

1. *Textually explicit*: the information required for answering the question is stated explicitly in the text (“reading the lines”)
2. *Textually implicit*: the question can be answered from the text, but the answer has to be inferred by integrating material from different parts of the text (“reading between the lines”)
3. *Scriptally implicit*: the information required for answering the question involves prior knowledge (“reading beyond the lines”).

Many educational researchers have been debating whether higher order questions promote learning and the development of thinking skills more than lower order questions, with differing results. Winne (1979) and Redfield and Rousseau (1981) performed meta-reviews of 18 (14, respectively) studies, 13 of which overlapped. While Redfield and Rousseau concluded that the use of higher order questions results in improved student achievement, Winne found no correlation between student achievement and type of question used. Other researchers who have performed reviews have found inconsistent results and some even report higher student achievement when lower order questions are used (Rosenshine, 1971; 1976; Dunkin & Biddle, 1974). Other researchers again (Samson, G. E., Strykowski, B., Weinstein and Walberg, 1987) reported a small positive effect of higher order questions on student achievement.

With regards to frequency of use of the different question types, study results are unanimous. Gall (1984) found that only about 20 percent of the questions posed in most classrooms require thinking at higher levels. Similarly, studies performed by Stevens

(1912), and Sirotnik (1983) have reported a higher incidence of lower order than higher order questions in classrooms. Research by Nystrand (1997), observing 58 eighth-grade language, arts and English classes, found that 64% of questions were lower order questions involving recitation and reporting of facts and only 36% of questions involved high-level thinking. Similar findings have been made by Guszak (1967); by analysing which types of questions 12 teachers asked in grades 2-6 in a school in Texas, he observed that 70% of questions were recognition or recall questions that focused on literal comprehension. More recent studies by Taylor et al., (2003, 2005) are still in line with this trend; in their study which examined question asking in schools in high-poverty areas, the authors observed that only 16% of teachers in grades 1 to 3 frequently asked higher-level questions.

While study results diverge with regards to whether higher order or lower order questions are more beneficial to student achievement, most researchers conclude that higher-level questions promote the development of thinking skills. However, this does not imply that lower order questions are not indeed beneficial for the learning process, or as Churches (2009) puts it:

“While the recall of knowledge is the lowest of the taxonomic levels it is crucial to learning. Remembering does not necessarily have to occur as a distinct activity. Remembering or recall is reinforced by application in higher level activities.”

How does this relate to automatic Question Generation? The system developed as part of this thesis generates factual, short-answer-style ‘wh’-questions (who, whom, whose, when, where, what) and these would typically be regarded as lower order questions. As can be seen from the literature review, lower order questions do not automatically imply lower

quality of questions or less learning effect. In fact, a user study was performed (Chapter 5), which highlighted that the proposed system can generate questions with high values for discriminating power (DP, an important psychometric measure for test questions from classical test theory). While manually created questions exhibited the highest DP, there is no statistically significant difference between system WLW and the state-of-the-art system, implying that questions generated by the system are as good as, if not better than, questions generated by the state-of-the-art system. It is important to bear in mind that the system developed as part of this thesis has not been designed to replace teachers in the Question Generation process, but to assist them, so the system can be used to support them in creating test questions, while teachers could supplement a number of higher order questions if needed.

2.2 Question taxonomies

As different applications of QG are based on different questions types, researchers in a variety of fields have proposed question classification schemes. Graesser, Person and Huber (1992, in Graesser et al., 2008) identified four different question types which occur in natural settings. *Sincere-information seeking (SIS)* questions are bona fide knowledge deficit questions which seek to bridge a specific gap in the questioner's knowledge. Other types of questions address communication and social interaction; *common ground questions* serve to establish whether knowledge is shared between participants in the conversation ("Did you mean...?", "Do you understand...?"), while *social coordination questions* are indirect requests to perform an action or to be allowed to perform an action

(“Could you...?”). Lastly, *conversation-control questions* are used to manipulate the flow of a conversation (“Can I ask you a question?”) (ibid.).

Graesser and Person (1994, in Graesser et al., 2008) suggested the following sixteen question categories:

1. *Verification*: invites a yes or no answer.
2. *Disjunctive*: Is X, Y, or Z the case?
3. *Concept completion*: Who? What? When? Where?
4. *Example*: What is an example of X?
5. *Feature specification*: What are the properties of X?
6. *Quantification*: How much? How many?
7. *Definition*: What does X mean?
8. *Comparison*: How is X similar to Y?
9. *Interpretation*: What is the significance of X?
10. *Causal antecedent*: Why/how did X occur?
11. *Causal consequence*: What next? What if?
12. *Goal orientation*: Why did an agent do X?
13. *Instrumental/procedural*: How did an agent do X?
14. *Enablement*: What enabled X to occur?
15. *Expectation*: Why didn't X occur?
16. *Judgmental*: What do you think of X?

The question types were also ranked according to complexity (ibid.); categories 1 to 4 are regarded as simple/shallow and can be answered without complex reasoning, categories 5 to 8 can be seen as intermediate, and categories 9 to 16 are considered complex/deep questions which require understanding of causal structures. This scale of depth also corresponds to Mosenthal's (1996, in Graesser et al., 2008) scale of question depth and Bloom's taxonomy of cognitive difficulty (1956).

The proposed framework generates mainly concept completion questions; however, sometimes the 'what' questions overlap with other question categories. For example, the question "What did Seth Putterman contemplate?" can be regarded as a causal consequence question, since in the documentary that this question is based on reveals that Seth Putterman made an observation during an experiment, which consequently led him to perform another experiment. The question "What should be produced at exactly the same billionth of a second if fusion was happening?" could be regarded as an interpretation question.

2.3 Question Generation in NLP

Question Generation (QG) as a sub-discipline of natural language generation (NLG) has only emerged in recent years. Question answering (QA) on the other hand, has been a focus in NLP research for a while, having become especially popular with the introduction of the question answering track in the Text Retrieval Conferences, beginning with TREC-8 in 1999 (Voorhees and Harman, 2000). Some QA systems rely on Question Generation to achieve their aim of providing an accurate answer to a question posed in natural language;

however, QG is mainly employed as an intermediate step with the goal of improving the output of QA systems. Ignatova et al. (2008), for example, have identified five major characteristics of ‘low quality human generated questions’ which influence the answer finding process on social Q&A sites such as Yahoo! Answers² and WikiAnswers³: misspellings, internet slang, ill-formed syntax, structurally inappropriate questions such as queries expressed by keywords, and ambiguity. The authors argue that improving these low quality questions using NLP techniques, will lead to an increase in the users’ chance of getting satisfactory answers to their questions (ibid.). Improved questions could also be used as input for automatic QA systems, as these rely on well formulated input (Rus, Cai and Graesser, 2007, in Ignatova, Bernhard and Gurevych, 2008).

Question Generation has now become a research area in its own right, with the Question Generation Workshops (Rus and Lester, 2009; Boyer and Piwek, 2010) aiming to promote research on issues related to Question Generation and the Shared Task Evaluation Challenge on Question Generation (Rus and Graesser, 2008) intending to increase the visibility of QG in the wider NLP community. Automated Question Generation finds application in many contexts; questions can, for example, be generated from information repositories to serve as candidates for Frequently Asked Questions (FAQs). They may be used in medical settings by patients and doctors or in legal settings by solicitors (Rus and Graesser, 2008). Question Generation is also very popular in educational environments. Amongst others, automatically generated questions can be used to promote and assess deeper learning by providing questions that human or computer tutors might ask and also

² <http://answers.yahoo.com/>

³ <http://wiki.answers.com/>

by suggesting questions that learners might ask themselves in their learning process, for example whilst reading (ibid.). QG systems which produce multiple-choice questions (MCQs) provide forms of assessment which allow for convenient and fast (self-) evaluation of student performance and can help to save time and resources (Mitkov, Ha and Karamanis, 2006).

Several systems for different learning purposes with different types of output questions have been developed. All the systems introduced are designed to generate learning content or promote learning in a certain way. First, systems which generate learning exercises are introduced. These systems make use of a variety of NLP tools, but the generated question types (e.g. MCQs, fill-in-the-blank questions) are different from the questions the framework developed as part of this thesis produces. Then a number of systems which are related to the framework is discussed as they make use of subtitles to generate learning material. These systems, however, lack the use of NLP tools and techniques. Finally, systems are introduced which, although aimed at reading comprehension, use NLP to output questions types similar to the ones the proposed framework produces.

A computer-aided system for the generation of multiple choice question items has been proposed by Mitkov and Ha (2003) and Mitkov, Ha and Karamanis, (2006). Multiple choice question items are generated from domain-specific source texts employing various NLP techniques, such as term extraction and shallow parsing. To generate items, the system first extracts important concepts from the source text. Nouns and noun phrases (NPs) with a frequency over a certain threshold are considered *key concepts*. The empirically determined threshold depends on several parameters such as the length of the text and the number of nouns it contains. In addition, key concepts are also those NPs

which have a key concept as their head and satisfy the regular expression $((AN|N)^+ | ((A|N)^*(NP)?)(A|N)^*N^4$. This regular expression has been found to capture a large range of structural patterns (Justeson & Katz, 1995). Questions are formed by transforming declarative source clauses into interrogative constructs. However, not all clauses are suitable for question generation; they have to fulfil certain eligibility conditions before they can be transformed. A clause-filtering module is used to identify eligible source clauses which (i) contain at least one key concept, (ii) are finite and (iii) are of SVO or SV structure (Mitkov, Ha and Karamanis, 2006.). Several simple rules are then used to generate a stem from a source clause; for example, the *subject-rule* transforms clauses into the stem “Which HVO” where H is a hypernym of a key concept extracted from WordNet (1998). In the final step, distractors, i.e. answer options which are semantically close but not identical to the key concept and which serve to distract the user from choosing the correct answer, are retrieved via WordNet (ibid.) and Wikipedia (2004). An evaluation has shown that the system can produce MCQs up to four times faster than a human domain expert without compromising quality.

Another popular application of QG is language learning, in particular vocabulary training. Brown et al. (2005) devised a system to automatically generate test items for vocabulary assessment. Six different types of questions are generated by the system: definition, synonym, antonym, hypernym, hyponym, and cloze questions (multiple choice fill-in-the-blank items). All questions are generated by exploiting the respective semantic relations in WordNet. For example, a definition question is generated by choosing the first definition

⁴“Where A is an ADJECTIVE, but not a determiner, N is a lexical noun (i. e. not a pronoun) and P is a preposition. In words, a candidate term is a multi-word noun phrase: it either is a string of nouns and/or adjectives, ending in a noun, or it consists of two such strings, separated by a single preposition” (Justeson & Katz, 1995)

from WordNet's synset which does not include the target word and so on. The cloze item requires the use of the target word in a specific context, either a complete sentence or a phrase. The example sentence or phrase is also retrieved from the gloss for a specific word sense in WordNet.

Hoshino and Nakagawa (2007) also designed a system for the generation of vocabulary assessment exercises. The system generates cloze items and employs online news articles as source texts which lend themselves well to such exercises because they generally exhibit correct spelling and use of grammar and their topics are suitable for classroom settings. The system assists users by helping to choose an article, highlighting grammar targets and suggesting possible choices for distractors. Questions are generated in which the distractors are *flat* (distractors may differ in inflection or meaning) or *symmetric* (distractors may differ in inflection and meaning). The input documents are pre-processed automatically; steps include term extraction, sentence splitting, tagging and lemmatising, synonym lookup, frequency annotation, inflection generation, grammar target mark up, grammar distractor generation and vocabulary distractor selection. All output is added to a pool of pre-processed articles which summarises information of all of the articles. User evaluations show that 80% of the generated items were deemed as appropriate.

Papasalouros et al. (2008) presented an approach to Question Generation which is based on domain-specific ontologies. Ontologies contain asserted knowledge in the form of definitions of terms, individuals belonging to these terms and relationships between these terms and individuals and they may also contain inferred knowledge, i.e. facts which are not explicitly defined. In the described approach, the asserted and inferred knowledge of ontologies as well as ontological axioms are exploited to generate multiple choice

questions. Several strategies for distractor selection are used. Distractors are always sentences with the same structure as the correct answer; amongst all sentences, the student has to identify the correct one. To exemplify, “Eupalinos is an Engineer” is a correct answer while “Polykrates is an Engineer” is a distractor. Papasalouros et al. (ibid.) employ class-based strategies (based on hierarchies), property-based strategies (based on roles between individuals) and terminology-based strategies to generate multiple choice items. An evaluation in terms of pedagogical quality, linguistic/syntactical correctness and number of questions produced for different domain specific ontologies was performed. Results show that while all reviewed items were satisfactory for assessment, not all questions were syntactically correct. It was found that in order to generate syntactically correct items, it is important that input ontologies adhere to certain conventions (e.g. properties’ names must be written as verbs). In addition, it was found that property-based strategies generate a large number of multiple choice items but are difficult to manipulate syntactically, while class and terminology-based strategies generate fewer questions but are easier to manipulate syntactically.

Research suggests that no framework which employs NLP techniques to generate factual questions from subtitles has been proposed to date; this is despite the fact that subtitles have been employed in several NLP tasks, such as machine translation (Xiao and Wang, 2009; Tiedemann, 2007; Flanagan, 2009; Aziz and Specia, 2012), estimation of word frequencies of spoken French (New, 2007) and automatic summarisation of videos (Agnihotri, et al, 2001).

Sano et al. (2008) discuss the automatic generation of multiple choice questions (MCQs) with accompanying picture from TV news programmes. While information is extracted

from subtitles in this approach, no NLP techniques are employed to process them. The suggested approach consists of several steps. Firstly, an image containing a distinctive subject is selected using computer vision techniques. Next, a sentence from the subtitles corresponding to the image is selected which is used as key (right answer). If a selected image depicts a person, the corresponding sentence should ideally contain the name of that person in order to generate a question about the person. If there is no subtitle text corresponding to a particular image, other Sections of the subtitles which are likely to contain a reference to the selected image will be searched. If the picture does not show a person, the word with the highest Inverse Document Frequency (IDF) from amongst the sentence's tangible nouns is determined to be the distinctive subject. Finally, 3 other sentences are chosen as distractors. An example of a question output by this system can be seen in Figure 3.



- (a) This is the earless seal appeared in Tokyo Bay.
- (b) A walrus is captured by fisherman.
- (c) Artificial rearing of sea lion has started.

Figure 3 Example of automatically generated questions as proposed by Sano et al. 2008

An evaluation showed that out of 199 generated MCQs (from 250 images) 39.2% were usable. Usable in this context means that for each image, one sentence could be extracted which relates to the image and which constitutes the correct answer, as well as three sentences which serve as distractors.

Another approach which employs subtitles is presented by Yang et al. (2009). The proposed system is aimed at Taiwanese learners of the English language. It provides

leaners with listening comprehension exercises focusing on reduced English verbal forms, as these often prove to be difficult for Taiwanese EFL learners. Similarly to Sano et al.'s system, no NLP techniques are used to process the subtitles as such. A corpus of film and TV subtitles is generated. Then a set of verbal contractions is given to a machine learning⁵ algorithm which produces a database of verbal reductions grouped into different categories. The system has a student interface which outputs media clips and five listening cloze items to a learner.

The two systems just described are distinct from the framework developed as part of this thesis; they use machine learning algorithms to gather information from the subtitles, but they do not process the subtitles with NLP tools in any other way, meaning that their approach only treats language very superficially and misses out on the deeper levels of linguistic processing possible by using NLP tools. In what follows, several systems are presented which have been developed to automatically generate questions from texts using a variety of NLP techniques, but which differ from the approach employed in this research in that they deal with different text types and are aimed at reading comprehension.

Heilman and Smith (2009) generate questions from reading material for educational practice and assessment using existing NLP tools. QG from complex sentences often leads to unnatural or senseless questions (ibid.); thus, several simplifying transformations are performed in the first stage of the approach. Amongst others, steps include removal of phrase types such as leading conjunctions, sentence-level modifying phrases, and

⁵ Machine learning is a scientific field which focusses on design and development of algorithms which allow computers to learn how to perform certain tasks based on empirical data. Often, machine learning methods are broken into two phases. Training: A model is learned from a collection of *training data*. Application: The model is used to make decisions about some new *test data*. (Hertzmann and Fleet, 2009)

appositives. Simplification steps are performed using Tregex and T-Surgeon (Levy and Andrew, 2006). Tregex is a utility which uses regular expressions to match patterns in trees and T-Surgeon is a tool which uses Tregex information as input in order to manipulate trees. Sentence-initial conjunctions, for example, are identified using the expression $ROOT < (S < CC=conj)$ and removed by deleting *conj* using T-Surgeon. In the second stage of the approach, declarative sentences are transformed into a set of possible questions. Answer phrases, which may be targets for wh-movement, are identified and converted into question phrases with the help of several transformational rules. The system can generate *who*, *what*, *where*, *when*, and *how much* questions and relies on the entity type annotations PERSON, LOCATION, DATE/TIME and MONEY respectively; these annotations are retrieved using the BBN Identifinder Text Suite (Bikel et al., 1999). Since one source sentence can give rise to a number of questions, some of which are less desirable, because they may, for example, contain errors, statistical ranking is performed using two approaches. In the first one, a discriminative reranker (Collins, 2000 in Heilman and Smith, 2009) based on a logistic regression model that defines a probability of ‘acceptability’ is employed. Questions may be unacceptable for various reasons, for example, ungrammaticality, the question has an obvious answer (i.e. the answer to the question is clearly the topic of the source text), the question is too vague (e.g. “What did Lincoln do?”), or a different question type should have been used (i.e. a question would have been acceptable if a different wh-word (when, where, who, etc.) had been used). In the first ranking approach (“Boolean”), question deficiencies are collapsed, meaning that a question containing any of the deficiencies is treated as unacceptable. In the second approach (“Aggregate”), separate conditional models of the probability of a given question

being acceptable according to each of the deficiencies are learnt and combined. The training and test data sets were created using corpora consisting of texts taken from the English Wikipedia⁶, Simple English Wikipedia⁷, and Wall Street Journal (WSJ) data in the Penn Treebank (Marcus et al., 1993 in Heilman and Smith, 2009). One test set also consisted of 100 questions evaluated by 15 native speakers. An evaluation shows that before ranking, 12.8% of all questions were rated as acceptable (this includes WSJ corpus), and after ranking, the percentage of all acceptable questions is 26.6% (this does not include WSJ) with a precision-at-10 of 43.3%.

Gates (2008) discusses the generation of reading comprehension questions from corpora of expository texts aimed at children. Like Heilman and Smith, several readily available NLP resources alongside a set of manually created transformation rules are used to generate wh-questions (who, what, where, when) which are directly derived from the text. Students are encouraged to use a look back strategy; as the answers to the questions are linked to the text, the students are required to go back to the relevant sentence in the text and click on the word which represents the key in order to answer a question correctly. In order to generate questions, the system first identifies named entities using BBN's *IdentiFinder* (Bikel, et al., 1998) and Prop-Bank semantic arguments (Palmer, et al. 2005) using *ASSERT* (Pradhan, et al. 2005) (Pradhan et al., 2004). Synset classes from WordNet⁸ are derived for each noun and the Stanford parser is used to obtain a parse tree with surface forms of words and to determine the stem or root of each word. The lexical, syntactic and semantic annotations are combined to produce annotated parse trees. T-Surgeon, alongside

⁶ <http://en.wikipedia.org/>

⁷ <http://simple.wikipedia.org/>

⁸ WordNet is a lexical database for the English language which groups words into sets of synonyms called *synsets*. It shows semantic relations between these synonym sets as well as short, general definitions (glosses)

a set of manually created transformation rules, is then used to generate wh-question and answer trees by transforming the candidate declarative sentence trees and converting to HTML format for displaying purposes.

Mostow and Chen (2009) describe an approach for an automated reading tutor for narrative text. Similarly to Gates (2008), the generated questions serve to promote a self-questioning approach with the aim of facilitating text comprehension in children. A four-step strategy, which gradually leads the child to learn how to ask good questions, is employed. In the first step, *describing*, the system points out to the child that question asking can help understanding. Next, in the *modelling* step, the child is given an example of a good question to ask themselves while reading. In the *scaffolding* step, the child is invited to generate questions from predefined question building blocks. Finally, in the *prompting* step, the system encourages the child to generate a question by themselves. Questions are asked where a character's mental state, e.g. belief, emotion, etc. is described in the text; this allows for good why-questions as often a change in a character's mental state occurs and inferential knowledge is needed in order to answer the question. To detect mental states, ten categories of modal verbs are chosen and their synsets extracted from WordNet. Questions are generated using a *mental state understanding system*; this consists of a parser that generates a semantic representation of the input text, and a mental state inference engine that expands the semantic representation of the text into a situation model of the story. Question types include what, why and how questions. A human evaluation classed 71.3% of questions as acceptable, where acceptability refers to grammatical well-formedness and 'appropriateness' with regards to the context of the story, but this is not further elaborated.

Based on this work, Chen, Aist, & Mostow, (2009) describe a method of automatically generating comprehension questions from informational texts aimed at schoolchildren. Since informational texts are inherently different to narrative texts, it is not possible to generate questions based on mental states. Instead, the authors make use of discourse markers as prompts for generating a question which can aid text comprehension by self-questioning. Three types of questions are proposed; questions about conditional context are asked after conditionality markers such as ‘if’, ‘even if’, etc. Questions about temporal information are asked after markers such as those regarding dates and times. Questions about modality are prompted by auxiliary verbs. Questions are then generated using a situation model and question templates. Examples of the question types can be seen in Table 1.

Question Type	Source sentence	Generated Question
Conditional	If humans removed all the kelp from the sea soon all the other sea life would start to suffer as well.	What would happen if humans removed all the kelp from the sea?
Temporal	Rainbows are seen after it rains and the sun is out.	When would rainbows be seen?
Modality	All goats should have covered shelters where they can escape the weather.	Why should all goats have covered shelters?

Table 1 Questions generated in Chen et al. (2009)

For the evaluation, the same criteria as in Mostow and Chen (2009) were used, i.e., a question has to be grammatically correct and it has to make sense in the context of the text. 180 system-generated questions were evaluated. Out of 88 questions about temporal information, 65.9% (58) were classed as acceptable. Out of 77 questions about modality, 87.0% (71) were classed as acceptable and out of 15 questions about conditional contexts, 86.7% (13) were classed as acceptable.

Chali and Hasan (2012) generate factual questions from topics based on the assumption that each topic is associated with a body of texts containing useful information about the topic. Questions are generated by exploiting named entity information and predicate argument structures of the sentences present in the body of texts. Similar to Heilman (2011), an overgenerate-and-rank approach is used in which all possible questions are generated for a given topic and then they are ranked in terms of topic relevance and syntactic correctness.

Despite the fact that some of the aforementioned systems make use of subtitles, they differ from the approach employed in this thesis in that they do not employ NLP techniques to process them. Those systems described that do use NLP techniques have a methodology similar to the one employed in the framework developed as part of this research, but they differ from it in that the type of text they process is, by nature, different from documentary subtitles. Just like other text types, documentary subtitles have unique features that require specific treatment; for example, very often meaning is conveyed in the visuals of the video, but not in the subtitles. This issue will be further elaborated Chapter 3. Nevertheless, because the methodology for Question Generation employed in the systems for reading comprehension is similar to the methodology developed as part of this thesis, Heilman's (2011) system, which in QG circles is widely regarded as the state-of-the-art, was used as the baseline for the experiments described in Chapters 5 and 6.

2.4 Summary

In this Chapter, the importance of questions in education was discussed. Bloom's taxonomy (1956), a framework of educational objectives developed in the 1950's, which has been revised several times and is still relevant in classrooms today, was introduced. Bloom's taxonomy distinguishes between so-called higher order and lower order thinking skills that ideally students should gain through classroom activities. Question asking is a major part of daily teaching. Teachers use different types of questions and it is possible to classify different types of questions as higher order questions, i.e. questions which require analytical thinking and lower order questions, i.e. questions which recall facts. There has been a longstanding debate whether one type is more beneficial than the other with regards to the learning effect. While study results diverge, most researchers conclude that higher order questions promote the development of thinking skills. However, this does not imply that lower order questions are not indeed beneficial, as the lower order skills are a building block required in order to gain higher order skills.

Question Generation is a relatively new area in NLP, but systems for automatic Question Generation can help teachers create questions as learning material. Several existing approaches have been described, the most notable ones being the one by Heilman (2011) and Mitkov, Ha, & Karamanis (2006). While the methodology is similar, the existing approaches focus solely on Question Generation for reading comprehension. To date, no system which generates questions from videos has been developed, despite the fact that video use is increasing in educational and training settings, as will be explained in the

following Chapter. The next Chapter will discuss documentary videos and subtitles in Question Generation.

CHAPTER 3: DOCUMENTARY VIDEOS AND SUBTITLES IN QUESTION GENERATION

In this Chapter, the use of documentary videos and their subtitles for Question Generation is discussed. Section 3.1 serves to examine the use of videos in educational settings. In Section 3.2, the reasons for choosing documentary videos for Question Generation are explained. In Section 3.3, documentary videos are defined and different genres of documentaries and their characteristics are described. In Section 3.4, a close look is taken at subtitles accompanying video documentaries and the way in which their features affect the QG process. Section 3.4.4 presents a qualitative analysis of subtitle texts compared to other types of texts using a number of different linguistic indices.

3.1 Multimedia learning: videos in the classroom

The use of videos in the classroom is not a new concept; teachers have been using videos in teaching sessions since the 1950s. Back then, 16mm film was the height of technology; however, technology has evolved hugely since then and nowadays even DVD players appear to be outdated. Today, teachers (and students) have a huge array of videos and video technologies available at their fingertips. Berk (2009) concludes that the popularity of videos in teaching has been brought about by changes in four areas:

1. there is a large variety of video formats available,
2. the ease with which the technology can facilitate theory application in the classroom,

3. the number of video techniques the instructor can use and
4. the research on multimedia learning that provides empirical support for their use as an effective teaching tool

An abundance of resources is freely available and easily accessible online; clips relating to almost any topic can be found on video streaming sites such as YouTube⁹ and Vimeo¹⁰. The former also features a dedicated Section with educational video content, YouTube EDU¹¹, which consists of a corpus of over 700.000 high quality educational videos. YouTube EDU targets three learning levels: primary and secondary school level, higher education level and lifelong learning level and also provides guides for teachers and schools to use the provided content most effectively. YouTube EDU is just one of many websites and resources focussing on delivering educational video content. Amongst the more well-known ones, which will be briefly introduced here, are iTunesU¹², TED-Ed¹³ and Coursera¹⁴.

Coursera offers free online courses including Humanities, Medicine, Biology, Social Sciences, Mathematics, Business, Computer Science. The courses consist of short video lectures delivered by lecturers from prestigious universities, such as Stanford, on different topics and assignments to be submitted, usually on a weekly basis. Originally just aimed at

⁹ www.youtube.com

¹⁰ www.vimeo.com

¹¹ <http://www.youtube.com/education>

¹² <http://www.apple.com/uk/education/itunes-u/>

¹³ <http://ed.ted.com/about>

¹⁴ <https://www.coursera.org/>

self-study, some of Coursera's courses have been approved for college credit at the time of writing.

A similar service has been provided by iTunesU (iTunes University) since 2007. Originally only aimed at college and university students, the service has been updated to also target children at pre-school level up to 12th grade. There are over 350,000 files available to download. Content includes course lectures, language lessons, lab demonstrations, sports highlights and campus tours provided by a number of accredited, international universities and colleges.

TED-Ed is an educational platform and spin-off of the popular TED.com site, a website with international talks on Technology, Entertainment and Design. The TED-Ed platform has a special feature; it allows users to take any educational video, and easily create a customised lesson around the video. In educational circles, this is part of a 'blended learning' technique called "flip teaching".

Flipped teaching or flipped classroom refers to a pedagogical model in which the typical lecture and homework elements of a course are reversed (Educause, 2012). The key element of flipped teaching is a short video lecture, which students watch at home before class. In-class time is then devoted to exercises, projects, or discussions. There is no single model for the flipped classroom— it applies to any class structure that provides pre-recorded lectures followed by in-class exercises. One common model is for students to watch multiple lectures of five to seven minutes each, which may be broken up with online quizzes or activities to test what students have learned. (ibid.) Immediate quiz feedback and the ability to rerun lecture segments may help clarify points of confusion. Flipped

teaching is an increasingly popular technique, because it allows the students to learn in their own time, with the ability to watch, rewind, and fast-forward as required and the class time can then be used to apply the knowledge, work on collaborative projects, etc. A study at an American high school has shown that flipped teaching can reduce fail rates in Maths and English by about 40 percent. Question Generation systems can help teachers in providing the assessment questions in flip teaching, as the manual creation of such content can be time-consuming.

With regards to learning outcomes, there are manifold reasons why videos can be useful in teaching (Berk, 2009). Amongst others, videos serve to

1. Grab students' attention;
2. Focus students' concentration;
3. Generate interest in class;
4. Create a sense of anticipation;
5. Energise or relax students for learning exercise;
6. Draw on students' imagination;
7. Improve attitudes toward content and learning;
8. Build a connection with other students and instructor;
9. Increase memory of content;
10. Increase understanding;
11. Foster creativity;
12. Stimulate the flow of ideas;
13. Foster deeper learning;

14. Provide an opportunity for freedom of expression;
15. Serve as a vehicle for collaboration;
16. Inspire and motivate students;
17. Make learning fun;
18. Set an appropriate mood or tone;
19. Decrease anxiety and tension on scary topics; and
20. Create memorable visual images.

A number of papers have been dedicated to analysing how the use of video affects the brain and student learning. According to Gardner (2000, in Berk, 2009), each student possesses an individual profile of unique 'core intelligences'. Three of those core intelligences, namely verbal/linguistic intelligence (i.e. learning by reading, writing, speaking, listening, debating, discussing and playing word games), visual/spatial intelligence (i.e. learning by seeing, imagining, drawing, sculpting, painting, decorating, designing graphics and architecture, coordinating colour, and creating mental pictures) and musical/rhythmic intelligence (i.e. learning by singing, humming, listening to music, composing, keeping time, performing, and recognising rhythm) can be tapped by using video.

Watching a video engages both hemispheres of the brain (Berk, 2009). Generally speaking, the left hemisphere is viewed as the logical and analytical side that processes information sequentially as in mathematics, logic, and language (Miller, 1997, in Berk, 2009), while the right hemisphere is responsible for creative processes and emotions, focussing on art, colour, pictures, and music (Jourdain, 1997; Polk and Kertesz, 1993, in Berk, 2009). When watching a video, the left side processes the dialogue, plot, rhythm, and lyrics; the right

side processes the visual images, relationships, sound effects, melodies, and harmonic relationships (Hébert and Peretz, 1997; Schlaug et al., 1995, in Berk, 2009).

3.2 Why use documentary videos?

It is crucial for a system to yield a high number of relevant questions, i.e., questions based on significant information in the source text (Vanderwende, 2008). For a system which generates factual sentence-to-text questions, it is necessary for the source texts to contain fact-based information; otherwise it will not be able to generate a high number of high-quality questions for its determined question types. For this reason, documentaries were chosen as input for the system, rather than films from other genres such as dramas or comedies. Documentaries lend themselves well to factual question generation, because they typically serve to depict aspects of reality and in doing so, they frequently present their viewers with factual information about a certain topic. For many applications of QG, the Question Generation procedure can be regarded as a three-step-process consisting of concept selection, question type determination and question construction (Nielsen, 2008). A premise for a high-quality QG system is the formation of questions based on key concepts, i.e. snippets of source text which carry vital information. While key concepts may vary depending on the context of the application, amongst other evaluation criteria, a high quality system is characterised by high precision and recall when identifying key concepts (Rus and Graesser, 2008).

Documentary subtitles are a transcript of what is being said; the textual information can be exploited using NLP techniques. Other QG systems which generate factual questions from

texts alone only deal with one layer of information. However, a system which generates questions from videos can also exploit a ‘second layer’ of information, the visual layer. Nichols (2001) points out that the images in a documentary are often distinct from the commentary; the images can illustrate, support or counter what is being said. There is a meaningful relationship between the commentary and the images; the images are organised through the commentary. The use of videos thus opens up a variety of possibilities for the creation of an application that combines different types of information. Not only can information be extracted in the form of text, but it is also possible to extract images, short video sequences, sounds, etc., to accompany the generated questions. For example, a question can be supported by supplying the answer (key) in the form of an image. If the key is a person, a picture of that person can be shown alongside the question stem. When generating multiple-choice questions, it is possible to extract distractors (i.e., answer options which are similar to the key, but are nevertheless incorrect, and serve to distract the user from choosing the correct answer) from image databases such as ImageNet (Deng et al., 2009). While other QG systems, such as the state-of-the art system by Heilman (2011), employ, at the core, a similar methodology to the one employed by system WLW (linguistic pre-processing and a rule-based approach), system WLW has the unique ability to make use of the visual layer by extracting a screenshot to be supplied alongside a question. This image is used to help test-takers answer more questions correctly by attracting and focussing their attention on important sections in the video. A user study (cf. Chapter 5) showed that supplying a screenshot alongside a question is indeed beneficial, as test-takers who answered questions accompanied by a screenshot had a statistically

significantly higher score on a comprehension test about a scientific documentary than those who did not receive these images.

Another reason for using documentary videos is the fact that they are frequently employed in classroom settings. There is an EU-funded project, ‘Activewatch’¹⁵, which is part of the ‘Lifelong Learning Programme’¹⁶ and is designed to provide workshops for educators on how to use documentary videos to their advantage in the classroom. The Lifelong Learning Programme has a budget of €7 billion to invest in educational projects; the fact that the use of documentary videos is being supported by this initiative goes to show that there is a belief that documentary videos can be beneficial in teaching. To the best of the knowledge, system WLW is the first of its kind which can generate factual questions from documentary videos and it can support teachers and educators in the tedious, expensive and time-consuming task of generating assessment materials.

3.3 Documentary genres

From a film-making perspective, there are no strict boundaries as to what defines “documentary” videos; however, the general view is that documentary videos serve to document aspects of reality, and to instruct and maintain a historical record. Documentaries can be divided into six sub-genres (Nichols, 2001); each of them has certain characteristics with which a specific type of effect can be achieved. *Poetic documentaries*, for example, draw on real historical facts but transform the material in distinctive artistic ways, while *reflexive documentaries* try to achieve a shift in awareness

¹⁵ <http://www.activewatch.ro/en/latest-news/workshop-for-teachers-use-of-documentary-film-in-the-classroom/>

¹⁶ http://ec.europa.eu/education/lifelong-learning-programme/doc78_en.htm

in the viewer by engaging with the audience, rather than the subjects. In *observational documentaries*, the film maker does not intervene at all, while in participatory ones, the film maker actively takes part in events, often by using interviews. *Performative documentaries* are subjective accounts which break away from factual accounting. The final type of documentary described by Nicholls (ibid.) is the *expository* one. This type is the one employed in the system proposed. Expository documentaries assemble facts into an argumentative frame. Information is typically relayed by the spoken word; two common techniques employed are the ‘voice-of-god’ commentary, a voice-over where the speaker is heard but not seen and the ‘voice-of-authority’, in which the speaker is heard and seen (ibid.).

3.4 Using subtitles in Question Generation

Automatic Question Generation is a far from being a straightforward process and there are many challenges to be overcome when generating questions from documentary video subtitles. One of these challenges is the fact that the nature of subtitles has not yet been fully understood. There are certain characteristics of subtitles that require special treatment from a QG system. This Section will describe the characteristics of subtitles and how the QG system accommodates them.

3.4.1 Accessibility of subtitles

Subtitles tend to be available in one of two forms; they are either hard-coded (embedded) into the visual layer of a video or can be accessed as a separate subtitle file. When subtitles

are hard-coded, they cannot easily be accessed and special software is required to extract them; this is time-consuming and it may not always be accurate. For this reason, hard-coded subtitles for are not used with the system; instead those that are available for download as a separate file alongside the video are employed. These subtitle files can for example be obtained from the online video streaming website BBC iPlayer¹⁷. Due to advances in automatic speech recognition, there has also been a trend towards automatic subtitling. YouTube implemented automatic subtitling functionality in 2009. In theory, this makes a plethora of videos available to use with system WLV, however, the automatically generated subtitles are often inaccurate, which is why videos and subtitles from the YouTube website are not used.

3.4.2 Format and structure of subtitles

Subtitles are available in a file format called srt, which adheres to a specific internal structure, as can be seen in Figure 4. The subtitle file contains the text in chunks of three or more lines. The first line is a line number, the second line is reserved for the time stamp indicating the exact seconds and milliseconds in which the text, in line three or sometimes four, is spoken. There are no fixed rules as to how many characters per line and whether two or more lines of

¹⁷ <http://www.bbc.co.uk/iplayer/>

```
6
00:00:43,040 --> 00:00:45,480
Your ability to talk.

7
00:00:45,480 --> 00:00:48,480
Or more precisely, the way you use language.

8
00:00:50,480 --> 00:00:53,520
THEY SPEAK IN NATIVE TONGUE
```

Figure 4 Screenshot of srt subtitle file

dialogue are used; amongst others, this depends on the subtitling workstation used. However, as readability is crucial when it comes to using subtitles, it has been suggested that an ideal subtitle is one sentence long with its clauses spread over two lines (Díaz Cintas and Remael, 2007). An advantage of the subtitle files provided by BBC iPlayer is that they adhere to grammar, spelling and punctuation rules and there is no need for correction. However, the files still need to be converted into a format which the QG system can process, i.e. the time stamps need to be removed and the lines of text need to be concatenated to form full sentences. This is necessary, because during the first step of the QG process with system WLW, the text is pre-processed with several NLP resources, such as a sentence splitter, part-of-speech tagger and dependency parser and in order for these resources to function correctly, the text needs to be input as a coherent whole. The time stamps are only removed temporarily, as they are still required at a later stage when the

system extracts a screenshot from the video to display alongside a question (the methodology will be explained in Chapter 4 and an evaluation of the approach in Chapters 5 and 6).

3.4.3 Challenges posed by using subtitles

While being able to exploit the visual layer of documentary videos with the QG system adds unique benefits to automatically generated questions, it also creates challenges. As mentioned in the previous Section, before subtitles can be linguistically processed, they need to be put into a specific format. Another difficulty arises when there is a mixture of voice-of-god and voice-of-authority commentary being employed. While during the voice-of-god commentary, what is being said is usually well-formed and grammatically correct, when the narrator is seen and heard, he often engages in dialogue with participants and consequently the text shifts from factual, informative text to a conversation or dialogue. Trying to generate questions from such sentences can lead to unusable questions. For example, in one scientific documentary on nuclear fusion (Horizon, 2005) the following discourse is encountered:

The two researchers who claimed to have made the breakthrough have made no comment themselves. There was a conflict situation really, the newspapers, which escalated in the University, hmm, it was very bad, hmm. Now, even Professor Fleischmann acknowledges he made a mistake.

The first and third sentences in this example are uttered by the narrator, whereas the second sentence is spoken by a scientist being interviewed in the documentary. As the subtitles do not always contain punctuation marks for direct speech, there is no easy way to distinguish between comments made by the narrator and comments made by people within the documentary. While the use of fillers such as ‘hmm’ could be used to identify direct speech, such fillers are not always used and sometimes there simply is no way of knowing who made an utterance without actually having watched the documentary. One way to identify whether an utterance is made by the narrator or a person within the documentary would be by using Computer Vision (CV) techniques. Computer Vision is a complex research area that is concerned with establishing methods for acquiring, processing, analysing, and understanding images and there is a wide range of applications. Whilst the implementation of CV into the QG framework could yield interesting new insights, it is a research area in its own right and thus out of the scope of this thesis. In addition, even if one was to find out who makes an utterance, it is not possible to simply ignore all sentences uttered by persons within the documentary; even if informal speech is employed, the sentences may still contain important information which may make that sentence a good candidate for a question. For example, from this sentence in a documentary about the history of Christianity: “Yep. All historical Christian churches face east, as you know. Yep, east, west, yep better than anybody else”¹⁸, it would be possible to generate the question: “Where do all the historical Christian churches face?”, even though at first, the sentence might seem unsuitable for question generation.

¹⁸ From “A history of Christianity” (BBC, originally aired 2009)

3.4.4 Comparing subtitles to other text types

The aforementioned challenges are far from an exhaustive list of the difficulties that may be encountered when generating questions from video documentaries. One of the problems is that subtitles as a text type are not yet fully understood. In order to gain a greater insight into the characteristics of subtitles and what distinguishes them from other texts, a qualitative linguistic analysis of several subtitles texts and comparable texts extracted from Wikipedia, using a tool called Coh-Metrix (Graesser et al., 2004), was performed. Developed by researchers at the University of Memphis, Coh-Metrix is a computational tool that produces indices of the linguistic and discourse representations of a text which can be used in a variety of ways to investigate the cohesion of the text. Altogether, Coh-Metrix calculates 108 different linguistic indices; the indices are subdivided into eleven groups: (1) Descriptive, (2) Text Easability Principal Component Scores, (3) Referential Cohesion, (4) LSA, (5) Lexical Diversity, (6) Connectives, (7) Situation Model, (8) Syntactic Complexity, (9) Syntactic Pattern Density, (10) Word Information, and (11) Readability.

For the analysis, Coh-Metrix was used to produce indices for the subtitles of five different documentaries: one historical, one scientific, one biographical documentary, one documentary about food and one about linguistics (in order to process the subtitles with Coh-Metrix, just like for use with the framework, the text from the subtitle files was first extracted and formed into coherent sentences). Then the indices were compared to the indices generated for five ‘comparable’ articles taken from Wikipedia; ‘comparable’ meaning ‘similar in topic’ and length. For example, as the biographical documentary

revolves around Steve Jobs, these subtitles were compared to the Wikipedia article for Steve Jobs. All texts were approximately 1500 words long. Wikipedia articles were chosen due to the ease with which they can be accessed, but for future work on a larger scale analysis, one could also compare the indices generated for textbook articles. The most important indices of the analysis can be seen in Table 2, while the full table can be found in Appendix A.

Coh-Matrix Index	History Subtitles	Science Subtitles	Biography Subtitles	Food Subtitles	Linguistics Subtitles	History Wikipedia	Science Wikipedia	Biography Wikipedia	Food Wikipedia	Linguistics Wikipedia
DESWC 'Word count, number of words'	1598	1549	1604	1628	1553	1579	1547	1643	1379	1716
DESSL 'Sentence length, number of words, mean'	12.388	16.479	12.73	9.988	20.169	19.256	24.952	21.064	17.456	22.88
PCNARz 'Text Easability PC Narrativity, z score'	0.345	-0.318	0.272	0.46	-0.961	-0.39	-1.138	-0.407	-1.675	-1.186
PCSYNz 'Text Easability PC Syntactic simplicity, z score'	0.675	0.296	0.605	0.853	-0.104	-0.31	-0.529	-0.262	0.237	-0.552
PCREFz 'Text Easability PC Referential cohesion, z score'	-0.846	-0.999	-1.235	-1.065	-0.985	-1.125	0.466	-0.914	-0.514	-0.267
PCVERBz 'Text Easability PC Verb cohesion, z score'	0.868	0.584	0.291	0.388	0.218	-0.557	-0.273	-0.204	-1.585	0.263
CNCAI 'All connectives incidence'	81.352	78.76	71.696	73.096	103.026	82.331	85.326	93.122	97.897	90.909
CNCCaus 'Causal connectives incidence'	25.031	32.924	28.055	21.499	19.961	15.833	38.785	17.651	26.106	16.9
CNCLogic 'Logical connectives incidence'	35.044	36.152	28.678	33.17	28.976	21.533	44.602	29.215	23.93	26.224
CNCADC 'Adversative and contrastive connectives incidence'	13.141	9.684	9.352	13.514	16.742	8.866	15.514	9.738	9.427	11.655
CNCTemp 'Temporal connectives incidence'	20.651	12.266	18.08	11.671	11.59	20.899	16.16	29.215	15.228	23.893
CNCTempx 'Expanded temporal connectives incidence'	16.896	18.722	17.456	16.585	28.332	16.466	16.807	15.216	20.305	18.648

CNCAdd 'Additive connectives incidence'	41.302	34.216	31.172	39.926	72.762	43.699	36.846	45.648	55.838	55.944
DRPVAL 'Agentless passive voice density, incidence'	6.258	10.975	4.988	5.528	3.863	8.866	14.221	8.521	10.152	8.741
WRDNOUN 'Noun incidence'	260.326	240.155	260.598	196.561	311.656	310.956	281.189	332.928	372.009	351.399

Table 2 Coh Metrix indices for subtitles and Wikipedia articles

The produced indices yield a number of interesting findings when comparing subtitles. Starting with the descriptive indices, the mean sentence length in words (DESSL) is higher for Wikipedia articles (\bar{x} = 21.1216) than subtitles (\bar{x} = 14.3508) and the difference is statistically significant ($p=0.001522$). One reason for this could be the fact that subtitles often contain dialogue, which often consists of short sentences, just like in this documentary about Pyramids¹⁹:

- *All lined in fine dressed masonry.*
- *Oh, my goodness!*
- *A little staircase.*

In the narration of the documentary, it is possible to observe both long and short sentences. Often but not always, when the narrator is seen (voice-of-authority), more colloquial language is used, while when a voice over is used (voice-of-god), the narrator reads out a script, which often results in the language used being formal, with syntactically complex sentences. For a QG system, it is generally easier to process simple sentences; the simpler the input sentence is from a syntactic perspective, the bigger the likelihood that the sentence is processed correctly by the various components. Conversely, due to the QG system following a specific order of steps to generate questions (as explained in Chapter 4), if a complex sentence is input to the system and errors in the processing occur in the early stages of processing (for example in the syntactic parsing stage), then this error will be cascaded into the other stages of the QG process and lead to more errors and eventually an unusable question.

¹⁹ “The Man who discovered Egypt”, BBC, (2011)

Another index calculated by Coh-Metrix is the narrativity score (PCNARz). Narrative texts tell stories and are characterised by characters, events, places and the use of oral language. Not surprisingly, the Wikipedia articles analysed score lower on narrativity than the subtitle texts, as documentaries often have a story-like structure and as mentioned before, contain oral language. Out of the subtitle texts, the history documentary subtitles scored highest in terms of narrativity, while the scientific documentary subtitles scored low as well.

In Coh-Metrix, a low syntactic simplicity score (PCSYNz) implies that the texts contain sentences with more words and complex, unfamiliar syntactic structures which are more difficult for the reader to process. In terms of syntactic simplicity, all Wikipedia texts apart from the one about food score lower than the subtitle texts. For a QG system this could mean that the likelihood of producing a larger amount of syntactically correct questions is higher when using subtitle texts, since more complex sentences can pose problems to the individual processing components and errors might be cascaded throughout the process.

Coh-Metrix produces several cohesion indices, one of them being referential cohesion (PCREFz). A text with high referential cohesion contains words and ideas that overlap across sentences and the entire text, forming explicit threads that connect the text, whereas low cohesion text is typically more difficult to process for humans because there are fewer connections that tie the ideas together. As there are often quickly changing scenes in documentaries, interlaced with switches between dialogues and narration, one could assume that subtitle texts score lower on referential cohesion values than the Wikipedia

articles. However, this is not the case; there is no statistically significant difference between the referential cohesion values for the two text types ($p=0.060$).

Another cohesion index is verb cohesion (PCVERBz), which is an indicator of the extent of verb overlaps in the text. If a text exhibits a large number of repeated verbs, it is likely to include a more coherent event structure which is characteristic of narrative texts. The subtitle texts exhibit a higher verb overlap than the Wikipedia articles and the difference is statistically significant ($p=0.027$). This is not surprising, as video documentaries have a narrative structure and often “tell a story”.

Connectives are another means of creating cohesive links between ideas and they also provide clues about the organisation of text. Coh-Metrix calculates an incidence score for five different types of connectives: causal (CNCCaus; *because, so*), logical (CNCLogic; *and, or*), adversative/contrastive (CNCADC; *although, whereas*), temporal (CNCTemp; *first, until*), and additive (CNCAdd; *and, moreover*). It would be fair to assume that Wikipedia articles have a higher incidence of connectives due to the written, formal language used as opposed to the documentary subtitles. However, when performing a significance test on the scores for the connectives, there is only a statistically significant difference between the text types when comparing the values for temporal connectives (causal: $p=0.255$, logic: $p=0.218$, adversative: $p=0.26$, temporal: $p=0.028$, additive: $p=0.28$) with the Wikipedia articles exhibiting a higher number of temporal connectives. The biographical Wikipedia article and the history articles for both subtitles and Wikipedia score the highest on temporal connectives (29.215, 20.615 and 20.899 occurrences per 1000 words, respectively). This might be due to the chronological structure that

biographical and historical texts often adhere to. Information about the incidence of discourse connectives could be relevant to the QG system in the future. In its current state, the system forms questions on a per sentence basis (see Chapter 4). The subtitle texts are ‘scanned’ for suitable ‘candidate sentences’ (for example, to form ‘who’ questions, sentences containing persons are used). This means that system WLV can currently only exploit explicit information contained in the sentence that has been selected as question candidate, but not beyond that. However, information is sometimes contained implicitly, spanning over several sentences or in a paragraph. For example, consider the sentences: “Thursday’s outcome in the Commons was an immense defeat. Not simply because Cameron lost the vote, but because he had lost the support of much of Parliament and of his own party.” Currently, while system WLV could generate the question “Who lost the vote?”, it would not be able to identify that these two sentences are logically related. A good question that could be generated from these sentences would be: “Why was Thursday’s outcome in the Commons an immense defeat?”. Generating this question would be possible, if the QG exploited connectives in combination with a discourse representation theory, such as Rhetorical Structure Theory (Mann and Thompson, 1987).

Another interesting finding is that there is a statistically significantly higher incidence of agentless passives (DRPVAL) in the Wikipedia articles ($p=0.00048$). Agentless passive constructions, such as “data were collected”, are typical of scientific texts, where the emphasis is put on what was done and not who did it. While system WLV can generate questions from agentless passive constructions (e.g. from “The Louvre was built in 1793.”, it is possible formulate the question “When was the Louvre built?”), higher incidence of

agentless passive constructions will lead to a reduced number of “who” questions, because there are no persons in subject position (a premise for generating a ‘who’ question).

Interestingly, however, the Wikipedia articles exhibit a larger incidence of nouns (WRDNOUN, $p=0.020$). A large number of nouns in the source texts is important for the QG system, as its methodology is based on recognising named entities (nouns) in the text and generating questions based on them (see Chapter 4). Thus, if a text exhibits a larger number of nouns and named entities, this will mean that a larger number of questions can be generated.

Although this analysis is by no means exhaustive, by comparing various linguistic indices of subtitle texts from video documentaries to comparable articles from Wikipedia, an insight was gained into the linguistic structure and makeup of subtitles as a text genre. To the best of the knowledge, subtitles have not been analysed from a linguistic perspective before and this has been a gap in the literature. This analysis helps to understand characteristics and idiosyncrasies of subtitles as a text genre.

3.5 Summary

In this Chapter, documentary videos and subtitles in Question Generation were discussed. First, a look was taken at the use of videos in the classroom. While the use of videos in teaching is not a recent development, there have been some developments which have led to a rise in the popularity of video use in educational settings. For example, a large variety of video formats is available and there are different ways of using technology to facilitate theory application in the classroom. In addition, instructors can use a number of video

techniques and the research on multimedia learning that provides empirical support for their use as an effective teaching tool has grown steadily.

Documentary videos are increasingly popular in classroom settings and there are even EU-funded projects helping instructors to incorporate them into their teaching activities.

The QG system developed as part of this thesis generates factual questions from the subtitle text that accompanies a video documentary. This Chapter described how subtitles can be accessed and which format they typically follow. While there are a number of advantages to using subtitles, the challenges that they pose for Question Generation, such as the difficulty of identifying who makes an utterance (the narrator or somebody depicted within the documentary) when only looking at the subtitle text, were also discussed. Finally, a qualitative analysis of subtitle text compared to other texts was presented and various linguistic characteristics of subtitle texts were described. The next Chapter describes the framework for Question Generation in detail.

CHAPTER 4: A FRAMEWORK FOR QUESTION GENERATION FROM DOCUMENTARY VIDEOS

This Chapter describes the Question Generation framework which has been developed as part of this research project. Section 4.1 gives an overview of the methodology employed. In Section, 4.2, the GATE architecture on which the development of the employed approach has been based, is described. Section 4.3 describes how subtitle files need to be prepared in order to be processed by system WLIV. In order to generate questions, several linguistic pre-processing steps are performed using the GATE architecture. These steps are described in Section 4.4. Once the text has been pre-processed, a rule-based approach is employed to identify question candidate sentences and to transform them into questions. This is described in Section 4.5. System WLIV has the unique ability to make use of the information contained in the visual layer by supplying a screenshot from the video alongside a question. This process is described in Section 4.6. Two error analyses are also described. The results of these preliminary evaluations have been used to improve the quality of the questions generated by the framework. The first error analysis, outlined in Section 4.7, describes an evaluation of a set of 258 questions by human experts. In the second preliminary evaluation, described in Section 4.8, a qualitative error analysis of the framework's questions compared to those of a state-of-the-art system is performed and the usability of questions per transformational rule is analysed. Finally, Section 4.9 provides a summary of this Chapter.

4.1 Overview

For many applications of QG, the Question Generation procedure can be regarded as a three-step process consisting of concept selection, question type determination and question construction (Nielsen, 2008). A premise for a high-quality QG system is the formation of questions based on key concepts, i.e. snippets of source text which carry vital information (Mitkov and Ha, 2003). While key concepts may vary depending on the context of the application, amongst other evaluation criteria, a high quality system is characterised by high precision and recall when identifying key concepts (Rus and Graesser, 2008).

It is crucial for a system to yield a high number of relevant questions, i.e., questions based on significant information in the source text (Vanderwende, 2008). For a system which generates factual questions, it is necessary for the source texts to contain fact-based information, otherwise it will not be able to generate a high number of high-quality questions for its determined question types. For this reason, it was chosen to employ documentaries as input for the system, rather than films from other genres such as dramas or comedies. Documentaries lend themselves well to factual question generation, as they typically serve to depict aspects of reality and in doing so, they frequently present their viewers with factual information about a certain topic.

Figure 5 shows the conceptual overview of the QG framework. In order to generate

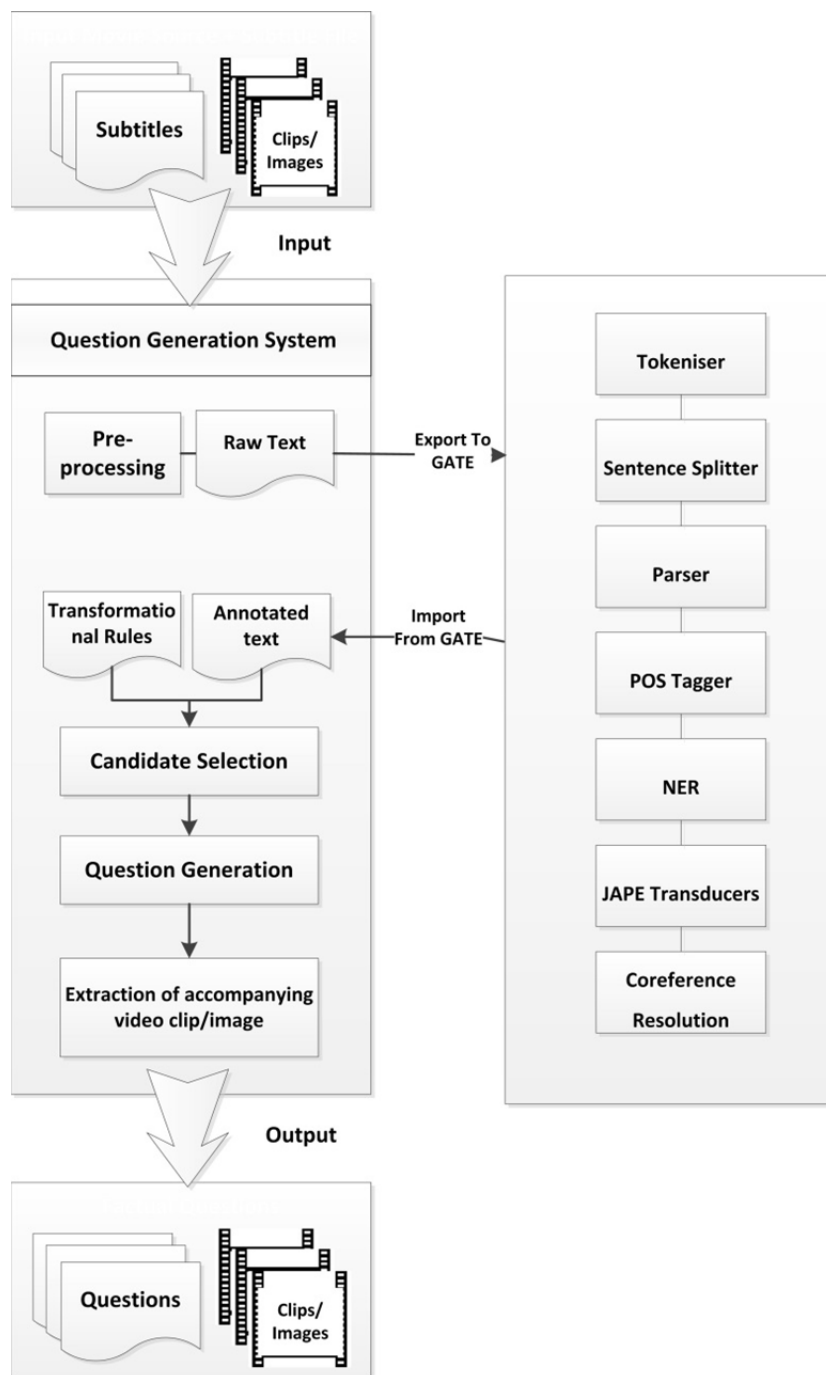


Figure 5 Conceptual design of the framework

questions, users are required to upload a subtitle file via the GUI. In the first step, the subtitle file is converted into a format that can be handled by GATE. After the Question Generation process, which involves a variety of processing tasks (described in Section 4.4), is completed, the questions and correct answers are presented to the user in a post-editing environment (see Figure 6) which allows for alterations to be made before exporting the question set. The post-editing environment allows users to freely alter the generated questions and their answers and it also allows users to see the context of a question, i.e. the source sentence of a question and its neighbouring sentences.

Questions generated for *Saving Britains Past*

Question	Answer		
Who was 21 when she witnessed the attacks?	Ruth Haskins	Context	<input type="button" value="Edit"/>
Who was The trailblazer of the town planning movement?	Patrick Abercrombie	Context	<input type="button" value="Edit"/>
An exhibition of whose plans for Bath was shown at the city's main art gallery in February 1945?	Patrick Abercrombie	Context	<input type="button" value="Edit"/>
Whose proposal was nothing if not ambitious?	Buchanan	Context	<input type="button" value="Edit"/>
When was the city of Bath awarded UNESCO World Heritage status?	In 1987	Context	<input type="button" value="Edit"/>
When did John Wood's son put on the finishing touch - The Royal Crescent?	1754	Context	<input type="button" value="Edit"/>
<input type="text" value="Where had Chamberlain built his political power ba"/>	<input type="text" value="In Birmingham"/>	Context	<input type="button" value="Save"/>

Content Clear All Highlights

Since World War Two, Britain's towns and landscapes have been transformed. The modern world has brought radical change and progress. It has also threatened to destroy historic buildings, countryside and centuries-old ways of life. People have fought back to save their heritage. The battles that have been won and lost reveal our changing attitudes to the past, present and future and have shaped how Britain looks today. It all began with the Blitz. In the early 1940s, Hitler sent the Luftwaffe to destroy cities, break morale and pave the way for Nazi invasion. The deadly night-time air raids left a trail of devastation and fear. Town squares, medieval streets and ancient cathedrals were all in the firing line. When the beautiful city of Bath was attacked in April 1942, it seemed nowhere was safe. As the bombing continued, panic spread over how to protect Britain's heritage and preserve what was left for future generations. Out of this crisis came the energy and desire to save all of Britain's most important buildings. Choices would have to be made - what to hold onto, what to let go. But would these decisions be enough to save Bath, Britain's most precious city? Bath is the jewel in the crown of British cities. It was built during a golden age for British

Figure 6 Screenshot of post-editing environment

4.2 The GATE architecture for Natural Language Processing

The framework employs the NLP environment GATE (Cunningham, et al., 2008), which provides a collection of tools for a large variety of human language processing purposes. The development of GATE started over 15 years ago and is one of the biggest open source language processing projects. GATE can be used as an integrated development environment (IDE, GATE Developer) or via an application programming interface (API, Gate Embedded); the latter allows developers to integrate GATE into one's own JAVA applications. By default, GATE includes a large number of NLP plugins²⁰; amongst others there are tools available for Information Extraction, Annotation, Alignment, Parsing, and Machine Learning. There are also tools to allow working with Ontologies as well as a number of language-specific tools.

GATE was designed assuming that Natural Language Processing tasks can be broken down into several subtask and executed by components referred to as 'resources' within GATE. There are three types of resources:

- Language Resources (LRs) represent entities such as lexicons, corpora or ontologies;
- Processing Resources (PRs) represent entities that are primarily algorithmic, such as parsers, generators or ngram modellers;
- Visual Resources (VRs) represent visualisation and editing components that participate in GUIs.

²⁰ The list of default plugins can be seen here: <http://gate.ac.uk/gate/doc/plugins.html>

The screenshot of the GATE Developer GUI (Figure 7) exemplifies the different types of resources. In the left hand pane, the subtitles of a history documentary have been loaded into a corpus as a Language Resource. Underneath, the several Processing Resources that are required in the Question Generation Process can be seen. The middle window, the annotation pane (Visual Resource), displays the subtitle text which has been processed with the PRs. One of GATE's main characteristics is that linguistic information is added to the text in the form of annotations. In the screenshot in Figure 7, the right hand side pane shows all the annotation types that have been added by the various PRs. One can see that the subtitle text has been tokenised (word boundaries have been identified). The annotation types, 'Date', 'Person' and 'Organisation', for example, have been added by the Named Entity Recognition (NER) Processing Resource. The way that GATE deals with annotations lets a user perform endless powerful operations and manipulations on text in a very user-friendly way. The users are not only restricted to using GATE's predefined annotation types, but also have the freedom to create their own annotations and use a GATE-specific language (JAPE, see Section 4.5) to automatically identify their own annotation types.

The main idea behind GATE is that language resources are run over texts sequentially and in each step the text is linguistically enriched with information in the form of stand-off XML annotations and the information gathered in previous steps is exploited in subsequent steps. One of the main reasons for choosing GATE over other readily available toolkits, such as NLTK (Bird, 2006) was the fact that GATE Developer could be used to easily run and visualise test scripts for Question Generation using JAPE, while GATE Embedded

was then used to design a bespoke Question Generation system. Another great advantage of the framework is the fact that it is easily customisable with regards to the transformational question rules and components employed. The rules are not hard-coded into the system; instead, each rule exists as a separate JAPE file which is simply put into a specific destination in the file directory of the QG system and then the QG system accesses it from there. When developing the framework, emphasis was put on the development of the rules and how the framework could be improved by using different rules. The chosen approach allows for easy adapt to easily adapt rules and make judgement about how each individual rule performs. In addition, as mentioned previously, an abundance of Processing Resources has been developed for GATE and whichever PRs a user decides to use, these can easily be integrated into the system. Thus, even though the system at the moment is only designed for English, if a user wishes to do so, they could easily integrate new question rules and processing resources for other languages.

The screenshot displays the GATE Developer 7.1 build 4485 interface. The main window shows a document editor with a text document containing a paragraph about Adela Pankhurst. The text is annotated with various entities and relations, such as dates, locations, and persons. A table below the text lists these annotations with their types, sets, start and end positions, IDs, and features.

Type	Set	Start	End	Id	Features
Location		1838	1845	5481	{locType=country, matches=[5476, 5477, 5480, 5481, 5489, 5494, 5505, 5528, 5531, 5537, ...]}
Date		1889	1903	5482	{kind=date, rule1=DateName, rule2=DateOnlyFinal}
Person		1926	1941	5483	{gender=female, matches=[5483, 5486, 5487, 5648, 14944, 14952, 14975, 14993], rule=Pe...}
Potential Question		1926	2005	14997	{Answer = Adela Pankhurst, Question =Who made her way to a park in Manchester called B...}
Person		1947	1950	14944	{ENTITY_MENTION_TYPE=PRONOUN, antecedent_offset=1926, matches=[5483, 5486, 54...]}
Location		1968	1978	5484	{locType=city, matches=[5484, 5512, 5517], rule1=InLoc1, rule2=LocFinal}
Person		2007	2010	14952	{ENTITY_MENTION_TYPE=PRONOUN, antecedent_offset=1926, matches=[5483, 5486, 54...]}
Person		2069	2086	5485	{gender=male, matches=[5485, 5497], rule=PersonFinal, rule1=PersonFull}
Person		2109	2112	14975	{ENTITY_MENTION_TYPE=PRONOUN, antecedent_offset=1926, matches=[5483, 5486, 54...]}

The right-hand side of the interface shows a list of annotation sets, including Date, Dependency, FirstPerson, JobTitle, Location, Lookup, Organization, Person, PleonasticIt, Potential Question, QuotedText, Sentence, SpaceToken, Split, SyntaxTreeNode, Title, Token, and Unknown. The 'Original markings' section is also visible.

Figure 7 Screenshot of the GATE Developer GUI

4.3 Setting up subtitles for use with GATE

Once a file has been uploaded to the system, several minor processing steps are required in order for the texts to be processed correctly by GATE. The original subtitle file contains the text in chunks of three or more lines. The first line is the line number, the second line is reserved for the time stamp indicating the exact seconds and milliseconds in which the text, in line three or sometimes four, is spoken. A script removes the line numbers and time stamps, as well additional information for the hard of hearing which describes actions or sounds and is written in capitals and then concatenates the remaining sentence fragments into one document.

4.4 Linguistic pre-processing

Several pre-processing steps using GATE's processing resources are performed. The resources are loaded into a so-called 'pipeline' and are run in a hierarchical order over the text. The order of the resources within the pipeline plays a crucial role, because each processing resource relies on information gained in the previous step(s). For example, tokenisation needs to be performed before sentence splitting can be performed and these two steps need to have been performed before Named Entity Recognition can take place.

4.4.1 Tokenisation

The first step performed is tokenisation. Tokenisation is a process of text segmentation in which a stream of text is divided into words, symbols or other meaningful elements, called

tokens. Generally, tokenisation is considered an easy task in comparison to other NLP tasks, because often simple heuristics are sufficient to divide text into tokens. Usually, tokens are continuous strings or alphabetic characters separated by whitespace characters, such as a space, a line break or a punctuation mark. Despite the fact that the task is considered to be easy, it is not always straightforward. Punctuation marks, for example, are not exclusively used to delimit sentences; they can also be used in abbreviations. Another issue is that abbreviations can have multiple meanings, for example ‘in’ can stand for ‘inches’, but it could also refer to the preposition ‘in’. This does not usually pose a problem to modern tokenisers, because they are trained with a list of common abbreviations and probabilistic models can help distinguish between different word senses. For the QG system and (all other NLP applications for that matter) it is important that the tokeniser output is as accurate as possible, as the list of tokens becomes input for further processing; errors made in the tokenisation stage will propagate into later stages and create problems.

4.4.2 Sentence splitting

The sentence splitter segments the text into sentences. Running this GATE module is compulsory if the POS tagger is employed. The POS tagger relies on the input from the sentence splitter. The sentence splitter annotates every sentence with the annotation type “Sentence” and each sentence break with the annotation type “split”.

4.4.3 Part-Of-Speech Tagging

Part-of-speech (POS) tagging is the process of assigning word category labels to text. Typically, in linguistics, these are thought of as nouns, verbs, adjectives, adverbs, etc; however, in a computational context, the distinction is often very fine-grained and thus many more labels are used. There are, for example, labels to distinguish between singular (NN) and plural nouns (NNS), as well as for different categories of verbs. VBD, for example, stands for verb, past tense and VBZ stands for verb, 3rd person singular present, etc. (Santorini, 1990). In GATE, POS tags are assigned using the ANNIE suite of processing resources, which is a bundle of processing resources aimed at Information Extraction, built by the GATE developers. The POS tagging process is a prerequisite for the Parser (in the subsequent step) to function correctly.

4.4.4 Syntactic Parsing

In Natural Language Processing, syntactic parsing refers to the process of computationally analysing a sentence's component categories and grammatical functions. A number of different parsers have been implemented for use with GATE; the parser used in the system is the Stanford Dependency Parser (Marneffe and Manning, 2008). The dependencies provide a representation of grammatical relations between words in a sentence and occur in the form of triplets: name of the relation, governor and dependent. To exemplify this, consider the following sentence:

The dog chases the cat.

The Stanford dependencies for this sentence would be:

det(dog-2, The-1)
nsubj(chases-3, dog-2)
root(ROOT-0, chases-3)
det(cat-5, the-4)
dobj(chases-3, cat-5)

Here, the first part is the name of the relation, followed by the governor and the dependent. The number behind a word constitutes the position of the word in the sentence. Thus, in the example, the relation between ‘dog’ and ‘the’ is that of a determiner. ‘Nsubj’ stands for nominal subject and signifies that the noun phrase “dog” is the syntactic subject of the clause. ‘Cat’ is the direct object (dobj) of this example sentence. In the QG framework, these dependencies play a major part in the identification of question candidates and the formation of questions. As will be explained in more detail in Section 4.5, the framework uses a rule-based approach. After the linguistic pre-processing, several manually created rules identify question candidates based on the information gained in the pre-processing steps. For example, in order to generate a ‘who’-question about a person, for one, those sentences need to be found that contain person names (this information is gained from the Named Entity Recogniser, described in the next Section) and secondly, person name also needs to be in an ‘nsubj’ relation.

4.4.5 Named Entity Recognition and Gazetteer Lookup

In Natural Language Processing, a Named Entity is a span of text that can be assigned a proper name (Jurafsky and Martin, 2008). Named Entity Recognition (NER) is concerned

with identifying proper names in text and assigning predefined category labels to them, such as the names of persons, organisations, locations, expressions of times, quantities, monetary values, percentages, etc. NER is a subtask of Information Extraction, but it finds application across a wide range of NLP contexts, because Named Entities often carry important information in text. NER also plays a crucial role for the QG system, as the key concepts in the text are named entities. As will be explained in detail in Section 4.5, after the pre-processing with GATE's resources, a set of manually created rules is employed to identify question candidates in the text and to transform them into questions. With the combined input from the NER module and the dependency parser, it is possible, for example, to generate "who" questions, by looking for people names in subject position. In the sentence "Einstein developed the general theory of relativity", the NER module would identify 'Einstein' as a person and from the dependency parser it would become clear that Einstein is the subject of the sentence. Thus, the system could generate the question "Who developed the general theory of relativity?", with 'Einstein' being extracted as the answer to the question.

4.4.6 Morphological analysis

In Linguistics, the study of Morphology is concerned with the analysis of the structure of words. Words are broken down into smaller, meaningful linguistic units, called morphemes, including roots and affixes. Language can be classified by the way morphemes are used (morphological typology). In GATE, there is a processing resource called 'morphological analyser'. It annotates each word in a text with morphological

information. For certain types of questions (where an auxiliary verb needs to be inserted), the QG system needs to identify the base form of verbs and this can be done using the information from the morphological analyser module. For example, consider the sentence “The First World War ended in 1918.”. If the sentence is transformed into a question, the auxiliary verb ‘did’ needs to be inserted and the main verb ‘ended’ needs to be used in present tense to generate the grammatically correct question “When did the First World War end?”

4.4.7 Pronoun resolution and sentence simplification

In Linguistics, an “anaphor” is an entity in text that refers back to a previous item (“antecedent”). The process of determining the antecedent of an anaphor is called Anaphora Resolution (Mitkov,1999). For example, in the sentence “Jake likes to play football and he does not like swimming”, “he” is the anaphor to the antecedent “Jake”. There are a number of different types of anaphora and when a personal pronoun is being resolved, as in the example sentence, this is referred to as Pronoun Resolution. While humans can often resolve anaphora intuitively, achieving the same computationally is still a big challenge in Natural Language Processing, due to a number of reasons. While semi-automatic anaphora resolution systems achieve an accuracy of around 78%, fully automatic systems still only achieve an accuracy of about 62% (Mitkov, 2001). In the framework, pronoun resolution is performed using the output from GATE’s pronominal co-referencer. Unfortunately, there is no data available about the accuracy of this GATE component, but it is important to bear in mind that the quality of questions generated by

the framework is dependent on the performance of the components it relies on. If the GATE pronominal co-referencer resolves a pronoun incorrectly, this will lead to an incorrect question. If it fails to resolve a pronoun completely, this means that no question will be generated from a sentence that might have led to a good question if the pronoun had been resolved as the system is designed to ignore sentences with unresolved pronouns, as will be explained next.

All questions generated by system WLV are formed from single sentences. If pronoun resolution was not performed, or sentences with unresolved pronouns were not ignored, some questions would be unusable because they contain personal pronouns and a test-taker might not know what they are referring to. From the sentence “He developed the general theory of relativity”, it would still be possible to form the question “Who developed the general theory of relativity?” but the system would as answer extract “he”, which would be meaningless. For this reason, in all simple sentences (i.e. sentences containing one independent clause, “The boy laughed”) and complex sentences (i.e. sentences with one independent and at least one dependent clause, “I read the newspaper before I go to work”) contained in the text, first-mention pronouns are replaced with the longest co-referent in the co-reference chain. The use of the longest co-referent ensures that answers about persons appear in a consistent format where different names are used to refer to the same person. For example, if a text about Albert Einstein refers to the physicist interchangeably as ‘Einstein’, ‘Albert Einstein’ and ‘Dr. Albert Einstein’, the pronoun resolver will replace all first-mention personal pronouns referring to Einstein with ‘Dr. Albert Einstein’. In independent clauses in compound sentences, not only first-mention pronouns, but all

subject personal pronouns will be replaced with their co-referents. For example, the sentence

Einstein discovered the theory of general relativity and he received the 1921 Nobel Prize in Physics would be transformed to

Dr. Albert Einstein discovered the theory of general relativity and Dr. Albert Einstein received the 1921 Nobel Prize in Physics.

In the following step, these compound sentences are then split into several sentences with initial conjunctions deleted, so that from the example sentences, two questions can be generated (*Who discovered the theory of general relativity? Who received the 1921 Nobel Prize in Physics?*). If the pronoun resolution was not performed and the sentence not split in this manner, the system would generate the syntactically awkward questions **Who discovered the theory of general relativity and he received the 1921 Nobel Prize in Physics?* and *Albert Einstein discovered the theory of general relativity and who received the 1921 Nobel Prize in Physics?* It would have been possible to design the transformational rules to output the question *WHO discovered the theory of general relativity and WHO received the 1921 Nobel Prize in Physics?*. However, it was decided that splitting the sentences before question generation would be the best option, as more complex sentences are more likely to be processed incorrectly by the GATE components and the transformational rules. By choosing this methodology, the introduction of potential mistakes in questions, which could cascade through other stages of the QG process, is avoided.

The splitting of complex sentences, together with the removal of all initial conjunctions, can be regarded as a form of simplification of the input text (albeit a minor one); the

investigation of other implementations which benefit the question formation process, such as dealing with restrictive and non-restrictive appositives and dependent clauses, is part of on-going research.

4.5 Rule-based approach to question generation

The framework generates questions by applying several algorithms, written in a GATE-specific format (JAPE), to the subtitles with the aim of finding question candidates and transforming them into questions. This step is performed after the pre-processing steps described in the previous section. The algorithms, or ‘JAPE rules’, as they will be referred to, consist of a left hand side (LHS), which is used to match a pattern in a GATE corpus (the subtitle text) using annotations gained in the pre-processing stage and regular expressions and a right hand side (RHS) which is used to indicate the action to perform and to manipulate the text/parse trees. LHS and RHS are divided by an arrow (“-->”). Figure 5 shows a very basic JAPE rule which looks for occurrences of the preposition “in” followed by a year in a corpus. If this is found, the sequence (e.g. “in 1985”) would then be assigned the annotation type “inYear”.

```
{Token.string == "in"}  
{Year}  
) :date  
-->  
:date.inYear = {rule = "inYear"},
```

Figure 5 Annotating Dates in text using JAPE

4.5.1 Question rules and helper rules

In the framework, there is a distinction between *question rules* and *helper rules*. Question rules are used to identify question candidates in the source text. Currently, question rules can generate questions from sentences in which persons occur in subject position (who), in object position (whom) and possessive constructs (whose), as well as temporal (when), spatial (where) and ‘what’ questions. Helper rules have been implemented to deal with tasks which need to be performed repeatedly, such as the identification subjects, adjuncts, groups of persons, etc. Implementing helper rules for the repeating tasks also helps to speed up the Question Generation process and to make the question rules easier to read. An overview of rules employed and their functions can be seen in Table 3.

By using the linguistic information made available in the pre-processing steps and the application of syntactic transformations (such as wh-movement and subject-auxiliary inversion) question candidates are transformed into questions. Question candidates are all those sentences that satisfy the conditions specified on the LHS of the question rules. Question and helper rules are run over each sentence in the subtitle and ‘fire’ (i.e., perform certain actions defined on the RHS) if they match a certain pattern specified on their LHS. All helper rules are run over the text first; question rules can then access the information added by the helper rules where applicable. The framework is designed so that all rules are run over every sentence in the source text; thus, one sentence can lead to one, several or no questions. Consider the following sentences: “Linguistics is the scientific study of language.”, “During the 20th century, Ferdinand de Saussure distinguished between the notions of *langue* and *parole*.” and “He was a Swiss linguist and semiotician“. From the

first sentence, the framework would generate the question: “What is Linguistics?”, as the question rule QR_What fires when there are NPs in source sentences which are not Persons, Dates or Locations (there are separate questions rules for each of these types). From the second sentence, the framework would produce the questions “Who distinguished between the notions of language and parole?”, “What did Ferdinand de Saussure distinguish between?” and “When did Ferdinand de Saussure distinguish between langue and parole?”. The order of questions generated is the same as the order in which they are presented in Table 3 (i.e. ‘who’ questions are always generated first). One could argue that in this case, the ‘what’ and ‘who’ questions about Saussure are better questions than the ‘when’ question, but this might not be the case for all sentences. For this reason it is left to the user to decide which question to use when one sentence leads to several questions. This approach raises some questions about effectiveness. Producing many questions per sentence and not ranking them in a specific way may mean that users end up wasting a lot of time discarding unwanted questions. One aspect is addressed in Chapter 6, where an experiment will show that post-editing (and discarding unwanted) questions is considerably faster than having a human expert create questions from scratch. On the effects on users of not ranking questions (as in Heilman’s (2011) approach) the experiments in Chapter 5 will highlight that the questions generated by the framework have similar psychometric values as questions generated by Heilman’s (ibid.) system and questions generated by human experts. Finally, from the third sentence, the framework would not generate a question at all, as this sentence contains an unresolved personal pronoun and questions with unresolved pronouns are not usable (as explained in Section 4.4.7)

As mentioned before, helper rules are run over the text first and the information gained in this step is then used by the question rules. The question rule QR_Loc (see Table 3) is used to generate questions about locations. It relies on the input of the helper rules HR_LocationGroup, HR_Adjunct, HR_Subj. The question rule first identifies question candidates in the source text and labels them as such; in this example, this applies to all sentences that satisfy the condition ($\{\text{SyntaxTreeNode.cat} == \text{"S"}, \text{SyntaxTreeNode.contains}\{\text{LocationGroup}\}\}$), i.e., all nodes of the parse tree that constitute a full sentence and the sentence contains a LocationGroup (with LocationGroup being one or more annotations of named entities of type Location preceded by certain prepositions, as specified in HR_LocationGroup). The sentences *Einstein was in Germany when he published his theory* and *Einstein taught in Germany and Switzerland* would both be identified as QuestionCandidates. For each QuestionCandidate, the rule then sets LocationGroup as Answer and extracts adjuncts (cf. Table 3). If an answer is contained in an adjunct, that adjunct is set as answer and all the remaining adjuncts are stored for question construction. Next, the subject is extracted using input from HR_Subj. From the remaining unlabelled parts, the verb phrase is extracted and decomposed into auxiliary and main predicate. Cases are treated differently depending on whether they contain certain auxiliaries (was, were), because these sentences require the insertion of ‘did’ in the question stem and a non-finite main verb. For example, *Einstein was in Germany when he published his theory* would turn into *Where was Einstein when he published his theory?*, while *Einstein taught in Germany and Switzerland* would turn into *Where did Einstein teach?*. All remaining unlabelled parts after this step are then labelled Predicate and the

question is then constructed by rearranging the labelled parts in the following order: Where
[did] +subject+mainVerb+predicate +adjunct(s)?.

Rule name	Description of action	GATE LHS	Example
QR_PersonSubj	<p>This rule looks for sentences in which a Person annotation occurs in (active and passive) sentences, with the person being the subject of the sentence. A <i>who</i> question is generated. Looks at the sentence constituents via the <code>SyntaxTreeNode.consists</code> feature and identifies the verb phrase. If the answer is contained within the verb phrase, it removes it from the verb phrase.</p> <p>Requires helper rules: HR_PersonGroup, HR_SyntaxPerson</p>	<pre>({SyntaxTreeNode.cat=="S", SyntaxTreeNode contains {PersonGroup.depType=="nsubj"} } {SyntaxTreeNode.cat=="S", SyntaxTreeNode contains {PersonGroup.depType=="nsubjpass"} }):sentence</pre>	<p>Abercrombie was commissioned by the British government to redesign Hong Kong → Who was commissioned by the British government to redesign Hong Kong?</p>
QR_PersonObj	<p>Where a Person annotation occurs in object position (direct or indirect object or object of a preposition), a <i>whom</i> question is generated by this rule.</p> <p>Requires helper rules: HR_PersonGroup, HR_SyntaxPerson, HR_possv</p>	<pre>(({SyntaxTreeNode.cat=="S", SyntaxTreeNode contains {PersonGroup.depType=="dobj"} }, !PersonGroup within {Possv}) {SyntaxTreeNode.cat=="S", SyntaxTreeNode contains {PersonGroup.depType=="iobj"} }, !PersonGroup within {Possv}) {SyntaxTreeNode.cat=="S", SyntaxTreeNode contains {PersonGroup.depType=="pobj"} }, !PersonGroup within {Possv})):sentence</pre>	<p>Bath's architectural renaissance began with the buildings designed by John Wood. → Bath's architectural renaissance began with the buildings designed by whom?</p>
QR_PersonPoss	<p>This rule looks for sentences in which a Person annotation is followed by a possessive marker. A <i>whose</i> question is generated. The rule has got no way of identifying the scope of the possessive. Therefore, if there are two persons in the subject, they will be grouped together within the answer. This may have the undesired effect of producing a wrong question in cases where the possessive only refers to the last person. However, this is also ambiguous when the source sentence is taken out of context, therefore it is not a drawback of the rule</p>	<pre>({SyntaxTreeNode.cat=="S", SyntaxTreeNode contains {Possv} }):sentPoss</pre>	<p>An exhibition of Abercrombie's plans for Bath was shown at the city's main art gallery. → An exhibition of whose plans for Bath was shown at the city's main art gallery?</p>

	generator. Requires helper rules: HR_PersonGroup, HR_SyntaxPerson, HR_possv		
QR_Temp	In sentences which contain certain types of Date annotation, a <i>when</i> question is generated. Valid Date types are those which express a specific point in time rather than a duration. Adjuncts are extracted. For each date, if it is contained in an adjunct, the adjunct is set as answer and all the other adjuncts are added to the question. If the date is not contained in the adjunct, the date is as the answer and all adjuncts added to the question. The question is constructed like this: When [did] +subject+mainVerb+” “+predicate +adjuncts? Requires helper rules HR_adj, HR_subj	({SyntaxTreeNode.cat=="S", SyntaxTreeNode contains {Date}}) :sent	In 1901, British troops fought a brutal war over the gold-rich territories of South Africa. → When did British troops fight a brutal war over the gold-rich territories of South Africa? <u>Exception:</u> The war continued for 3 years. → *When did the war continue?
QR_Loc	This rule looks for sentences which contain a Location annotation. A <i>where</i> question is generated. Not all types of Location annotations can be used in order to generate useful location question; location annotations that are not preceded by a preposition such as 'in', 'across' or an expression such as 'all over' are not suitable. Adjuncts are extracted. For each location, if it is contained in an adjunct, the adjunct is set as answer and all the other adjuncts are added to the question. If the location is not contained in the adjunct, the location is set as the answer and all adjuncts added to the question. If the Location is	({SyntaxTreeNode.cat=="S", SyntaxTreeNode contains {LocationGroup}}) :sent	Paul had been killed in Rome. → Where had Paul been killed? <u>Exception:</u> The battles have shaped how Britain looks today → *The battles have shaped how where looks today?

	<p>within a prepositional phrase that starts at the same offset as the Location, the whole prepositional phrase is included as the adjunct (e.g. in Birmingham's Town Hall). The questions are constructed like this: Where [did] +subject+mainVerb+" "+predicate +adjuncts?</p> <p>Requires helper rules: HR_LocationGroup, HR_adj, HR_subj</p>		
QR_What	<p>Generates question of the question type 'what'. Question candidates are sentences which contain a NP group which are not a person, a date or a location (these cases are already covered by the other rules). The tokens of the verb are scanned in order to identify the auxiliary verb and the main predicate. If the verb contains a <i>be</i> or <i>have</i> auxiliary, the question is constructed like this: <i>What +aux + subject+predicate?</i> Otherwise, it is constructed like this: <i>What did +subject+main_verb+predicate?</i></p> <p>Requires helper rules: Rule_adj, Rule_subj, Rule_NPGroup</p>	<pre>({SyntaxTreeNode.cat=="S", SyntaxTreeNode contains {NPGroup.depType=="dobj"}} {SyntaxTreeNode.cat=="S", SyntaxTreeNode contains {NPGroup.depType=="iobj"}} {SyntaxTreeNode.cat=="S", SyntaxTreeNode contains {NPGroup.depType=="pobj"}}):sentence</pre>	<p>What is the Staffordshire hoard?</p> <p>→ One of the most significant discoveries of Anglo-Saxon art ever made What does enrichment involve?</p> <p>→ It involves separating out the useful, lighter atoms from the less useful heavier atoms.</p>
HR_Subj	<p>Annotates the subject of a sentence. If a subject contains several tokens, those NPs that are direct children of a sentence root and are contained within a subject dependency (nsubj/nsubjpass) are labelled as subject.</p> <p>Requires: Dependency ,SyntaxTreeNode</p> <p>Annotation created: <Subject></p>	<pre>({SyntaxTreeNode.cat=="S", SyntaxTreeNode contains {SyntaxTreeNode.cat==NP}}):sentWithSubj</pre>	<p><i>The war</i> began in 1939. → When did <i>the war</i> begin?</p> <p><i>The Luftwaffe's bombing raids over Bath in April 1942</i> had wrecked 19,000 buildings. → When had <i>the Luftwaffe's bombing raids over Bath</i> wrecked 19,000 buildings?</p>

HR_Adjunct	<p>Labels PPs and ADVPs that are direct children of a sentence root as Adjunct. ADVPs that consist only of one determiner (DT) or adverb (RB) are excluded. Adjuncts can appear in initial, medial or final position. Adjuncts are detected by this rule and moved to final position when the question is formed. If several adjuncts occur, they are moved to the end of the question in the order they appear.</p> <p>Requires: SyntaxTreeNode (from StanfordParser)</p> <p>Annotation created: <Adjunct></p>	<pre>({SyntaxTreeNode.cat=="S", SyntaxTreeNode contains {SyntaxTreeNode.cat=="PP"} } {SyntaxTreeNode.cat=="S", SyntaxTreeNode contains {SyntaxTreeNode.cat=="ADJP"} } {SyntaxTreeNode.cat=="S", SyntaxTreeNode contains {SyntaxTreeNode.cat=="ADVP"} }) :sentWithAdjunct</pre>	<p><i>Very early on</i> → ADVP ADVP <i>in 1942</i> → PP <i>In Milan</i>, Mozart wrote the opera <i>Mitridate</i>, re di Ponto. → Who wrote the opera <i>Mitridate</i>, re di Ponto <i>in Milan</i>?</p> <p><u>Exceptions:</u> <i>It all</i> began in 1944. → DT Peter had <i>almost</i> died in the war. →RB</p>
HR_PersonGroup	<p>This helper rule locates Person annotations brings together Person annotations that are coordinated through commas or conjunctions. In such sentences, all Persons make up an answer. The rule also assigns a grammatical number marker to ensure auxiliaries are used in singular form in the question if they occur as plural in the source sentence.</p> <p>Requires: SyntaxTreeNode (from StanfordParser), Person</p> <p>Annotation created: <PersonGroup number="sg/pl" rule="Rule_PersonGroup"></p>	<pre>({Person} ((AND)[0,1]:and {Person})*):person_list</pre>	<p><i>Emily and Martha were</i> highly respected in political circles. → Who was highly respected in political circles? Answer: Emily and Martha</p>
HR_Poss	<p>The purpose of this helper rule is to find constructions of the type <i>PersonGroup</i> followed by a possessive marker and to label this as <i>Answer</i>.</p> <p>Requires: Dependency ,PersonGroup,Token</p> <p>Annotation created: <Possv>, <AnswerP></p>	<pre>({PersonGroup}):answerP ({Token within {Dependency.kind=="possessive"} })):possessive</pre>	<p><i>Mary and Emily's</i> house has a nice patio. → Whose house has a nice patio? Answer: Mary and Emily's</p>
HR_LocGroup	<p>Similar to HR_PersonGroup, this helper rule annotates sequences of Location entities coordinated by commas or</p>	<pre>((PREP)[0,1] {Location})</pre>	<p><i>In Cheltenham, Edinburgh and London</i>, people took to the streets in protest. → Where did people take to the streets in</p>

	<p>conjunctions.</p> <p>Requires: SyntaxTreeNode, Location</p> <p>Annotation created: <LocationGroup rule="Rule_LocationGroup"></p>	<p>((AND):and {Location})*):location_list</p>	<p>protest? Answer: Cheltenham, Edinburgh and London</p>
--	---	--	--

Table 3 Rules employed by the system - prefix QR denotes a question rule, HR denotes a helper rule

4.6 Extracting images to accompany questions

The Question Generation system has the unique ability to extract screenshots from the video documentary which relate to a question and can be supplied alongside it. It is hypothesized that these images can help focus test-takers' attention and thus increase their performance on a comprehension test. In Chapter 5, the results from a user study are described in which, amongst others, this hypothesis is tested. The findings show that supplying an image alongside a question does indeed aid test-takers' performance.

The methodology for obtaining screenshots is described as follows. After questions have been generated, the source sentence of a question (i.e. the sentence which gave rise to a question) is mapped back to the time stamp contained in the subtitles. Next, a screenshot is taken from the video at the respective time a source sentence occurs in the video. For example, in a documentary about nuclear fusion, the sentence "It was Mike Saltmarsh's task to work out whether the neutrons detected could indeed be from fusion or were simply background neutrons from the neutron generator" which occurred 29 minutes and 15 seconds into the video, gave rise to the first question and screenshot in Table 4 .


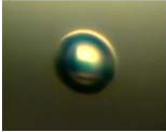

Question	Answer	Screenshot
Whose task was to work out whether the neutrons detected could indeed be from fusion or were simply background neutrons from the neutron generator?	Mike Saltmarsh	
What should be produced at exactly the same billionth of a second if fusion was happening?	Fusion neutrons	
What is nuclear fusion?	A nuclear reaction in which atoms are forced together until they fuse, giving off massive amounts of heat, light and energy.	

Table 4 Extracted screenshots for questions

At present, the screenshots are used to attract test-takers' attention, but there is a wide array of possibilities for using images. For example, if a multiple choice question (MCQ) format was used, it would be possible to use screenshots and images from other sources as distractors (similar but incorrect answers). Testing this scenario was out of the scope of this research project but might form part of future research.

4.7 Error Analysis 1: Human expert opinion

This Section presents an evaluation of a set of 258 questions generated by system WL.V. Two evaluators assessed the questions' usability in terms of grammaticality and overall usefulness according to a 4-point scale employed by Mitkov, Ha and Karamanis (2006). The goal of this small-scale evaluation was to estimate the percentage of usable questions generated by the framework, in order to be able to compare it to usability rates of other systems. In addition, based on the feedback gathered from the human experts, an error analysis was performed which allowed to draw conclusions on how the system output can be improved, since at the time of evaluation, the system was still in development. The error analysis also highlighted the limitations of the approach and which conclusions might be drawn from this for other QG systems which employ a similar methodology.

4.7.1 Methodology

Prior to this evaluation and error analysis, an internal, preliminary evaluation on a set of 90 questions was performed. This was done in order to get an impression of the quality of items the system produces as well as to make sure there were no complications when the final set of questions was given to the human expert evaluators. The results of this internal evaluation are described in Section 4.7.2.1 and compared with the new results in Section 4.7.2.2.

Two graduates of Linguistics and TESOL (Teaching English to Speakers of Other Languages) evaluated a total of 258 questions generated by the system. The questions were

generated using the subtitles of three documentaries on the topics of London, the history of Britain and Christianity, respectively²¹. The questions presented to the evaluators can be found in Appendix C.

The evaluators were asked to assign a score using a 4-point scale, categorising questions either as *unusable* or *usable*, with *usable* questions being further divided into *usable with major revision*, *usable with minor revision* or *usable without revision*. This evaluation scheme has also successfully been employed by Mitkov, Ha and Karamanis (2006), who have developed a state-of-the art system for the computer-aided generation of multiple choice questions (MCQs). The evaluators were given guidelines as to what rendered a question unusable or usable; for example, highly ungrammatical questions, those with obvious or no answers and those that were grammatical but were not informative should be classed as unusable, while minor revisions could be issues with punctuation and major revisions could be the deletion, addition or rearrangement of one or several words. The scores assigned by the evaluators for each question can be found in Appendix D. The evaluators classed 144 questions (55.8%) as *usable with or without revision*. Out of the *usable* questions, 81 (31.4% of the total) were *usable without revision*, 39 (15.1% of the total) were *usable with minor revision* and 24 (9.3% of the total) were *usable with major revision*. 114 (44.1%) questions were classed as unusable. Inter-annotator agreement was fair with a kappa value (Cohen, 1960) of $k=0.44$. The primary goal of this evaluation was not to assess the agreement between evaluators, but the fair agreement shows that even with two evaluators, the assessment of questions is still imperfect, highlighting that what

²¹ The documentaries and subtitles were downloaded from BBC iPlayer, BBC's online on-demand video player (<http://www.bbc.co.uk/iplayer/>)

constitutes a high-quality question may be a subjective decision with many variables to consider.

Generated Question	Score
Who hopes to create an identical modern survey of his own by flying exactly the same route the RAF did 60 years ago to create the first aerial surveys? Answer: Chris Going	Usable without revision
Who described the People’s Budget as pure socialism.. , an immensely wealthy landowner who’d also been Liberal prime minister? Answer: Lord Rosebery	Usable with minor revision
When be the port of London grown to the largest in the world? Answer: The late 1930s	Usable with major revision
Where were used for cultivating food in fact? Answer: In London	Unusable

Table 5 Examples of questions assessed by evaluators

4.7.2 Discussion

4.7.2.1 Internal evaluation

An internal evaluation of the system output, in which a set of 90 questions generated for a historical documentary had been assessed by me, gave a brief insight into the shortcomings of the proposed approach. In this evaluation, 42% of questions had been deemed usable. A number of improvements that needed to be implemented were identified, particularly with respect to using other levels of linguistic processing. Especially anaphora resolution needed to be explored, since a large proportion of questions (29%) had been deemed unusable for this reason. Unresolved anaphora, especially inter-sentential anaphora, i.e. cases, where the antecedent is situated in a sentence different from the one containing the anaphor (Mitkov, 2002) pose a problem to sentence-based QG systems like the one proposed, because the source sentence is taken out of context when generating the

question. For system WLV, resolution of pronominal anaphora is of particular importance. It is not possible to generate a usable *who*-question from the sentence *He graduated in 2009*, since theoretically the answer to this question would be “*he*” and this answer is meaningless. Resolving pronominal anaphora increases the number of question candidates from which meaningful questions may be generated.

While the types of errors discovered during the first experiments were by no means exhaustive, it quickly became apparent that the errors could be categorised into three super-classes. The first class of errors originated through the fact that the question rules were not refined enough. For example, in questions which require the insertion of auxiliary verbs, an error occurred because the tense of the main verb incorrectly remained unchanged (**When did it seemed nowhere was safe in London?*).

The second super-class consisted of errors caused due to the limitations of the NLP tools employed. Shortcomings of the named entity transduction component of GATE in combination with limited gazetteer lists meant that some questions were syntactically incorrect, because they had been assigned an incorrect annotation type. For example, *Queen in Houses in Queen’s Square were flattened*, is incorrectly labelled as a person, resulting in the unusable question *Whose Square were flattened?*

The final super-class consisted of errors caused by certain characteristics of the source texts. Certain sentences, for example, do not convey sufficient meaningful information to warrant a question to be classed as usable; using generic sentences, such as *Peter was shocked* is not optimal, because the question could hold true for several answers. This poses a particular problem for systems which generate MCQ systems (Skalban, 2009).

After the preliminary evaluation, a variety of measures were implemented to counteract the observed errors. For example, a module for pronoun resolution (as described in Chapter 4) was introduced, as well as a morphological analyser component which is used to retrieve the infinitive forms of main verbs in questions with auxiliaries. In addition, a number of helper rules were introduced to deal with repetitive tasks as well as to further alleviate certain errors.

4.7.2.2 Expert opinion

A first internal error analysis highlighted the need for additional processing resources. It is described how the system output has improved since the implementation of these resources. Despite the attempt to eradicate certain error types, the categorisation types into three super-classes still holds true. Table 6 shows the classes of error types and how many times errors based on these classes were observed in the dataset.

Error type cause	Observed
Limitations of transformational rules	58
Limitations of NLP tools	23
Characteristics of source text	17

Table 6 Error type classes and numbers of times errors were observed in the dataset

58 questions contained syntactical errors as a result of the transformational rule not being able to cope with a specific type of sentence or linguistic phenomenon and consequently failing. An example for such a highly ungrammatical question is **Who called made her way to a park in Manchester called Boggart Hole Clough a young woman on 15th July 1906?* generated from the source sentence *On 15th July 1906, a young woman called Adela Pankhurst made her way to a park in Manchester called Boggart Hole Clough.* In

this example, the error occurred due to the rule not being able to cope with the restrictive clause *a young woman called*. At other times, questions contain syntactical errors because a rule is applied to a complex or complex-compound sentence. For example, *And so, on the evening of July 30th 1909, Lloyd George decided he had no choice but to take his People's Budget directly to the people.* turns into the syntactically flawed question **When had decided he no choice but to take his People 's Budget directly to the people so?* Different degrees of severity of syntactical errors can be observed; while the two aforementioned examples are so ungrammatical that they are unusable, the apposition *Archaeologist Chris Going* causes a syntactical error in **Who has been documenting the changing face of London from the air for the last five years Archaeologist?* which does not render the question unusable since it only requires minimal revision.

Errors due to limitations of the NLP tools employed, in particular with respect to limitations of the named entity recognition and gazetteer lookup in GATE, occurred 23 times in the question set. While this error can in some cases be alleviated by adding the incorrectly annotated or unknown term to the gazetteer list, it is not possible to eradicate this error completely by populating the gazetteer lists, since it is not possible to foresee each and every case where the employed components will fail. In the question sets, unusable questions were generated, for example, when a location was incorrectly annotated as a Person; from *Edessa is special, because its ruler King Abgar set an important precedent here.* the system thus incorrectly generated **Who is special, because its ruler King Abgar set an important precedent here who?* An even more complex example of this error can be observed in *Christian Edessa has long since disappeared.* Christian in this case was intended as an adjective, Edessa again as location; however, the

tokens are wrongly annotated as a person. This example also shows, that the system would benefit from measures for word sense disambiguation (WSD).

17 questions were deemed unusable because they were generated from unsuitable, generic source clauses. Examples of such sentences include *The grid forms we see in New York* (*Where did the grid forms we see?) and *Further south was Alexandria in Egypt* (*Where was further South Alexandria?). At the moment, it is not possible for the system to avoid generic sentences as the question rule fires once a specific pattern in a source sentence is matched, based on named entities in the source sentence. In *The grid forms we see in New York* the location rule is fired due to the location contained in the sentence.

4.7.3 Conclusion

A preliminary, internal error analysis of questions produced by the framework developed as part of this thesis had shown that 42% of the assessed questions were usable, but also highlighted that a number of improvements needed to be implemented, particularly with respect to using other levels of linguistic processing, such as pronoun resolution. This preliminary evaluation was performed in order to make sure no complications would arise when another set of questions was given to two human expert evaluators and also to get a general idea of the quality of questions produced by the framework. The analysis had shown that errors can be grouped into 3 super-classes: errors due to limitations of the transformational rules, errors caused by limitations of the NLP tools employed and finally errors which occur due to certain characteristics of the source text.

After a second evaluation performed by two human evaluators, a usability rate of 55.8% was reported. An error analysis based on the evaluators' feedback showed that despite the larger proportion of usable questions produced, errors could still be observed in the output and they also still corresponded to the three super classes identified previously. This observation might imply that any QG system which employs such a methodology might be prone to such errors.

The error analyses and evaluation described were by no means exhaustive, but they allowed to gain insight into the challenges and limitations of such an approach. While some suggestions for improvement were made, there are a number of issues that have not yet been fully understood and that need to be researched, e.g. the filtering out of generic sentences which render questions unusable, issues surrounding anaphora resolution and how the use of other NLP resources can affect the QG process.

Since the current approach is an over-generation approach (one source sentence can give rise to several questions), ways of ranking the generated question need to be investigated, like other researchers, such as Heilman and Smith (2010), have done.

At the centre of this research still remains one question which will always be of utmost importance: what constitutes a good question and how can this be measured? Research suggests that subtasks and evaluation in QG are application dependent. While, as pointed out in Chapter 5, question quality can be measured with a number of internal and external metrics, it is hypothesized that the most important criterion for a high quality system-generated question should be indistinguishability from human generated questions in order for a QG system to be of real use to the end user. To this end, in the experiments described in Chapters 5 and 6, several psychometric measures of system-generated questions in

comparison to human-created ones are computed and the operations underlying the post-editing process are analysed.

4.8 Error Analysis 2: Performance of transformational rules

The previous Section, described a small-scale evaluation which had been performed during the development and improvement phase of the system to estimate the percentage of usable questions produced. A subsequent error analysis helped to identify super-classes of errors observed in the generated questions. The purpose of this error analysis is to compare the quality of the output of two QG frameworks in terms of overall usability, analyse the performance of the transformational rules employed (i.e. how many questions per transformational rule are usable) and to provide a more fine-grained classification of errors. The first framework (system WLW) is the framework developed as part of this thesis for generating questions from video documentaries. The second system (system CMU), proposed by Heilmann (2011), is a state-of-the-art system which generates factual questions (wh-questions) from informational texts (see Chapter 2 for a more detailed description of this system). System CMU also employs a rule-based approach and several readily-available NLP tools to generate questions. Both systems employ an ‘overgeneration approach’ meaning that one source sentence in the text can give rise to several questions. In addition, system CMU applies statistical ranking methods after the question generation process in order to separate high-quality questions from lower quality ones. Since both systems employ a similar methodology to generate questions and since the types of questions generated are largely the same, this evaluation is performed in order to compare the quality of questions generated by the own system, namely system WLW, to that of system CMU.

4.8.1 Methodology

A source text taken from the subtitles about the history of Britain, about 500 words in length, was uploaded to both systems and questions were generated based on this source text. A total of 36 questions were generated by system WLV, while system CMU output 260 questions in total. In order to fairly assess the quality of the questions generated, only those questions were chosen for evaluation which were based on source sentences used in both systems. In other words, if a source sentence gave rise to questions in system WLV, then all those questions generated from system CMU based on this sentence were used for evaluation. This meant that out of system CMU's 260 questions, 156 questions were selected for this evaluation.

In the analysis, there is a distinction between “usable” and “unusable” questions using a 4-score scale. Questions which were usable without the need for post-editing were scored 1, questions which required minor changes before being deemed usable were scored a 2, questions which needed major alterations before being classed usable were scored 3 and unusable questions were assigned a score of 4. Just like in the first error analysis, minor revisions could be issues with punctuation and major revisions could be the deletion, addition or rearrangement of one or several words. Table 7 shows results for usable and unusable questions for both systems broken down per question type. While system CMU generates a higher number of questions in total, the proportion of usable questions is similar for both systems; for system WLV, 22 (61%) out of 36 questions are deemed usable, while 105 (67%) out of 156 of system CMU's output were classed as usable. In

system CMU, a total of 71 errors were observed and in system WLW a total of 26 errors and it is important to note here that one question could contain multiple errors.

Question type	Usable		Unusable		Total per type	
	WLV	CMU	WLV	CMU	WLV	CMU
Who	1	11	0	7	1	18
Whose	2	0	0	4	2	4
Whom	1	0	1	0	2	0
What	15	43	8	29	23	72
When	2	11	4	2	6	13
Where	1	2	1	1	2	3
Yes/No	0	38	0	8	0	46
Total	(61%) 22	(67%) 105	(39%) 14	(33%) 51	36	156

Table 7 Usable and unusable questions generated by the two systems broken down by question types (all question types)

The question types generated by both systems are mainly the same; however, system WLV does not generate ‘yes/no’ questions and system CMU does not generate ‘whom’ questions. For this reason, in Table 8, only the results for shared question types, i.e. question types that both systems are able to generate, are considered. As only two questions of system WLV are affected, the usability rate remains at 61% for system WLV. However, system CMU’s question usability rate is reduced to 60%. This is because a large number (46) of CMU’s set of selected questions consists of yes/no questions. These questions are often classed as usable simply because they are syntactically correct. However, unlike for wh-questions, for yes/no questions no answer phrase gets extracted from the sentence and the answer needs to be inferred, which may not always be possible.

Question type	Usable		Unusable		Total per type	
	WLV	CMU	WLV	CMU	WLV	CMU
Who	1	11	0	7	1	18
Whose	2	0	0	4	2	4
What	15	43	8	29	23	72
When	2	11	4	2	6	13
Where	1	2	1	1	2	3
Total	(61%) 21	(60%) 67	(39%) 13	(40%) 43	34	110

Table 8 Usable and unusable questions generated by the two systems broken down by question types (shared question types)

In general, the usefulness of yes/no questions to assess knowledge is disputable, at least here, because these questions are only ever generated from affirmative declarative sentences, thus the answer would always be ‘yes’. For example, from the source sentence: “*Einstein was born in 1879*”, the generated question would be “*Was Einstein born in 1879*” and the inferred answer would be ‘yes’. If such questions were used in knowledge assessment, test participants would quickly realise that the answer to the yes/no questions is always ‘yes’. In order to generate questions to which the answer would be no, questions would have to be generated from source sentences which contain negations. For example, from the sentence: “*Buzz Aldrin was not the first to set foot on the moon.*” a QG system could generate the question “*Was Buzz Aldrin the first to set foot on the moon?*” with the (inferred) answer being ‘no’. However, (factual) statements are typically affirmative; negation is marked and used to signal something unusual or an exception, and thus it could be difficult to find candidate sentences for such questions. In addition, negation can pose difficulties in NLP and is a complex topic in its own right, which will not be further

discussed here. It is, however, important to mention that in CMU’s system yes/no-questions are ranked lower than wh-questions in the ranking system, which are preferred (i.e. ranked higher).

4.8.2 Error analysis

In addition to the 4-score scale used to assess the quality of the generated questions, an error analysis was also performed to examine why questions were deemed unusable or which sort of post-editing they required. The error types observed in system CMU’s questions largely agree with those described in Heilmann (2011), even though not all error types were observed (due to the limited sample set). The error types observed can be found in Table 9. Just like in the previous experiment, the error types observed can also be classified into super-classes of errors.

Error type	Description	Example
Wh-word error	An incorrect wh-word was chosen.	“ When began with the buildings designed by John Wood?” → “what” should be used as wh-word
Pronoun resolution error (type 1)	A pronoun was not resolved	“It was built during a golden age for British architecture.”
Pronoun resolution error (type 2)	A pronoun was incorrectly resolved	Source: “His ambition was nothing less than to revive the splendour of ancient Roman cities.” Question: “Whose ambition was nothing less than to revive the splendour of ancient Roman cities?” Extracted answer: “Jane Austen’s” (not antecedent of “his”)
Syntax error	The resulting question is ungrammatical	Source: “And so, on the evening of July 30th 1909, Lloyd George decided he had no choice but to take his People’s Budget directly to the people” Question: “When had decided he no choice but to take his People’s Budget directly to the people so?” Answer: “On the evening of July 30th 1909”
Answer error (type 1)	Answer phrase is generic/irrelevant	Question: “What was the English class system set in in Bath?” Answer: “in stone”

Answer error (type 2)	Wrong answer phrase for generated question	Source: “Since World War Two, Britain’s towns and landscapes have been transformed.” Question: “What have Britain’s towns and landscapes been transformed since World War Two?” Answer: “World War Two”
Formatting error	The generated question contains typographical errors	“What did other architects continue, creating a city of unmatched glory, CHURCHBELL RINGS order and scale?”
Decompositional error	Error which occurs when breaking down of MWEs, or other complex units of meaning fails	“And, , whose son put on the finishing touch - The Royal Crescent after his death in 1754?” from “And, after his death in 1754, John Wood’s son put on the finishing touch - The Royal Crescent.”

Table 9 Observed error types in the output of 2 Question Generation frameworks

4.8.2.1 Pronoun resolution errors

The main reason for questions from the CMU set being deemed unusable stemmed from pronoun resolution errors. In total, 23 questions (15% of all questions, 57% of unusable questions) contained this error type and 22 questions were deemed unusable because of it. Pronoun resolution in system CMU is performed using the ARKref system (developed by Brendan O’Connor and Michael Heilman, 2013). ARKref is a tool for noun phrase co-reference resolution which relies on syntactic information from the Stanford Parser (Klein and Manning, 2003) and semantic information from an entity recognition component (supersense tagger) to identify a set of antecedent candidates for a given mention. The candidate with the shortest tree distance from the target is selected as the antecedent.

Two different types of pronoun resolution errors have been observed. Pronoun resolution error type 1 was classed as those cases where a pronoun was not resolved and pronoun resolution error type 2 as those cases, where a wrong antecedent was identified for a given pronoun. Error type 1 occurred 10 times in the sample question set and in each case

rendered the question unusable. From the sentence: *“It was built during a golden age for British architecture.”* system CMU generated the question *“When was it built?”* In this case, it is not possible to identify what “it” is referring to. Unresolved anaphora can lead to unusable questions in sentence-to-question generation systems, because each question is based on one source sentence, without regard to its context. If the source sentence contains an unresolved anaphor, so will the question that is generated from it and even as a human it is often not possible to identify an antecedent for a pronoun in such cases.

This error type can affect the extracted answer phrase. From the sentence: *“And, after his death in 1754, John Wood’s son put on the finishing touch - The Royal Crescent.”* system CMU generated the question *“When did John Wood’s son put on the finishing touch- The Royal Crescent?”* with the answer phrase being “after his death in 1754”. Thus, at first inspection, the question itself appears to be without fault, however, when examining the answer phrase it becomes clear that the answer phrase is ambiguous. “His” refers to “John Wood” in this case, but without knowing the context of this source sentence, this cannot be said for sure. However, in this case, the question can easily be edited to make it usable, by reducing the answer phrase to “in 1754”.

Pronoun resolution error type 2, i.e. incorrect resolution of pronouns, affected 13 questions. From the source sentence *“His ambition was nothing less than to revive the splendour of ancient Roman cities.”* system CMU generated *“Whose ambition was nothing less than to revive the splendor of ancient Roman cities?”* and as answer phrase “Jane Austen’s ambition” was extracted. While the question is syntactically sound, the extracted answer phrase renders the question unusable. Even without looking at the source

sentence in context, it is clear that “Jane Austen” is not the correct antecedent for the personal pronoun “his” in the source sentence.

From analysing only the system output, it is not possible to directly identify why pronoun resolution error types 1 and 2 occur or which steps needed to be undertaken in order to prevent them from occurring.

In system WL_V, pronoun resolution is performed using the co-reference module supplied by GATE as well as by using several JAPE rules (see Section 4.4.7). In all simple and complex sentences contained in the text, first-mention pronouns are replaced with the longest co-referent in the co-reference chain. In independent clauses in compound sentences, not only first-mention pronouns, but all subject personal pronouns will be replaced with their co-referents, as these will afterwards be split into several shorter sentences.

In system WL_V's output questions, only pronoun error resolution type 1 was observed 6 times. For example, the system-generated the question: “*What was it to the Georgians?*”. This question is unusable, because it is not possible to identify what the antecedent of the pronoun “it” is.

The current implementation of system WL_V only resolves male and female personal pronouns, as these typically have a person's name as antecedent and are in many cases easy to resolve; however, in the future more sophisticated ways of resolving anaphora need to be explored.

4.8.2.2 Wh-word error

In the question sets of both systems, some questions were formed using an unsuitable wh-word. For example, from the sentence: *“The modern world has brought radical change and progress”* system CMU generated *“When has brought radical change and progress?”* when the correct wh-word should have been “what”. In all cases observed in this error analysis, it is possible to generate a usable question by changing the wh-word. This error type occurred accounted for 11.27% of errors in system CMU’s example set and 7.69% in system WLV.

4.8.2.3 Syntax error

Sometimes generated questions are ungrammatical. For example, from the sentence: *“And so, on the evening of July 30th 1909, Lloyd George decided he had no choice but to take his People’s Budget directly to the people”* the question *“When had decided he no choice but to take his People ‘s Budget directly to the people so?”* In this case, the resulting question is ungrammatical because the source sentence is complex and the transformational rule cannot accommodate complex sentence structures like these. This error was observed 12 times in the WLV dataset and 18 times in the CMU dataset.

4.8.2.4 Answer phrase error

Two types of answer error have been observed. In answer error type 1, the extracted answer is generic or irrelevant. For example, the answer to the question: *“What was the English class system set in in Bath?”* is “in stone”, which obviously does not make for a good question. In answer error type 2, based on the source sentence, the wrong answer is

extracted. For example, from the source sentence: “*Since World War Two, Britain’s towns and landscapes have been transformed.*” the question: “*What have Britain’s towns and landscapes been transformed since World War Two?*” was generated, alongside the extracted answer “*World War Two*”.

4.8.2.5 Formatting error

Questions which contain formatting errors are affected by certain typographical issues. Even though not observed in system CMU’s sample question set, Heilmann (2011) describes this error type in his work and explains that the error type occurs due to certain characteristics of the source text, e.g. the QG system might sometimes generate questions from captions which may lead to ungrammatical questions. While this error type could strictly be regarded as a parsing error, Heilmann created a specific error category for this type of error as it can often be overcome by implementing simple filtering rules. System WLV has already got implementations of such filtering rules specifically geared towards the characteristics of subtitles; subtitles often contain additional information for the hard of hearing, for example, “DOG BARKING”. This information is typically written in capital letters and can thus easily be filtered out. Before filtering, this information leads to ungrammatical questions, such as:

“What did other architects continue , creating a city of unmatched glory , CHURCHBELL RINGS order and scale?”

4.8.2.6 Decompositional error

For the question generation process, a source sentence needs to be broken down into constituents, so that it can later be rearranged to form a question. Decomposition errors occur where the breaking down of the source sentence into smaller constituents fails. A decompositional error was observed once in system WL_V's sample questions: “*And, , whose son put on the finishing touch - The Royal Crescent after his death in 1754?*” This question was generated from the source sentence: “*And, after his death in 1754, John Wood's son put on the finishing touch - The Royal Crescent.*” On the surface, it appears that this error is simply a formatting error; however, when examining the source clause, it becomes apparent that this error occurred when breaking down the sentence into its constituents.

4.8.3 Conclusion

In this experiment, the performance of the transformational rules of the Question Generation framework were analysed, i.e. how many questions per question type can be deemed as ‘usable’. An error analysis with a fine-grained classification of errors was performed and compared the results not only for system WL_V, but also for that of the state-of-the-art system proposed by Heilman (2011). The findings show that there is a similar usability rate for the questions generated by both frameworks and that the observed error types occur in both datasets. Generally, the categorisation of errors is difficult and far from straightforward, because questions can contain multiple errors and often it is difficult to assign one error to one specific category.

4.9 Summary

In this Chapter, the system for automatic Question Generation from documentary videos which was developed as part of this thesis was described. The GATE architecture which forms the basis of the system was introduced and it was explained why employing GATE has several advantages. GATE Developer, the graphical development environment of GATE, allowed to easily create a prototype system, which was then transformed into a bespoke QG system using GATE Embedded, the Java class library.

The methodology used in order to generate factual questions from documentary video subtitles was described. After formatting the subtitle files in a way in which they can be used by the system, several linguistic pre-processing steps are performed; these include tokenisation, sentence splitting, POS tagging, syntactic parsing, NER, morphological analysis and pronoun resolution. Following the pre-processing, a set of manually crafted transformational rules is used to identify question candidates in the text and to transform them into questions. A feature unique to the system was also described; namely the ability to make use of an extra layer of information – the visual layer of the documentary. The system can provide a screenshot taken from the video alongside a question. This can help test takers' performance on a comprehension test by focussing their attention, as the experiment in Chapter 5 highlights. Due to the modularity of GATE and the structure of the framework, it is very easy to make adaptations to the system itself and its rules. Users can easily change the processing resources employed and as the transformational question rules are not hard-coded, they can easily be adjusted or enhanced.

This Chapter also described two error analyses which were performed to evaluate the proposed Question Generation framework and to improve the generated questions based on the findings made. In the first experiment, two evaluators judged a set of questions generated by the system and based on the judgements an error analysis which highlights challenges in QG, which can also be applicable to other types of QG systems, was performed. The second experiment was an analysis of the performance of each of the transformational rules; it was analysed how many questions generated per rule could be deemed usable and the different error types observed were analysed in a fine-grained scheme. The results were compared to a state-of-the art system. In the next Chapter, evaluation in Question Generation will be discussed.

CHAPTER 5: EVALUATION 1 - PREQUESTIONS AND PSYCHOMETRIC PARAMETERS

This Chapter is divided into two Sections. In the first Section (5.1) general notions about evaluation in Question Generation are described. First, three aspects of evaluation in Question Generation, namely key concept identification (Section 5.1.1), question type determination (Section 5.1.2) and question realisation (Section 5.1.3) are explained. Shared evaluation tasks are discussed in Section 5.1.4. The second part of this Chapter (Section 5.2) describes a user study which was performed using a novel evaluation methodology to investigate a large number of research questions; amongst others the feasibility of using system-generated questions as so-called ‘pre-questions’ is discussed and the psychometric parameters of system and manually generated questions is analysed.

5.1 Evaluation: QG as a three-step-process

Evaluation in QG may not be as straightforward as in other disciplines of NLP. The reason for this is that evaluation in QG is application-dependent; since there are different application contexts in which the generated questions serve different goals, evaluation criteria need to be matched to these goals in order to determine a QG system’s performance.

For many applications of QG and especially those concerned with dialogue, the Question Generation procedure can be regarded as a three-step-process, consisting of *key concept identification*, i.e. the identification of concepts in the source text from which questions will be generated, *question type determination*, i.e. the selection of the most suitable

question type given the source text and a target concept and *question realisation*, i.e. the creation of the surface form of the questions based on the prior steps (Nielsen, 2008). As each of these subtasks have different objectives, it is necessary to evaluate them in ways appropriate for each task.

In the proposed framework,

5.1.1 Evaluation of the key concept identification task

A premise for a high-quality QG system is the formation of questions based on *key concepts*, i.e. spans or snippets of source text which carry vital information (Mitkov and Ha, 2003). The objective of the key concept identification task is to output such key concepts in the form of annotations in the source text (Nielsen et al., 2008). While key concepts may vary depending on context of the application, a high quality QG system is characterised by high precision (it should only find key concepts, not false positives) and recall (it should find all key concepts) when identifying key concepts (Graesser, Rus and Cai, 2008). In theory, if the evaluation of the *key concept identification* task in QG was to be performed automatically using the standard F-score (Van Rijsbergen, 1979), this would imply that key concepts can be dichotomised. However, it is problematic to class concepts as either *key* or *non-key*, since, in practice, the importance of concepts can vary along a scale (Nielsen et al. 2008) and certain concepts that are important in one context might not be regarded as significant as in another. For example, in a source text about morphology, ‘affix’ and ‘language’ are important concepts, while the first could be thought of as more

important in the context of morphology. At the same time, in a source text about general linguistics, the term ‘language’ might be regarded as more important than ‘affix’.

Human evaluation in the key concept identification task would provide a gold standard and would allow drawing conclusions about the reliability of human annotation using inter-annotator agreement (ibid.). While accurate, this process would be costly and time extensive. For this reason, Nielsen et al. (ibid.) have suggested a modified version of the F-measure developed by Lin and Demnher-Fushmann for question answering (2005, in Nielsen et al. 2008). Here, it is assumed that at least two domain experts annotate the key concepts in a set of test documents, considering *vital snippets* from which a high quality QG system should generate questions as well as *optional snippets* which are concepts deemed to be reasonable for a QG system to identify (Nielsen et al. 2008). A good QG system should be able to identify only vital and optional snippets, without annotating any false-positives, i.e. concepts that are neither vital nor optional (ibid.).

In the proposed F-measure, each vital concept is weighed equally, independent of its length. Recall constitutes the coverage of vital snippets and precision constitutes the extent to which a system tagged snippet overlaps with a single human annotated snippet, vital or optional (ibid.). The human judgments could either be measured in a binary way (i.e., identified versus not identified), or as ratings of the extent to which a concept was covered by the system. There is no suggestion as to how the extent of the overlap could be measured exactly, but measures for semantic similarity or n-gram could be used for this purpose. While the binary decision method would be easier and cheaper to perform, the overlap measure could provide more sensitive measurements of the performance of a QG system.

The scores also take *facets* into consideration, which are any fine-grained component of the semantics of an utterance or text but which can be used for other underlying units of meaning, for example, syntactic relations from a dependency parse (Bethard et al., 2007, in Nielsen et al. 2008). The F-measure can be calculated as follows (ibid):

Let k be the number of vital snippets, m be the total number of annotated snippets across all human annotators, n be the total number of system-tagged snippets, V_i , A_i , and S_i be the set of semantic facets in the vital, human-annotated (vital and optional), and system-tagged snippets, respectively. The metric calculates the Instance Recall (IR) for each vital snippet and Instance Precision (IP) for each system-tagged snippet as:

$$IP_j = \frac{\max_{i=1..m} |S_j \cap A_i|}{|S_j|}$$

$$IR_i = \frac{\max_{j=1..n} |V_i \cap S_j|}{|V_i|}$$

Make the overall recall and precision equal the average instance recall and precision and calculate the F-measure as usual, where β assigns a relative importance to precision and recall:

$$P = 1/n \sum_{j=1}^n IP_j$$

$$R = 1/k \sum_{i=1}^k IR_i$$

$$F_{\beta} = \frac{(1 + \beta^2)P \cdot R}{(\beta^2P + R)}$$

5.1.2 Evaluation of the question type determination task

Question type determination, the selection of the most suitable question type given the source text and a target concept, can be a subjective process as the selection of question types may depend on pedagogical theories, dialogue context and user models, but in the course of the QG STEC, question types have been regarded independently of context (Nielsen et al. 2008). The proposed evaluation method for this task is for annotators to enumerate the types of questions that would be suitable given a certain target concept. After adjudication and labelling of *vital* and *optional question types*, these lists would provide a gold-standard of question types. A system's performance could then be adjudged using F-measure; the number of vital questions types output by a QG system in relation to all vital question types for that target concept shows recall, while all of the question types selected by the system that were identified as either vital or optional in the gold-standard annotation show the precision of a QG system (Nielsen et al. 2008).

5.1.3 Evaluation of the question realisation task

The *question realisation* task is concerned with finding a suitable surface form for a question to be generated based on the input of the *key concept identification* and the *question determination task*. In order to fairly evaluate only the output of this subtask, the

input data (target concept and question type) should be provided with gold standard annotations. However, the evaluation of the system output as a whole would be performed application dependent (ibid.), since different questions serve different purposes. For example, in educational environments, questions may be used to assess students' knowledge of a subject matter and thus a measure that can be investigated is discriminating power (Lin and Miller, 2005, in Nielsen, 2008). For an intelligent tutoring system, learning gains may be evaluated, since the objective of such systems is to maximise learning gains (Nielsen, 2008). When dealing with MCQs in particular, measures from classic test theory, such as distractor usefulness and item difficulty (Isaacs, 1994), may be analysed, or the efficiency of the system in terms of quality of questions vs. time taken to produce them (Mitkov, Ha and Karamanis, 2006).

Intrinsic measures also provide valuable insight into the quality of a QG system. Such measures include grammaticality, use of anaphora, clarity, interestingness and others (Nielsen et al. 2008). A form of evaluation for this subtask is proposed (ibid.) in which gold standard questions generated by human experts are compared to system-generated ones. This process would involve analysis of n-gram overlaps, a technique which is commonly used in the areas of machine translation and automatic summarisation. In addition, Nielsen et al. (ibid) propose to utilise *facets* as was suggested for the key concept identification task; with the help of facets at bigram level, it would be possible to judge the degree to which a system-generated question is a paraphrase of a gold standard question.

These techniques, discussed at the QG STEC, are one approach to evaluation in question generation. In the case of this research, some of the basic assumptions underlying the

design of the proposed framework necessitated an alternative approach. Regarding the identification of key concepts, the proposed framework generates questions based on Named Entities in the text and transformational rules which identify whether a source sentence fulfils the criteria for being a candidate sentence (for example, to generate a ‘who’ question, the source sentence needs to contain a Named Entity of the person and that person needs to be in subject position). The framework has been designed based on the assumption that Named Entities of specific types are key concepts in the documentaries. This was not explicitly tested, but it is a reasonable assumption that, even if a human annotator would not annotate, for example, all persons in the text as vital, they would still annotate them as optional snippets. This in turn would mean that the system would score high on precision in the question realisation task. On determining question type, again due to the nature of the chosen approach, the method proposed in QG STEG is of limited value. The framework generates a question based on criteria that need to be fulfilled in the source sentence, so certain question types can only occur if certain conditions are satisfied in the source sentence. So while the framework can generate different questions types from one sentence, the evaluation method proposed in QG STEC only evaluates vital and optional question types. Again, if not classed as vital, the question types by the framework would most likely be classed as optional. In addition, the findings from the error analysis in Section 4.8.2.2 support this assumption, as the amount of questions analysed in which a wh-word error occurred was not significant.

For evaluating the framework, it is more important to focus on quality in the question realisation task. The proposed evaluation method, is used in an experiment described in the second part (5.2) of this Chapter and a further experiment in Chapter 6. The results show

that the questions produced by the framework have psychometric values comparable to questions generated by human experts and a state-of-the-art system, and that generating questions with the framework and post-editing them is faster than manual creation. Thus the results of the evaluation performed shows that the framework performs well on quality and effectiveness.

5.1.4 QG shared evaluation tasks

The identification and definition of shared tasks can be helpful in promoting research advancements. At the Question Generation Shared Task and Evaluation Challenge (QG STEC) (Silveira, 2008) four Question Generation tasks were defined.

In the *Text-to-Question* task, the aim of a Question Generation system is to exhaustively generate a set of text-question pairs given a source text. Formally, this can be summarised as follows: “Given a Text T, create n Text-Questions pairs, each represented as a (K_i, Q_i) pair, where K_i , the target text, indicates which text segment from T represents the answer and the Q_i represents a question that would elicit K_i ”(ibid.).

In the *Tutorial Dialogue task*, the QG system takes as input a tutorial dialogue history and a target set of propositions. Based on these, the aim of the QG system is to generate questions which, if put to a student, would be answered by the student in such a way that the specified propositions are contained in the answer (ibid). Formally this can be summarised as: ‘Given a tutorial dialogue history H and a set of expected propositions P to be covered, create Question Q such that if answered by the student, it would induce the student to state P in the context of H’ (ibid.). To illustrate this, in a teacher-student

dialogue about morphology, a proposition could be *a morpheme is the smallest meaningful component of a word*. The QG system is expected to generate questions which would prompt the student to state the proposition. For example, such a question could be “What is a morpheme?” or “What is the smallest meaningful component of a word?”

In the *Assessment* task, a QG system takes as input a text and, optionally, a dialogue. The objective would be to select an important concept in the source text, determine a suitable question type, and to generate a textual question which can be used for assessment (ibid).

In the *Query-to-Question* task, a QG system takes as input a query consisting of keywords with the aim to convert the query into a canonical form of a natural language question (ibid.). For example, given the keywords *best, Italian, restaurant, Wolverhampton* a QG system could generate the question *Where is the best Italian restaurant in Wolverhampton?*

The framework has undergone several cycles of development, evaluation and improvement. Two error analyses were performed to estimate the percentage of usable questions generated by the framework (see Chapter 4). In the first error analysis, a set of 258 questions was evaluated by human experts in terms of grammaticality and overall usefulness. A 4-point scale was used, categorising questions either as *unusable* or *usable*, with *usable* questions being further divided into *usable with major revision*, *usable with minor revision* or *usable without revision*. In the second error analysis, the performance (percentage of usable questions) of each of the transformational rules employed in the framework was analysed and compared the results to those of a state-of-the-art system. A fine-grained classifications of errors that can be observed in system-generated questions was presented. In Chapters 5 and 6, two experiments are described. In the first experiment,

a user study is outlined which served not only to investigate the psychometric parameters of the questions generated by system WLIV compared to those generated by a state-of-the-art system and manually generated questions, but also to investigate the feasibility of using two types of so-called ‘pre-questions’ (text-based and image based). In the second experiment, the efficiency of post-editing system-generated questions versus the creation of questions manually by human experts was examined. A number of usability statistics and qualitative feedback was gathered and the edit operations that post-editors perform when post-editing questions were analysed.

5.2 Experiment 1

In this Section, an experiment is described designed to investigate whether questions generated automatically by two NLP-based systems (one developed by myself, the other the state-of-the-art system developed by Heilman, 2011) can successfully be used to assist multimedia-based learning. The feasibility of using a QG system’s output as ‘pre-questions’ is examined, with different types of pre-questions used: text-based and with images. The psychometric parameters of the automatically generated questions by the two systems and of those generated manually are compared. Specifically, the effect such pre-questions have on test-takers’ performance on a comprehension test about a scientific video documentary is analysed. The discrimination power of the questions generated automatically is compared to that of questions generated manually. The results indicate that the presence of pre-questions (preferably with images) improves the performance of test-takers. They indicate that the psychometric parameters of the questions generated by

system WLV are comparable to if not better than those of the state-of-the-art system. Qualitative feedback is gathered from some of the test-takers about the questions and the quiz.

5.2.1 Background

Research in education (Hamilton, 1985; Klauer, 1984; Rothkopf, 1982; Hamaker, 1986; Anderson & Biddle, 1975) has shown that *pre-questions, i.e. questions which are supplied to test-takers before receiving learning material*, can have beneficial effects on student learning in reading activities. Pre-questions can help focus learners' attention on the learning material targeted by the questions and they also increase the learning effect through repetition (Thalheimer, 2003). The manual creation of questions is time-consuming and requires the knowledge of domain experts. Research in Natural Language Processing indicates that systems for Question Generation can assist teachers in this laborious task, thus saving time and resources. Semi-automatic QG systems can produce test questions up to 4 times faster than a human expert, without compromising quality (Mitkov, Ha and Karamanis, 2006). This experiment examined whether the questions produced by the framework can successfully be used as pre-questions and thus support creators of assessment materials. Two different types of pre-questions are investigated: text-based and with supporting image. This experiment also serves to test whether pre-questions have a beneficial effect in combination with audio-visual learning material as opposed to reading material; the effect pre-questions have on test-takers' performance on a comprehension test about a scientific video documentary is analysed. It is also examined

whether or not questions generated automatically by system WL_V and the state-of-the-art system developed by Heilman (2011) have the same psychometric parameters as those generated manually by human experts. The psychometric parameters of questions, such as their discrimination power, are among the most important measures of the quality of the questions.

5.2.2 Methodology

This Section describes the experimental design and procedures, as well as definitions and research questions.

5.2.2.1 Definitions

Pre-questions are supplied to test-takers before receiving learning material (here: the documentary video). Pre-questions are non-scoring and do not require an answer. Pre-questions can be text-only or can be accompanied by a relevant image. In this experiment, images are screenshots extracted from the video.

Post-questions are presented to the test-takers after receiving learning material (here: after watching a documentary). Post-questions are generated either manually by a human expert or automatically. The post-questions employed in this experiment are short answer style questions.

System A is the QG framework designed by the author, as described in Chapter 4. *System B* is the QG system developed by Heilman (2011). Its methodology is explained in the literature review in Chapter 2.

5.2.2.2 Research questions

The aim of the experiment is to answer the following research questions:

1. a) Whether the presence of text-based pre-questions helps test-takers to answer post-questions more accurately (i.e. more questions are answered correctly).
- b) Whether the presence of pre-questions *with screenshots extracted from the video* helps the test-takers to answer post-questions more accurately.
2. a) Whether the presence of text-based pre-questions affects the time taken to answer post-questions.
- b) Whether the presence of pre-questions *with screenshots extracted from the video* affects the time taken to answer post-questions.
3. What are the psychometric parameters of questions generated by system A when compared to system B and manually generated questions?

5.2.2.3 Selection of system-generated post-questions

Due to the nature of their QG approach, both QG systems produced more questions (A: 139, B: 567) than required for the experiment. Only 9 questions were needed from each method for the participants to complete the experiment in approximately one hour. As system B uses certain heuristics to output questions ranked in terms of quality, the top 3 questions corresponding to the respective parts of the video were selected for use in the experiment. A human expert, a high school teacher of English, watched the documentary and was instructed to select the best 3 questions per part from system A's pool of questions. It was decided that this was a better approach than random selection, as system

B's questions are automatically ranked in terms of quality and so also only its best questions are used.

5.2.2.4 Generation and selection of human-generated questions

The manually generated questions were obtained from a high school teacher of English and Media. The teacher was given access to the documentary video and a transcript and was asked to produce comprehension questions that they would also use in their classroom were they to utilise this video in one of their teaching sessions. The teacher was also instructed to generate the questions in such a way that they could be answered solely with information from the video and did not require any additional knowledge. The human expert generated 22 questions in about 80 minutes, 9 of which were selected for the experiment at random.

5.2.2.5 Selection of pre-questions

For the first two hypotheses, the focus is on whether or not pre-questions help the performance of test-takers, rather than the generation method of pre-questions. As a result, pre-questions were selected manually from system A's pool of generated questions. Pre-questions were selected based on two premises. Firstly, a question was deemed a suitable pre-question if it revolved around an important concept in the documentary. Secondly, a question was selected as a pre-question if the same or a similar question was also generated by one or more of the other systems. For example, the question "What is nuclear fusion?" was selected as a pre-question because it revolves around a central concept in the documentary. In addition, the same question was generated by the human expert. An

example for similar questions generated by all three methods can be seen in Table 10. The development of automatic selection methods for pre-questions and their evaluation will be left to future research.

	Question	Answer
System A	What did some scientists suspect that Rusi Taleyarkhan’s fusion neutrons could in fact be coming from?	From his own neutron generator
System B	What did Mike Saltmarsh think that any fusion finding could be explained by?	From the pulse neutron generator
Manual	What did the other scientists criticise about Taleyarkhan’s first experiment?	Other scientists criticised that the neutrons detected in the experiment might be background neutrons from the neutron generator.

Table 10 Questions with similar content generated by all three QG methods

5.2.2.6 Selection of images

The screenshots are extracted using the following process. After questions have been generated, the source sentence of a question (i.e. the sentence which gave rise to a question) is mapped to the time stamp contained in the subtitles. Then a screenshot is taken from the video at the respective time a source sentence occurs in the video. For example, the sentence “It was Mike Saltmarsh’s task to work out whether the neutrons detected could indeed be from fusion or were simply background neutrons from the neutron generator” which occurred 29 minutes and 15 seconds into the video gave rise to the first question and screenshot in Table 11.


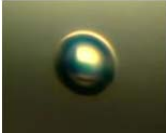

Question	Answer	Image
Whose task was to work out whether the neutrons detected could indeed be from fusion or were simply background neutrons from the neutron generator?	Mike Saltmarsh's	
What should be produced at exactly the same billionth of a second if fusion was happening?	Fusion neutrons	
What is nuclear fusion?	A nuclear reaction in which atoms are forced together until they fuse, giving off massive amounts of heat, light and energy.	

Table 11 Pre-questions with screenshots extracted from the video

5.2.2.7 Participants and interface

29 students took part in the experiment. All participants were final year undergraduate students at a university in Spain reading translation with a major in English. The participants had access to the experiment via an online interface²². Instructions for the experiment (e.g. note-taking was allowed, but participants should watch the video only once) were displayed in the interface. The interface provided access to the video and tracked each participant's answers and time spent to answer each question. In Appendix E, a few sample screenshots from the interface can be seen.

5.2.2.8 Experimental Design

The video used was a documentary on 'nuclear fusion' (Horizon, 2005). The experiment consisted of three parts, each corresponding to a 10-minute section of the documentary.

²² The experiment can be accessed at: <http://www.bootlace.eu/quiz/randq/>

The participants were divided into three groups. Before each part of the video was shown, participants were given either three pre-questions containing a screenshot extracted from the video, three text-only pre-questions or no pre-questions, depending on their group (cf. Table 12). After each part of the video, the students were asked to answer nine comprehension questions (post-questions) about what they had just seen in the video. Three of those questions had been generated by system A, three by system B and three by a human expert. The post-questions were identical for all participants.

	Group 1	Group 2	Group 3
Part 1	Pre-questions + screenshots	Pre-questions no screenshots	No pre-questions
Part 2	Pre-questions no screenshots	No pre-questions	Pre-questions + screenshots
Part 3	No pre-questions	Pre-questions + screenshots	Pre-questions no screenshots

Table 12 Pre-question scenarios. All scenarios contained 9 identical post-questions, 3 generated by each method (WLV, CMU and Manual)

The experimental design employed has several advantages. With the setup, a unique way of economising the experiment and maximising the amount of research questions investigated was created, whilst eliminating problems that can be incurred by cross-group-performance comparison and cross-question-comparison.

The experiment followed best practices. The groups in the experiments were randomised; when the test-takers accessed the experiment interface, they were randomly assigned to one of the three groups. The test-takers were unaware which question was generated by which system and there was no influence on which test-taker was assigned which question, making the experiment a double-blind study. The experiment exhibits all the advantages of a crossover experiment with a repeated measures design; the same measures are collected multiple times for each test-taker, in a balanced way (all test-takers received the same

number of questions and went through the same number of pre-question scenarios). This design allows for the elimination of cross-question-comparison issues in the experiment, because the post questions given to the test-takers are identical across groups, generated based on one documentary (rather than comparable ones). For example, the post-questions that Group 1 received in part 1 of the experiment were identical to the ones that Group 2 and Group 3 received, but the difference is in the pre-question setting. In parts 2 and 3 of the experiment, the questions are identical across the groups again, but the pre-question setting differs again. Not only does the experiment examine and isolate the effect of different types of pre-questions (and use no pre-questions as a control setting), at the same time, it also evaluates, in a fair and valid way, the psychometric parameters of the post-questions generated by system WLIV, a state-of-the-art system and manually generated questions. This novel and unique evaluation method is an improvement over existing approaches, such as those presented in Mitkov, Ha and Karamanis (2006) and Heilman (2011), by eradicating variables which could influence the test results and at the same time examining a large number of research questions at the same time.

5.2.3 Results

5.2.3.1 Answering research question 1: accuracy

Firstly, a χ^2 test of independence was used to determine whether the performance across the groups differed significantly; there was no evidence to suggest so. Table 13 shows the breakdown of correctly and incorrectly answered post-questions for each pre-question type (Q_{np} =no pre-questions, Q_{tp} =text-based pre-questions, Q_{sp} =pre-questions with screenshots).

Due to time constraints, not all test-takers answered all questions, which is the reason for the total number of questions answered varying for each pre-question type. Proportionally, the highest number of correctly answered questions is observed where test-takers were given pre-questions with screenshots, followed by text-based pre-questions. Test-takers who did not receive any pre-questions at all produced the smallest proportion of correct answers.

Pre-question type	Correct	Incorrect	Total	% correct
Q_{np}	75	113	188	39.83
Q_{tp}	86	85	171	50.29
Q_{sp}	84	60	144	58.33
$(Q_{tp}+Q_{sp})$	(170)	(145)	(315)	(53.97)

Table 13 Breakdown of correct and incorrect answers per pre-question type

A χ^2 test was performed to determine whether these results are statistically significant. When comparing the performance of students who did not receive pre-questions (Q_{np}) to the performance of students who received only text-based pre-questions (Q_{tp}), the result is statistically significant ($p=0.047$). The same applies when the performance of students who did not receive pre-questions is compared with that of students who had received pre-questions with screenshots (Q_{sp}); the difference is statistically significant by a lower p-value ($p=0.00085$). When text-based pre-questions and pre-questions with screenshots are grouped together ($Q_{tp}+Q_{sp}$) and compared to no pre-questions (Q_{np}), the result is also statistically significant ($p=0.00225$). However, when comparing the performance of students who received text-based pre-questions with that of those who received pre-questions with screenshots, no statistically significant difference ($p=0.1537$) was found. It

can thus be concluded that test-takers who receive pre-questions (with or without image) tend to perform better on a comprehension test than those who receive no pre-questions at all.

5.2.3.2 Answering research question 2: time taken to answer post-questions

For each test taker, the time to answer a question was measured. It was hypothesized that the presence of pre-questions would affect the time taken to answer post-questions. The highest mean value (cf. Table 14) was observed in the pre-questions with screenshots condition (Q_{sp}), followed by text-based pre-questions (Q_{tp}). The lowest average time required to answer a question was observed in the no pre-questions condition (Q_{np}). However, there appears to be no significant difference between the means of the different conditions, which is confirmed by a single-factor analysis of variance. It can thus be concluded that the presence of pre-questions, with or without screenshot, does not affect the time taken to answer post-questions significantly.

	Min t in s	Max t in s	Mean	SD
Q_{np}	2	237	53.26	44.38
Q_{tp}	3	403	54.84	55.07
Q_{sp}	5	306	58.57	46.49

Table 14 Seconds taken to answer post-questions depending on pre-question type

5.2.3.3 Answering research questions 3: psychometric parameters

Classical test theory can provide information about the effectiveness of a question (also referred to as ‘item’). One measure is item discriminating power (DP) (Gronlund, 1982).

DP describes the relationship between student performance on a particular item and their total exam score. DP ranges from -1.0 to 1.0; the higher the value, the more discriminating the item. A high DP means that test takers with overall high scores answered the item correctly, whereas test takers who performed poorly overall did not answer the item correctly. On the converse, a low DP indicates that poorly performing test takers answered an item correctly whereas test takers with overall high scores did not answer an item correctly; this means that the item may be confusing for better scoring test takers. Items with near zero or negative DP should not be used for assessment. To calculate DP, test results need to be ranked from highest to lowest score. Two equal-sized groups are formed, the ‘upper group’ containing the tests with the highest scores, and the ‘lower group’ containing those with the lowest scores. DP is calculated as follows:

$$DP = \frac{R_U - R_L}{\frac{1}{2}P}$$

Where DP is the discriminating power, RU is the number of right answers from the upper group, RL is the number of right answers from the lower group, P is the number of total participants. The results for the discriminating power for the three QG methods can be seen in Table 6.

	Min	Max	Mean DP
System A	-0.15	0.44	0.16
System B	-0.22	0.22	0.07
Manual	0.15	0.59	0.37

Table 15 Discriminating powers for all three QG methods

The manually created questions exhibit the highest average DP, followed by system A and lastly system B. The application of Student’s t-test shows that there is a statistically

significant difference between system A's mean DP and the manual questions' mean DP ($p=0.0434$). The same applies when comparing system B's mean DP to that of the manual questions. However, no statistically significant difference could be observed between system A's and system B's mean DPs ($p=0.356988$). While this means that neither automatic system's questions are as good as questions generated by human experts at distinguishing between well and poorly performing students. It also means that system A's questions are as good as, if not better than, those generated by the state-of-the-art system.

5.2.3.4 Qualitative feedback from test-takers

At the end of the quiz, the test-takers were asked to answer a number of feedback questions. As the quiz was taken during a lecture and thus there were time constraints on the participants, answering these feedback questions was not compulsory, as the aim was to make sure that as many students as possible answer all the comprehension questions. Even though only 7 students answered the feedback questions, this qualitative feedback provides a unique insight into what test-takers thought about the quiz. The following questions were:

1. **How challenging did you find the questions in this quiz?** (A: Most questions too challenging/ B: Some questions too challenging/C: They were just right/D: Some questions not challenging enough/E: Most questions not challenging enough)
2. **Did you find reading some of the questions before the video helpful?** (Yes/No)
3. **Did you find the images in some of the questions helpful at all?** (Yes/No)
4. **Did you have any prior knowledge about 'nuclear fusion'?** (Yes/No)

5. **What is your English proficiency level?** (Native, Advanced, Intermediate, Beginner)
6. **Are you male or female?**
7. **Please tell me your age in years**
8. **Finally... Do you have any other feedback for us? We are particularly interested in what you thought about the questions in this quiz.**

Table 16 shows the responses given by the students as well as their overall score on the quiz. One student found most of the questions too challenging, while four students found some of the questions too challenging. One student felt they were 'just right'. Interestingly, the student who felt that most questions were too challenging, achieved a better score on the quiz (60% of questions answered correctly) than the student who felt they were just right (52% of questions answered correctly). The students who felt that some of the questions were too challenging, answered 33%, 72%, 56% and 56% correctly, respectively. All but one student found the text-based pre-questions helpful and 4 students found the supplied images helpful. One student stated that the text-based pre-questions were very helpful as they helped her focus on the important aspects in the video; however she also noted that she paid more attention at the beginning of the video and had the impression that most questions asked about information depicted at the beginning of the video and she suspected that she might not have answered questions correctly if they had asked about information in the later sections of the video, due to not paying attention anymore.

Student	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Quiz Score
A	B	NO	-	-	-	-	-	-	33%
B	B	YES	YES	YES	Adv	F	21	I thought some of the questions were difficult, but I found it helpful since it helped have some knowledge about what nuclear fusion is.	72%
C	B	YES	YES	NO	Adv	F	23	I'm glad to help you. I'm pretty sure that I would have done it better if I had taken some notes while listening.	56%
D	C	YES	NO	NO	Adv	F	22	The questions before helped a lot, because it helped to focus on important issues. I paid more attention at the start of the video and most questions were about the beginning of the video. Maybe if the questions were taking from a later section, I might have got them wrong due to not paying attention anymore. Also, some questions were similar to what was said in the video and I found these questions not challenging.	52%
E	B	YES	YES	NO	Adv	F	22	For an english speaker beginner some questions could have been problematic due to their grammatical construction.	56%
F	-	YES	NO	NO	Adv	M	21	I found some questions really redundant, an some of them asked for really specific data, so I had to look for these watching again some parts of the videos.	76%
G	A	YES	NO	NO	Adv	F	21	Some of the questions was difficult to understand because they were very long and ambiguous.	60%

Table 16 Qualitative feedback from test-takers

One student, who scored highly on the quiz, remarked that he had previous knowledge about nuclear fusion and that this might have helped him on the quiz. Conversely, student F who had no prior knowledge about nuclear fusion, scored even higher (76%). When setting up the experiment, influence due to prior knowledge was a concern and while this can never completely be eradicated, a documentary was chosen that dealt with a very specific topic, so as to keep this influence at a minimum. All students considered themselves to be of advanced English proficiency level and one student noted that for

beginner learners of English, some questions might be difficult to understand due to their grammatical constructions and another student mentioned that some questions were difficult to understand because they were long and ambiguous.

5.2.4 Conclusion

A novel evaluation approach was presented which enabled to the investigation of a large number of research questions, while at the same time eliminating variables such as cross-group-performance and cross-question-performance. The findings show that both text-based pre-questions and pre-questions with images lead to a larger number of correctly answered post-questions (as opposed to using no pre-questions). Supplying a screenshot alongside a pre-question will result in a statistically more significant difference of correctly answered questions when compared to no pre-questions. The ability to supply a screenshot alongside a question is unique to system WLW. The average time taken to answer a question is not statistically significantly different between the pre-question settings. It was examined whether questions generated by system WLW exhibit a discriminating power (DP), comparable to that of questions generated by human experts and a state-of-the-art system. It was found that manually created questions exhibit the highest DP and there is no statistically significant difference between system WLW and the state-of-the-art system, implying that questions generated by system WLW are as good as, if not better than, questions generated by the state-of-the-art system. A number of issues need to be addressed in future research. The feasibility of automatically or semi-automatically choosing pre-questions needs to be explored. Furthermore, there is a need to

investigate whether other images taken from other sources (e.g. Google Image search) can also be used in pre-questions. A large-scale experiment investigating the productivity of generating questions (in terms of time taken to post-edit questions vs. time taken to generate questions from scratch) is presented in the next Chapter.

5.3 Summary

In this Chapter, evaluation in Question Generation was discussed and an evaluation methodology was proposed to assess the framework. While evaluation in Question Generation is application-dependent because of the different aims that Question Generation systems might serve, researchers have proposed that for many applications of QG the Question Generation procedure can be regarded as a three-step-process, consisting of *key concept identification*, *question type determination* and *question realisation*. Just like in other areas of NLP, humans can be used to provide a gold standard; however, while human created gold standards can be very accurate, creating them is costly and time extensive. A modified version of the F-measure was discussed which takes into account human-annotated vital and optional snippets of information. Several tasks which have been created as part of the Question Generation Shared Task and Evaluation Challenge (QG STEC) have been introduced. The evaluation techniques discussed at the QG STEC are merely suggestions of how evaluation in question generation could be performed, rather than prescriptive instructions. The next Chapter deals with the evaluation of the proposed framework; two experiments that were performed to assess different aspects of the approach to Question Generation from Video Documentaries are described. The evaluation

methods discussed in this Chapter are not employed; instead, a new evaluation method is proposed which was used in an experiment described in the second part of this Chapter. In the experiment, a user study, the feasibility of using different types of pre-questions (with image and text-based) was explored and the psychometric parameters of questions generated by system WLW, a state-of-the-art system and manually generated questions were analysed. Using a novel evaluation approach, the results show that both text-based pre-questions and pre-questions with images lead to a larger number of correctly answered post-questions (as opposed to using no pre-questions). Supplying a screenshot alongside a pre-question will result in a statistically more significant difference of correctly answered questions when compared to no pre-questions. The average time taken to answer a question is not statistically significantly different between the pre-question settings. It was analysed whether questions generated by system WLW have a discriminating power (DP), comparable to that of questions generated by human experts and a state-of-the-art system. It was found that manually created questions exhibit the highest DP and there is no statistically significant difference between system WLW and the state-of-the-art system, implying that questions generated by system WLW are as good as, if not better than, questions generated by the state-of-the-art system.

CHAPTER 6: EVALUATION 2 - POST-EDITING VERSUS MANUAL GENERATION

This Chapter describes a second experiment which was performed in order to evaluate the quality of the questions output by the framework. The efficiency of post-editing system-generated questions versus creating questions manually is examined. Using a state-of-the-art post-editing tool, a number of question usability statistics were gathered and the operations that post-editors perform when post-editing questions were analysed.

6.1 Experiment 2

In this experiment, the efficiency of the QG system was analysed. The hypothesis is that it is faster to post-edit questions that the QG system-generated as opposed to creating the questions manually and that the quality of the questions is not compromised. A post-editing environment named PET (Aziz et al., 2012) was used to gather a number of statistics about the post-editing process, such as Human-targeted Translation Error Rate (HTER, Snover et al., 2006), usability scores of questions before and after post-editing and perceived post-editing effort. This data allowed to analyse how the human experts post-edited questions and what they change to make questions ‘usable’.

6.1.1 Background and related work

Previous research in QG (Mitkov, Ha and Karamanis, 2006) found that systems for computer-aided Question Generation can generate multiple choice test questions faster

than a human expert can do from scratch without compromising quality. In their experiment, Mitkov, Ha and Karamanis (2006) observed that the manual creation of a test question took on average 6 minutes, while the post-editing of a system-generated question took on average 1 minute and 40 seconds. While Mitkov, Ha and Karamanis (ibid.) list a few operations that evaluators performed during post-editing, such as the removal of discourse words, they do not provide an exhaustive analysis of the changes that evaluators make to system-generated questions. However, such an analysis is useful in order to understand what distinguishes system-generated questions from the ones humans generate and could ultimately help to improve the output of QG systems.

The aim of this experiment is to fill this gap in the literature and to extensively analyse the operations humans perform when post-editing system-generated questions, as this has not been done in detail before. The post-editing tool PET (Aziz et al., 2012) was used for this purpose. PET has been designed as a post-editing tool for Machine Translation which allows the post-editing or revision of translations from any MT system and collects segment-level information from this process, such as translation quality scores and post-editing time. As PET is highly customisable, it was possible to use it for this experiment's purposes (how PET was used to fit the purposes is described in the following section).

6.1.2 Methodology

Firstly, two teachers, one primary, and one secondary school teacher, were asked to watch three video clips from three different documentaries, each about 10 minutes in length. From here on after, these teachers will be referred to as “Creator 1” and “Creator 2”. For

each part of each documentary, the creators were asked to generate at least 3 factual comprehension questions about the information seen in the videos along with the correct answers to these questions.

Next, two other school teachers (both secondary school teachers) were asked to watch the documentary videos and then post-edit 574 questions each using PET. From here on after, they will be referred to as “post-editor 1” and “post-editor 2”. In addition to the manually created questions by the Creators, just like in the previous experiment described in Section 5.2, the question pool to be evaluated contained equal amounts of questions generated by the system developed as part of this thesis (system WLV) and by the state-of-the-art system (system CMU) described by Heilman (2011). The Post-editors were unaware which question was generated by which method.

Figure 8 is a screenshot of PET and shows how the questions were presented to the Post-editors. The screen is divided into 2 columns. In a Machine Translation context, these columns would be used to compare a machine translated sentence with its source sentence. In the context of the experiment, the question to be evaluated can be seen in the right hand column, while the source sentence which gave rise to it can be seen in the left hand column. The answer to a question can be found in the top row. Unfortunately, it cannot be edited; but post-editors were instructed to leave a comment in the feedback section if an answer required post-editing.

By clicking onto one of the questions, PET starts recording the editing time for that question alongside several other post-editing statistics. Once a post-editor edited a question and is content with it, they can proceed to the next question by clicking the corresponding

next button on the right hand control bar. If a question does not require post-editing (because it is either good enough as it is or it is so bad that it should be rejected), the evaluator can press the next button to skip to the next question. Every time a question has been edited, the post-editor is taken to an assessment screen (see Figure 10) in which they are asked to judge the post-editing effort as well as the usability of the question before and after post-editing. Due to PET being highly customisable, it is possible to set the feedback questions to whatever is required. In the context of this experiment, the post-editors were asked first to set a flag 'accept' or 'reject' for a question and to provide a comment if a question was rejected. Next, they were asked to judge the usability of a question before it was post-edited, using a scale from 1 to 5, or 'rejected' (5= very good, 4= good, 3= borderline, 2= bad, 1= unusable). Usability refers to the quality of a question with regards to syntactical soundness and whether it would be a good question to be used in a comprehension test. Post-editors were instructed to score questions as 'rejected', if the question could not be altered to make a usable question, whereas a score of 1 was to be assigned when a question was unusable in its pre-post-editing state, but with post-editing could be transformed into a usable question. Then the evaluators were asked to quantify the amount of post-editing effort required, on a scale of 1 to 3, or 'rejected' (1= no modifications were made, 2= some modifications were made, 3= major modifications were made). Finally, the Post-editors were asked to judge the quality of the question after post-editing, with a scale from 1 to 5 or 'rejected', or 'no post-editing' if they selected 'no modifications were made' in the previous question (5= very good, 4= good, 3= borderline, 2= bad, 1= unusable).

editing... partial: 25s revisions: 0 total: 0s

The Origin of species		2/184 2 saved 04:12:16
Like all nuclear devices, Garwin's exploited an extraordinary property of Einstein's equation, that tiny amounts of matter would contain massive amounts of energy.	What did Garwin's device exploit?	
In fact, we are only recently beginning to tease apart, nanosecond by nanosecond, the precise process by which nature performs this magic called photosynthesis.	What are we only recently beginning to tease apart, nanosecond by nanosecond, the precise process by which nature performs this magic called photosynthesis in fact?	
For Watson, the benefits of this moment far outweigh any risk.	What outweigh any risk for Watson?	
Photosynthesis is the process by which plants take sunlight, combine it with water and carbon dioxide and create energy.	What is Photosynthesis by which plants take sunlight, combine it with water and carbon dioxide and create energy?	
On this day, Charles Darwin published The Origin Of Species.	What was the name of Charles Darwin's influential book on evolutionary biology?	
With just five characters, Albert Einstein revealed an extraordinary truth of our world.	Who revealed an extraordinary truth of our world with just five characters?	
Perhaps most significantly, the Periodic Table can help us understand our relationship to the nearest thing we have to a life force.	What can the Periodic Table help miraculous processes understand to the nearest thing we have to a life force?	
This is the Natanz Nuclear Centre in Iran.	What is the Natanz Nuclear Centre in Iran?	
James Dewey Watson was one of the team that, in 1953, unravelled one of nature's deepest secrets.	Who was one of the team that, in 1953, unravelled one of nature's deepest secrets?	
By what percentage has the carbon dioxide in the atmosphere increased in the last 100 years?	When did the carbon dioxide in the atmosphere increase?	

Figure 8 Post-editing questions in PET Screenshot

Finish

MT

What was the name of Charles Darwin's influential book on evolutionary biology?

Post-edited MT

What was the name of Charles Darwin's influential book on evolutionary biology?

Accept or reject this question? If you reject a question, please give a reason in the comment box.

Accept

Comment on ' Accept or reject '

How good was the question before post-editing? Please select 'rejected' if you rejected the question (5=very good, 4=good, 3=borderline, 2=bad, 1=unusable)

5

Comment on ' Usability before '

The question is grammatically correct and can be used as it is. It is a good question in the context of the documentary.

How much post-editing was required? Please select 'rejected' if you rejected the question(3= major changes were made, 2=some modifications were made, 1= no modifications were made)

1

Comment on ' PE effort '

How good is the question after post-editing? Please select 'rejected' if you rejected the question or 'no post-editing' if you did not need to make any changes (5=very good, 4=good, 3=borderline, 2=bad, 1=unusable)

no post-editing

Comment on ' Usability after '

Figure 9 Assessing post-editing effort and question quality in PET

6.1.3 Results

6.1.3.1 Time spent post-editing questions versus generating from scratch

Firstly, a test of independence was performed to determine whether the obtained results differed significantly between the documentaries; there is no evidence to suggest there is. Appendix F shows an extract of sample output created by PET. Table 17 shows the number of questions post-edited and the number of questions manually created for the different documentaries while Table 18 shows the results for the time taken to post-edit questions versus time taken to generate questions from scratch.

	Post-editor 1			Post-editor 2			Creator 1	Creator 2
	WLV	CMU	Total	WLV	CMU	Total		
Documentary 1	76	99	175	76	99	175	9	9
Documentary 2	89	90	179	89	90	179	10	9
Documentary 3	88	90	178	88	90	178	9	9
Total	253	279	532	253	279	532	28	27

Table 17 Numbers of question post-edited and manually created

Some of the questions the Post-editors edited were manually created ones. These were not included in the analysis of post-editing time; the post-editing time is the amount of time taken to post-edit system-generated questions (by system WLV and system CMU). Creator 1 generated 28 questions in total in 2 hours and 49 minutes, averaging at ~362 seconds per question (6 min 2 secs). Creator 2 generated 27 questions in 2 hours and 35 minutes, averaging at ~344 seconds (5 mins 44 secs) per question. These timings are similar to the findings made by Mitkov, Ha and Karamanis (2006). The results also indicate that post-

editing questions is faster than the manual creation from scratch, with the average post-editing time per question being ~12 seconds. An analysis of variance confirms that there is no statistically significant difference between the post-editors' times. However, there is a statistically significant difference when comparing post-editing time to creation from scratch.

	Post-editor 1		Post-editor 2		Creator 1		Creator 2	
	Secs/ question	Time, total	Secs/ question	Time, total	Secs/ question	Time, total	Secs/ question	Time, total
Documentary 1	12.95	00:37:51	12.50	00:36:35	366.66	00:55:00	320	48:00
Documentary 2	17.29	00:51:46	13.07	00:39:10	372	01:02:00	340	51:00
Documentary 3	8.02	00:23:50	15.04	00:44:16	346.66	00:00:52	373.33	56:00
All	12.76	1:53:27	13.78	2:00:01	362.14	02:49:00	344.44	02:35:00

Table 18 Comparison between post-editing times and manual creation

Unfortunately, due to time and other constraints, another user study in which these questions were employed in a comprehension test could not be performed; however, the experiment in Section 5.2 showed that system-generated questions exhibit psychometric scores similar to human-generated ones. It can thus be concluded that system-generated questions are not only as good as human-generated ones in terms of quality, it is also considerably faster to post-edit system-generated questions as opposed to creating them manually. Consequently, Question Generation systems can help to save money and resources in educational settings.

6.1.3.2 Usability Before Post-editing

Using PET, the post-editors were asked to assign a score for each post-edited question, judging the quality of the question before and after post-editing and the perceived post-editing effort.

Table 19 shows the ratings for ‘usability before post-editing’ for both post-editors broken down by documentary and system and the sum of scores. A number of interesting findings can be made from these statistics. In total, post-editor 1 rejected 403 out of 574 questions (70%) and post-editor 2 rejected 298 out of 574 questions (51%). For both post-editors, most rejected questions had been generated by system CMU (56% and 55% of rejected questions, respectively). 43% and 44% of rejected questions, respectively, were those generated by system WLV.

As mentioned before, post-editors were unaware which question was generated by which method. Interestingly, both post-editors each rejected one manually generated question (making up the remaining 1% of rejected questions). The rejected manually generated questions were not identical; post-editor 1 rejected a question from documentary 1, while post-editor 2 rejected a question from documentary 3. Post-editor 1 rejected the question “What could be an advantage of the Human Genome Project?”, commenting that the question was not important in the context of the documentary, while post-editor 2 rejected “What did Rusi Taleyarkhan do before he sent his results to Science Magazine?”, commenting that the question was grammatically correct, but irrelevant.

Post-editor 1	Doc1	1	2	3	4	5	Reject	Post-editor 2	Doc1	1	2	3	4	5	Reject	
	Manual	0	0	0	4	4	1		Manual	0	1	1	2	5	0	
	WLV	6	3	10	7	2	48		WLV	6	8	14	3	6	38	
	CMU	3	4	2	7	1	82		CMU	3	4	2	7	1	82	
	Doc2	1	2	3	4	5	Reject		Doc2	1	2	3	4	5	Reject	
	Manual	0	0	0	2	10	0		Manual	0	0	3	1	8	0	
	WLV	8	4	0	9	1	64		WLV	1	9	15	8	9	47	
	CMU	7	5	5	3	0	72		CMU	7	4	10	8	10	51	
	Doc3	1	2	3	4	5	Reject		Doc3	1	2	3	4	5	Reject	
	Manual	0	0	1	8	13	0		Manual	0	1	1	2	17	1	
	WLV	6	2	4	10	5	61		WLV	5	9	14	2	11	47	
	CMU	6	1	3	4	1	75		CMU	4	18	16	12	8	32	
Total	1	2	3	4	5	Reject	Total	1	2	3	4	5	Reject			
Manual	0	0	1	14	27	1	Manual	0	2	5	5	30	1			
WLV	20	9	14	26	8	173	WLV	12	26	43	13	26	132			
CMU	16	10	10	14	2	229	CMU	14	26	28	27	19	165			
	36	19	25	54	37	403		26	54	76	45	75	298			

Table 19 Ratings for ‘usability before post-editing’ assigned per documentary and system

On average, post-editor 1 assigned a usability before post-editing score of 4.61 (post-editor 2: 4.5) for the manually created non-rejected questions, an average score of 2.90 (post-editor 2: 3.12) for questions generated by system WLV and 2.53 (post-editor 2: 3.1) for questions by system CMU. Inter-annotator agreement on all questions (rejected and accepted) was calculated using Cohen’s weighted Kappa (Cohen, 1960). For documentary 1, the inter-annotator agreement is $k=0.414$, which is a moderate agreement. For

documentary 2, the inter-annotator agreement is $k=0.363$, which is also a moderate agreement. For documentary 3, the inter-annotator agreement was moderate, with a value of $k=0.436$. This shows that rating questions is a subjective task.

6.1.3.3 Usability After Post-editing

Table 20 shows the ratings assigned by the post-editors, per documentary, judging the usability of the questions after they have been post-edited. Note that the number of rejected questions remains unchanged, post-editors had been instructed to score questions as ‘rejected’, if the question could not be altered to make a usable question, whereas a score of 1 was to be assigned, when a question was unusable in its pre-post-editing state, but with post-editing could be transformed into a usable question.

Post-editor 1	Doc1	1	2	3	4	5	Reject	Post-editor 2	Doc1	1	2	3	4	5	Reject
	Manual	0	0	0	4	4	1		Manual	0	0	1	2	6	0
	WLV	0	0	10	8	9	49		WLV	0	3	11	13	10	38
	CMU	0	0	6	9	2	82		CMU	0	0	4	11	2	82
	Doc2	1	2	3	4	5	Reject		Doc2	1	2	3	4	5	Reject
	Manual	0	0	0	2	10	0		Manual	0	0	1	1	10	0
	WLV	0	0	1	11	10	64		WLV	0	1	8	22	11	47
	CMU	0	0	5	12	3	72		CMU	0	0	11	15	13	51
	Doc3	1	2	3	4	5	Reject		Doc3	1	2	3	4	5	Reject
	Manual	0	0	1	8	13	0		Manual	0	1	1	2	17	1
WLV	0	0	5	13	9	61	WLV	0	6	10	13	12	47		
CMU	0	0	4	9	2	75	CMU	0	11	17	19	11	32		
Total	1	2	3	4	5	Reject	Total	1	2	3	4	5	Reject		
Manual	0	0	1	14	27	1	Manual	0	1	3	5	33	1		
WLV	0	0	16	32	29	173	WLV	0	10	29	48	33	132		
CMU	0	0	15	30	7	229	CMU	0	11	32	45	26	165		
	0	0	32	76	63	403		0	22	64	98	92	298		

Table 20 Ratings for ‘usability after post-editing’ assigned per documentary and system

Thus, the most interesting aspect is to investigate how many questions changed from score 1 to a higher score after post-editing and how post-editing affected the average scores of questions. Figure 10 and Figure 11 show the usability ratings before and after post-editing side by side for the two post-editors, respectively.

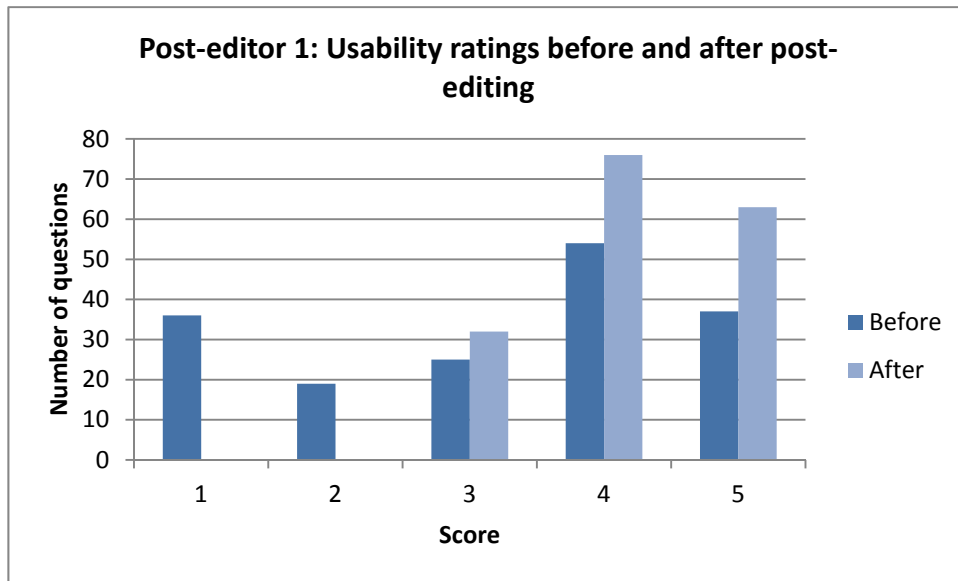


Figure 11 Usability ratings before and after post-editing assigned by post-editor 1

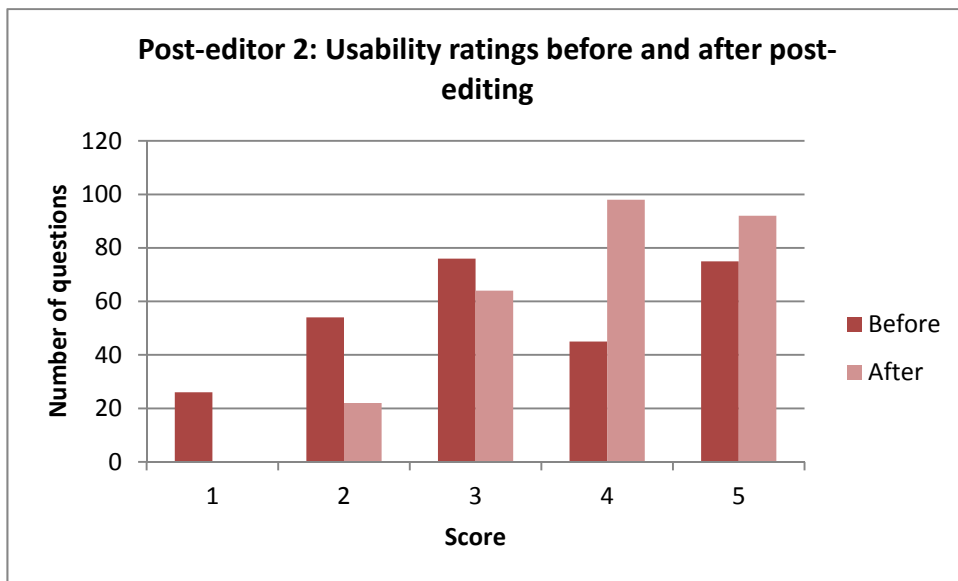


Figure 12 Usability ratings before and after post-editing assigned by post-editor 2

While before post-editing, the post-editors classified 36 (26) questions as unusable (score 1), after post-editing, neither post-editor assessed a question as unusable. Interestingly, in

post-editor 2's dataset, several questions were judged as being unusable before post-editing but as "bad" (score 2) after post-editing, while post-editor 1's question all scored at least 3 (borderline) after post-editing.

The change of scores from 1 to 2 occurred twice in the dataset, once with the question "What was he convinced it simply had to be from?" and once with "What did most groups across the globe agree that all they could find was?". Both questions were questions generated by system CMU. Unfortunately, the post-editor left no comments about the post-editing process of these two questions. As the idea of post-editing is to improve the question in such a way that it is usable in a comprehension test about a documentary, it is unclear why the post-editor decided not to reject the question straight away, as a question that is 'bad' even post-editing, would be unlikely to be chosen for a comprehension test.

Just like before, inter-annotator agreement on all questions (rejected and accepted) was calculated using Cohen's weighted Kappa (Cohen, 1960). For documentary 1, the inter-annotator agreement was $k=0.346$, which is a fair agreement. For documentary 2, the inter-annotator agreement was $k=0.282$, which is also a fair agreement. Only for documentary 3 the inter-annotator agreement was moderate, with a value of $k=0.404$.

Table 21 shows the average usability scores from before post-editing compared to the average usability scores after post-editing. Apart from the manual questions from post-editor 1, an increase in average score can be observed in all cases.

Post-editor 1		Before	After	Post-editor 2		Before	After
	Manual	4.61	4.61		Manual	4.5	4.66
	WLV	2.90	4.16		WLV	3.12	3.86
	CMU	2.53	3.84		CMU	3.09	3.75

Table 21 Average scores assigned by post-editors before and after post-editing, per system

The number of points a question improved by was also calculated by subtracting a question’s score before post-editing from its score and after post-editing. These results can be found in Table 22. Post-editor 1 made changes to 67 questions; 104 questions were accepted without post-editing and their score thus did not change. An example for a question that was accepted without changes with a score of 4 (“good”) is the question “What was unearthed in 1934 on the Northbourne estate?” which was generated in relation to a documentary about Anglo-Saxon Art History. The majority of post-edited questions improved by 3 points. Ten questions were changed from “bad” (score 1) to “very good” (score 5). An example for such a modification is the following:

The question “What did they become known as?” was changed into “What did the elite scientists that worked as part of a secret organisation founded in the 1960s become known as?” It is clear that the question has been changed significantly by the post-editor. The post-editor considered the resulting question so good, that it was not outweighed by the post-editing effort. In the comments section, the post-editor explained that he had remembered the context of the question and deemed it an important part of the documentary and thus created a very good question. This shows that one of the advantages of a QG system can also be to “trigger” post-editors’ memory to generate very good

questions from questions that were unusable before post-editing and that even seemingly unusable questions can be beneficial in the Question Generation process.

Post-editor 2 made changes to 96 questions and accepted 180 questions without post-editing. The majority of post-edited questions were improved by one point. One such question that was improved by one point was a manually generated question; post-editor 2 changed the question “What did the boar symbolise for Anglo-Saxons?” to “In the pagan beliefs of early Anglo-Saxons, what did the board symbolise?”, commenting that the question was not accurate/too broad before. By adding the additional information to the question, the post-editor felt that the question had improved from “borderline” (score 3) to “good” (score 4).

Improved by (points)	1	2	3	4	No PE	Edited
Post-editor 1	15	15	27	10	104	67
Post-editor 2	43	28	23	2	180	96

Table 22 Difference in question scores before and after post-editing

6.1.3.4 HTER

One of PET’s capabilities is to compute HTER (Snover et al., 2006) values. HTER stands for “Human-targeted Translation Error Rate” and is an edit-distance measure used to evaluate machine translated sentences by comparing them to a human-generated reference translation, calculating the fewest modifications (edits) required to the system output, so that the complete meaning of the reference is captured (Libermann, 2008). The formal definition is:

$$HTER = \frac{(Substitutions + Insertions + Deletions + Shifts)}{Reference\ Words}$$

Although HTER was developed with the evaluation of machine translations in mind, in the experimental scenario, it is used to compare system-generated questions to the ones post-edited by the evaluators. Bernhard et al. (2012) use HTER in the evaluation of their QG system; however, they do not provide a comprehensive analysis of how post-editors actually change system-generated questions. As it was already established that post-editing questions is faster than generating them entirely manually, the HTER values show how and to which extent system-generated questions differ from the ones post-edited by humans (the gold standard). Table 23 displays the number of different types of operations that have been performed as well as the average HTER value for all questions that were post-edited.

	Post-editor 1		Post-editor 2	
	Sum	Avg.	Sum	Avg.
Deletions	246	6	244	5.42
Insertions	73	2.28	91	2.33
Substitutions	112	2.29	125	1.89
Word Shifts	16	1.33	27	1.28
Phrase Shifts	18	1.5	51	2.43
HTER	0.65		0.5	

Table 23 Sums and averages of edit operations and HTER values per post-editor

The average HTER value is 0.65 for post-editor 1 and 0.5 for post-editor 2. The highest HTER value obtained was 4, in which case the system-generated question “What is Photosynthesis by which plants take sunlight, combine it with water and carbon dioxide

and create energy?” was post-edited to the question “What is photosynthesis?”. In this case, 16 words were removed in order to change the question from its original state to its post-edited state. The lowest HTER value obtained in the dataset was 0.06, in which case the question “When has Professor Fleischmann found it hard to get papers published in scientific journals?” was changed to “Since when has Professor Fleischmann found it hard to get papers published in scientific journals?”.

The average value for the operations is similar for both post-editors, indicating that the amount of post-editing work performed was similar. At this point, it is important to note that while many post-edited questions were changed using only one type of edit operation, some questions exhibit several types of edit operation simultaneously.

For both post-editors, the most frequently performed operation was ‘deletion’. This refers to the cases in which system-generated questions contained ‘extraneous words’, which the post-editor removed, as seen in the example about photosynthesis. This finding matches the observation by Bernhard et al. (2012). The authors state that the reason for this is often the reported speech contained in the source sentences. However, the data observed as part of this research does not support this claim; subtitle text hardly ever contains reported speech and the reason that some system-generated questions needed extraneous words removing is because the source sentences are often complex-compound sentences, i.e. sentences with several independent clauses and one or more dependent clauses. For example, from the source sentence “Like all nuclear devices, Garwin’s exploited an extraordinary property of Einstein’s equation, that tiny amounts of matter would contain massive amounts of energy.” the unusable question “What does Garwin’s exploited an

extraordinary property of like all nuclear devices?” was generated. In this case, the post-editor was able to make a usable question by changing it to “What did Garwin exploit?”. While the framework developed as part of this thesis includes means to perform the simplification of complex sentences, it currently cannot simplify complex-compound sentences. The fact that the majority of edits is about the removal of extraneous words from system-generated questions highlights that correct handling of these sentences is clearly a crucial issue that needs to be further investigated as it could lead to higher quality questions and reduced post-editing time.

Substitutions constitute the second most frequent edit operation and a number of interesting observations can be made by looking at the system-generated questions and their post-edited counterparts. When post-editors made substitutions, there was often a change in question type. For example, from the source sentence “Surprisingly, much of the Anglo-Saxon art in British museums was actually discovered less than a century ago” the system-generated question “What was actually discovered less than a century ago?” was created. The post-editor changed the question from a “what” question, to a “when” question: “When was much of the Anglo-Saxon art in British museums discovered?”. In another example, the system question “What was the popular conception for centuries that the Dark Ages were?” was changed to a “why” question: “Why were the Dark Ages referred to as dark?”. In some cases, the changes made by substitutions changed the question significantly, as can be seen in this example: “What finally succeeded in producing enough energy to light up a few houses?” was changed to “What did a team in Oxford succeed in?”. At other times, the changes through substitution were only minor, as

in this example: “What did evolution by natural selection states?”, which was changed to “What does evolution by natural selection state?”.

Word and phrase shifts occur relatively infrequently in the dataset, although post-editor 2 made more use of phrase shifts than post-editor 1. An example for a word shift can be observed when comparing the system-generated question “Where been the biggest haul of Anglo-Saxon gold ever discovered found in a field?” to the post-edited version “Where has the biggest haul of Anglo-Saxon gold ever discovered been found?”. Here, the auxiliary “been” was moved from second to penultimate position. An example of a phrase shift can be observed in the question: “What was Norman art all about?”, in which ‘all about’ was shifted from its original position in the system-generated question “Who were all about building permanent, public art?”

It was hypothesized that the perceived post-editing effort score increases as the HTER value for a question increases. To test this, the Pearson product-moment correlation coefficient was calculated. Pearson’s correlation coefficient between two variables X and Y measures their linear dependence and is defined as the covariance of these two variables divided by the product of their standard deviations. Pearson’s correlation coefficient gives a value between +1 and -1, where 1 signifies total positive correlation, 0 signifies no correlation, and -1 signifies negative correlation. The formula to calculate Pearson’s correlation coefficient is as follows:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

The computed value for the HTER score and perceived post-editing effort is $r = 0.6543$. This indicates a moderate positive correlation between the perceived post-editing effort and HTER score. The correlation scores for perceived post-editing effort and ‘question improvement score’ (i.e. the score for usability after post-editing minus score for usability before editing) were also calculated. The value obtained in this case is 0.8048. This indicates a high positive correlation and implies that the more effort was required to post-edit a question, the more it improved through post-editing, or conversely, a larger amount of perceived post-editing is required to change a question from a low to a high usability score as compared to perfecting a question that already scores highly.

Against the expectations, there is hardly any correlation between perceived post-editing effort and post-editing time ($r=0.2417$). It was hypothesized that the amount of time required to post-edit a question would increase with growing perceived post-editing effort; however, the correlation coefficient computed shows that this is not the case.

6.1.4 Conclusion

The results indicate that post-editing questions is faster than the manual creation from scratch, with the average post-editing time per question being ~12 seconds, while the average time spent to generate a question manually was approximately 6 minutes.

Using the post-editing tool PET, two evaluators were asked to post-edit a set of 574 questions generated by the framework developed as part of this thesis, a state-of-the-art system and a set of questions that had been created manually by two human experts. The post-editors were asked to assign a score for each post-edited question, judging the quality

of questions before and after post-editing and perceived post-editing effort. The average usability score assigned before post-editing was 4.61 (4.5) for the manually created questions, 2.90 (3.12) for questions generated by system WLV and 2.53 (3.09) for questions generated by system CMU. After post-editing, the respective average scores were 4.61 (4.66), 4.16 (3.86), 3.84 (3.75).

The number of points a question improved was calculated by subtracting a question's score before post-editing from its score and after post-editing. The majority of post-edited questions improved by 3 points. It was also found that some questions can be transformed from 'unusable' to 'very good', highlighting that a QG system can also be a "trigger" post-editors' memory to generate very good questions from questions that were unusable before post-editing and that even seemingly unusable questions can be beneficial in the Question Generation process.

Using PET made it possible to compute HTER values for the post-edited questions which allowed to analyse how post-editors change system-generated questions. It was found that the most commonly used edit operations are deletions and substitutions and sample questions to exemplify the different edit processes were provided. Finally, the correlations between HTER, perceived post-editing effort, post-editing time and question improvement score were computed. A moderate positive correlation between perceived post-editing effort and HTER score can be observed. The value obtained for perceived post-editing effort and 'question improvement score' indicated a high positive correlation and implies that the more effort required to post-edit a question, the more it improved through post-editing. It was assumed that the amount of time required to post-edit a question would

increase with growing perceived post-editing effort; however, hardly any correlation was found.

The experiment provided interesting data about the post-editing process and the efficiency of the post-editing process. Unfortunately, the post-edited questions have not been employed in a user study. This will have to form part of future research.

6.2 Summary

This experiment was concerned with the efficiency of generating questions manually versus post-editing system-generated questions. It was found that post-editing questions is faster than the manual creation of questions by human experts. A number of usability statistics and other data, such as the edit-distance measure HTER, was gathered, which allowed to get a unique insight into the post-editing process. Feedback was collected from the post-editors about several aspects of the post-editing process. The combination of quantitative and qualitative data that was gathered in the experiments has provided a unique understanding of the Question Generation Process.

CHAPTER 7: CONCLUSION AND THESIS REVIEW

This Chapter summarises the research undertaken as part of thesis and looks at directions for future research. Section 7.1 presents a summary of the preceding Chapters and reviews the main research findings and contributions of this study. Section 7.2 discusses how the findings of this research could be applied in future work.

7.1 Review of contributions

The main aim of this thesis was to investigate how NLP tools and techniques can be harnessed to generate questions from multimedia learning materials, in particular videos, to be used in educational contexts and to support educators in the laborious and time-consuming task of generating assessment materials. As part of the research, several research questions will be answered:

1. How can Natural Language Processing tools and techniques be used when automatically generating questions from multimedia learning materials?
2. What are the characteristics of video documentaries and their subtitles and how do they affect the Question Generation process?
3. How can the effects of system-generated questions be evaluated in educational settings?
4. How do system-generated questions differ from those created by human experts?

In order to answer these research questions, this thesis was organised into two main parts. **Part 1** consists of Chapters 1 to 3 which present the background information for the research. These Chapters are used to describe the motivation for the research, explain common terminology, and provide a comprehensive review of existing approaches in Question Generation. **Part 2** consists of Chapters 4 to 6, which describe the proposed framework and experiments performed.

Chapter 1 presented a short introduction to the history of Natural Language Processing and its sub-discipline Question Generation. It highlighted the aims and contributions of this research and provided an overview of the thesis.

In **Chapter 2**, a detailed review of existing approaches to Question Generation in Natural Language Processing was given, partly answering research question 1 (it is also partly answered in Chapter 4). The value of questions in educational settings, the incidence of questions in classroom use, the different types of questions and their learning effect was discussed.

Chapter 3 was used to describe the characteristics of video documentaries and their use in educational settings. Different genres of documentaries were described and the benefits of using videos for teaching were explained. The characteristics of documentary subtitles were discussed, highlighted their advantages and challenges for Question Generation and also performed a qualitative analysis comparing subtitles to another text type. The findings of this Chapter answer research question 2.

In **Chapter 4**, a framework which automatically generates factual questions from video documentaries was presented. The methodology is explained in detail, thus drawing conclusions which provide answers to research question 1. The framework has undergone several cycles of developments, evaluations and improvements. **Chapter 4** also described 2 error analyses which helped identify error types in the automatically generated questions. Based on these error analyses, several improvements to the framework and its transformational rules were made.

Chapter 5 was used to discuss evaluation approaches in Question Generation as well as shared evaluation tasks. **Chapters 5 and 6** were then used to evaluate the framework. Two experiments were performed. The first experiment was a user study in which a novel evaluation methodology was used to test a variety of hypotheses about system-generated questions. The findings and novel evaluation methodology proposed in this experiment answer research question 3. The last research question was answered by the last experiment in which 2 human experts editors were assigned to post-edit a set of questions produced by the system developed as part of this thesis (system WLIV) and a state-of-the-art system (system CMU). A large number of qualitative and quantitative feedback about the post-editing process was gathered.

In summary, this thesis produced the following original contributions: The **first original contribution** is a Question Generation framework which generates shallow factual questions from video documentaries. The questions generated by the system can be used as a quick way of testing students' comprehension of what they have learned from the documentary. The system uses several readily available NLP tools to generate questions

from the subtitles accompanying a video documentary. Although several text-based QG systems had already been developed, these differ from the proposed approach in that the type of text they process is, by nature, different from documentary subtitles. There are different types of documentaries and certain features of the documentary subtitles affect the Question Generation process. Not all genres of documentary videos are suitable for factual Question Generation and some will yield a larger number of useful questions than others. Thus, the **second original contribution** is the analysis of the characteristics of documentary videos and their subtitles; the framework was adapted to be able to exploit these characteristics. In a user study described in Chapter 5, a novel evaluation methodology, the **third original contribution**, was proposed. The evaluation approach employed is a double-blind, randomised, controlled crossover study to investigate whether questions generated automatically by the system developed as part of this thesis and a state-of-the-art system can successfully be used to assist multimedia-based learning. The feasibility of using a QG system's output as pre-questions was examined, with different types of pre-questions used: text-based and with images. The psychometric parameters of the automatically generated questions by the two systems and of those generated manually were also compared. Specifically, the effect such pre-questions have on test-takers' performance on a comprehension test about a scientific video documentary were analysed. The discrimination power of the questions generated automatically against that of questions generated manually was compared. The results indicate that the presence of pre-questions (preferably with images) improves the performance of test-takers. They indicate that the psychometric parameters of the questions generated by system WLV are comparable to, if not better than those of the state-of-the-art system. In addition, the

ability to extract images from the video is a feature that is unique to system WL.V. Not only did the user study provide quantitative data about automatically and manually generated questions, qualitative data was also gathered in the form of user feedback, which provides an insight into how users perceived the quality of questions.

The evaluation method employed is a novel and unique approach to investigate a large number of research questions in one experiment, whilst at the same time eliminating variables that could influence the results, such as cross-group-performance and cross-question-performance.

In another experiment, the **fourth original contribution**, the productivity of questions in terms of time taken to generate questions manually vs. time taken to post-edit system-generated questions was analysed. A post-editing tool which allows tracking several statistics such as the number of keystrokes used, minimum edit distance and others was used. The quality of questions before and after post-editing was also analysed. The experiments provide a unique insight into the nature of automatically generated questions by combining quantitative analyses as well as qualitative feedback from users and human expert evaluators.

7.2 Future work

The preceding Section described the current state of research completed as part of this PhD thesis. The following Section describes potential areas of future research. While the implementation of different NLP resources into the framework would be one direction for future research, the main focus is on performing a number of experiments investigating different aspects.

7.2.1 Larger experiments

While a large amount of valuable data about the post-editing process in the experiment described in Chapter 6 was gathered, the post-edited questions were not used in a user study. It would be interesting to perform this study with a large number of participants, gather statistics about the psychometric parameters of the questions and to compare the results to previous experiments.

7.2.2 Different presentation of questions

There are a number of other potential experiments which would provide an insight into the usability of questions. For example, it would be worth investigating how test-takers' performance would be affected if the questions were displayed *during the video* instead of after having watched the video. Similarly to the experiment with pre-questions, one could hypothesize that doing so would catch test-takers' attention, letting them focus on the

important information in the documentary and consequently answer a larger proportion of questions correctly.

7.2.3 Images as distractor

A number of experiments could be undertaken focussing on the use of accompanying images. For example, one could explore the effect of using images from different sources. Currently, the images stem from the video itself, but it would be possible to use images from other sources, for example, Google Image Search or image databases, such as ImageNet²³. Another interesting experiment would be to use images as ‘distractors’, i.e. answers which are similar to the correct answer but nevertheless wrong, in a multiple choice question format. For example, if a question is asked about a person, then the test-taker would have to choose amongst several images, which could be extracted from the video or another source, for the correct answer.

7.2.4 Use of other NLP resources

As mentioned in previous Chapters, a major advantage of the framework is its modularity. It’s modular structure allows users to easily adapt the question rules and add NLP resources in the form of GATE plugins. Future work could explore how the use of other NLP resources and further work on the question rules improves the system output.

²³ <http://www-cs.stanford.edu/content/imagenet-large-scale-hierarchical-image-database>

BIBLIOGRAPHY

Anderson, L. W. and David R. Krathwohl, D. R. (2001) *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Boston, Allyn & Bacon.

Anon (2005) The Association for Computational Linguistics: What is Computational Linguistics?, [online] Available from: <http://www.aclweb.org/archive/misc/what.html> (Accessed 27 September 2013).

Aziz, W. and Specia, L. (2012) PET: a Tool for Post-editing and Assessing Machine Translation., In *The 16th Annual Conference of the European Association for Machine Translation, EAMT '12*, Trento, Italy, [online] Available from: http://www.lrec-conf.org/proceedings/lrec2012/pdf/985_Paper.pdf (Accessed 27 September 2013).

Berk, R. A. (2009) Multimedia teaching with video clips: TV, movies, YouTube, and mtvU in the college classroom, *International Journal of Technology in Teaching and Learning*, **5**(1), pp. 1–21.

Bernhard, D., Viron, L. De, Moriceau, V. and Tannier, X. (2012) Question Generation for French: Collating Parsers and Paraphrasing Questions, *Dialogue & Discourse*, **3**(2), pp. 43–74, [online] Available from: <http://elanguage.net/journals/dad/article/download/2151/2833> (Accessed 7 October 2013).

Bethard, S., Nielsen, R. D., Martin, J. H., Ward, W. and Palmer, M. (2007) Semantic Integration in Learning from Text, In *Proceedings of Machine Reading AAAI Spring Symposium*.

Bloom, B. S. (1956) *Taxonomy of educational objectives, Handbook I: The cognitive domain*, NY, Guilford.

Burstein, J. and Attali, Y. (2006) Automated Essay Scoring With E-rater V. 2.0, *The Journal of Technology, Learning and Assessment*, **4**(3), [online] Available from: http://origin-www.ets.org/Media/Products/e-rater/erater_IAEA.pdf (Accessed 20 September 2013).

Chali, Y. and Hasan, S. A. (2012) Towards Automatic Topical Question Generation, In *Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai, India, pp. 475-492

Chen, W., Aist, G. and Mostow, J. (2009) Generating Questions Automatically from Informational Text, In *Workshop Proceedings of the AIED 2009: 14th International Conference on Artificial Intelligence in Education*, pp. 17–24.

Chong, M. and Specia, L. (2012) Linguistic and Statistical Traits Characterising Plagiarism., *COLING (Posters)*, **2**(December 2012), pp. 195–204, [online] Available from: <http://pers-www.wlv.ac.uk/~ex0233/pub/COLINGpaper.pdf> (Accessed 20 September 2013).

Churches, A. (2009) Blooms's Digital Taxonomy, [online] Available from: <http://edorigami.wikispaces.com/Bloom's+Digital+Taxonomy> (Accessed 23 September 2013).

Cohen, J. (1960) A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, **20**, pp. 37–46.

Commeyras, M. and Sumner, G. (1998) Literature questions children want to discuss: What teachers and students learned in a second-grade classroom, *The Elementary School Journal*, **99**, pp. 129–152.

Cotton, K. (2001) Classroom questioning, *School improvement research series*, [online] Available from:

http://rsd.schoolwires.com/145410515152938173/lib/145410515152938173/Classroom_Questioning_by_Cotton.pdf (Accessed 24 September 2013).

Deng, J., Li, K., Do, M., Su, H. and Fei-Fei, L. (2009) Construction and Analysis of a Large Scale Image Ontology, *Vision Sciences Society (VSS)*.

Díaz Cintas, J. and Remael, A. (2007) *Audiovisual Translation: Subtitling*, Manchester, St. Jerome.

Dunkin, M. J. and Biddle, B. J. (1974) *The study of teaching*, New York, Holt, Rinehart and Winston.

Educause (2012) 7 Things You Should Know About Flipped Classroom, [online] Available from: <http://net.educause.edu/ir/library/pdf/eli7081.pdf> (Accessed 27 September 2013).

Galassi, J. P., Gall, M. D., Dunning, B. and Banks, H. (1974) The use of written versus videotape instruction to train teachers in questioning skills, *Journal of Experimental Education*, **43**, pp. 16–23.

Gall, M. D. (1984) Synthesis of research on teachers' questioning, *Educational Leadership*, **42**, pp. 40–47.

Gall, M. D. (1970) The use of questions in teaching, *Review of Educational Research*, **40**, pp. 707–721.

Gall, M. D. and Rhody, T. (1978) Review of research on questioning techniques, In *Questions, questioning techniques, and effective teaching.*, Wilen, W. W. (ed.), Washington, DC: National Education Association.

Gardner, H. (2000) Can technology exploit the many ways of knowing?, In *The digital classroom: How technology is changing the way we teach and learn*, Gordon, D. T. (ed.), Cambridge, MA, President and Fellows of Harvard College, pp. 32–35.

Gates, D. M. (2008) Automatically Generating Reading Comprehension Look-Back Strategy, *Workshop on the Question Generation Shared Task and Evaluation Challenge*, Carnegie Mellon University, [online] Available from:

<http://www.lti.cs.cmu.edu/Research/Tech Reports/CMU-LTI-08-011 Automatically Generating Reading Comprehension Look-Back Strategy Questions from Expository Texts.pdf> (Accessed 27 March 2010).

Graesser, A. C., McNamara, D. S., Louwerse, M. M. and Cai, Z. (2004) Coh-Metrix: Analysis of text on cohesion and language., *Behavior Research Methods, Instruments, and Computers*, **36**, pp. 193–202.

Graesser, A. C., Person, N. and Huber, J. (1992) Mechanisms that generate questions, In *Questions and information systems*, Lauer, T., Peacock, E., and Graesser, A. C. (eds.), Hillsdale, NJ, Erlbaum.

Graesser, A., Rus, V. and Cai, Z. (2008) Question classification schemes, In *Workshop on the Question Generation Shared Task and Evaluation Challenge*, [online] Available from: <http://www.cs.memphis.edu/~vrus/questiongeneration/16-GraesserEtAl-QG08.pdf> (Accessed 27 September 2013).

Guszk, F. J. (1967) Teacher questioning and reading, *The Reading Teacher*, **21**, pp. 227–234.

Hébert, S. and Peretz, I. (1997) Recognition of music in long-term memory: Are melodic and temporal patterns equal partners?, *Memory and Cognition*, **25**, pp. 518–533.

Heilman, M. (2011) Automatic factual question generation from text, [online] Available from: <http://www2.lti.cs.cmu.edu/Research/Thesis/heilman, michael.pdf> (Accessed 24 September 2013).

Heilman, M. and Smith, N. (2010) Good question! statistical ranking for question generation, *Proc. of NAACL-HLT*, [online] Available from:

<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.162.2267> (Accessed 1 September 2010).

Hinrichs, P. D. E. W. (2005) Introduction to Computational Linguistics, [online] Available from: <http://www.sfs.uni-tuebingen.de/~fr/teaching/ws05-06/icl/slides/lecture2.pdf> (Accessed 27 September 2013).

Ignatova, K., Bernhard, D. and Gurevych, I. (2008) Generating High Quality Questions from Low Quality Questions, In *Workshop on the Question Generation Shared Task and Evaluation Challenge*.

Isaacs, G. (1994) *Multiple choice testing: a guide to the writing of multiple choice tests and to their analysis*, Campbelltown, HERDSA.

Jourdain, R. (1997) *Music, the brain, and ecstasy: How music captures the imagination*, NY, Avon Press.

Jurafsky, D. and Martin, J. H. (2008) *Speech and Language Processing*, 2nd ed, Pearson Prentice Hall.

Libermann, M. (2008) Language Log: HTER, [online] Available from: <http://languagelog.ldc.upenn.edu/nll/?p=193> (Accessed 27 September 2013).

Liddy, E. (2003) Natural Language Processing, In *Encyclopedia of Library and Information Science*, 2nd ed, NY, Marcel Decker, Inc., [online] Available from: <http://www.columbia.edu/itc/hs/medinfo/g6080/misc/articles/spyns.pdf> (Accessed 26 September 2013).

Marneffe, M.-C. de and Manning, C. D. (2008) The Stanford typed dependencies representation, In *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*.

Miller, M. (1997) *Brain styles: Change your life without changing who you are.*, NY, Simon and Schuster.

Mitkov, R. (2002) *Anaphora Resolution*, Longman.

Mitkov, R. and Ha, L. A. (2003) Computer-aided generation of multiple-choice tests, *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing* -, Morristown, NJ, USA, Association for Computational Linguistics, pp. 17–22, [online] Available from: <http://portal.acm.org/citation.cfm?doid=1118894.1118897> (Accessed 27 September 2013).

Mitkov, R., Ha, L. A. and Karamanis, N. (2006) A computer-aided environment for generating multiple-choice test items, *Natural Language Engineering*, **12**(02), p. 177, [online] Available from: http://www.journals.cambridge.org/abstract_S1351324906004177 (Accessed 27 September 2013).

Mosenthal, P. (1996) Understanding the strategies of document literacy and their conditions of use, *Journal of Educational Psychology*, **88**, pp. 314–332.

Mostow, J. and Chen, W. (2009) Generating Instruction Automatically for the Reading Strategy of Self-Questioning, In *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, IOS Press, pp. 465–472, [online] Available from: <http://www.cs.cmu.edu/afs/cs.cmu.edu/Web/People/listen2/pdfs/AIED2009-self-question-final-A4.pdf> (Accessed 27 September 2013).

Nielsen, R. D., Buckingham, J., Knoll, G., Marsh, B. and Palen, L. (2008) A Taxonomy of Questions for Question Generation, In *Workshop on the Question Generation Shared Task and Evaluation Challenge*.

Nystrand, M. (1997) *Opening Dialogue: Understanding the dynamics of language and learning in the classroom.*, New York, Teachers College Press.

Patrick, J., Wang, Y. and Budd, P. (2006) Automatic Mapping Clinical Notes to Medical Terminologies, *Australasian Language Technology ...*, pp. 75–82, [online] Available from: <http://www.aclweb.org/anthology/U/U06/U06-1.pdf#page=83> (Accessed 24 September 2013).

Pearson, P. D., & Johnson, D. D. (1978) *Teaching reading comprehension.*, 1978, New York, Holt, Rinehart and Winston.

Polk, M. and Kertesz, A. (1993) Music and language in degenerative disease of the brain, *Brain and Cognition*, **22**(1), pp. 98–117.

Pradhan, S., Ward, W., Hacioglu, K., Martin, J. and Jurafsky, D. (2004) Shallow semantic parsing using support vector machines, In *Proceedings of HLT/NAACL-2004*, [online] Available from: <http://acl.ldc.upenn.edu/N/N04/N04-1030.pdf> (Accessed 27 September 2013).

Raphael, T. E. and Wonnacott, C. A. (1985) Heightening fourth-grade students' sensitivity to sources of information for answering comprehension questions, *Reading Research Quarterly*, **20**, pp. 282–296.

Redfield, D. L. and Rousseau, E. W. (1981) A meta-analysis of experimental research on teacher questioning behavior., *Review of Educational Research*, **51**, pp. 237–245.

Van Rijsbergen, C. J. (1979) *Information Retrieval*, 2nd ed, Butterworth.

Rosenshine, B. (1976) *Classroom instruction.*, Chicago, University of Chicago Press.

Rosenshine, B. (1971). (1971) *Teaching behaviors and student achievement*, *Teaching behaviors and student achievement*. Windsor: NFER/Nelson., Windsor, NFER/Nelson.

- Rus, V. (n.d.) The Question Generation Research Website, [online] Available from: <http://questiongeneration.org/> (Accessed 25 September 2013).
- Rus, V., Cai, Z. and Graesser, A. C. (2007) Experiments on Generating Questions About Facts, In *Proceedings of CICLing*, Gelbukh, A. F. (ed.), Springer, pp. 444–455.
- Rus, V., Graesser, A. and Cai, Z. (2008) Question Generation: Example of A Multi-year Evaluation Campaign, In *Workshop on the Question Generation Shared Task and Evaluation Challenge*, [online] Available from: <http://www.cs.memphis.edu/~vrus/questiongeneration/5-RusEtAl-QG08.pdf> (Accessed 20 September 2013).
- Samson, G. E., Strykowski, B., Weinstein, T. and Walberg, H. J. (1987) The effects of teacher questioning levels on student achievement: A quantitative synthesis., *Journal of Educational Research*, **80**, pp. 290–295.
- Schlaug, G., Jancke, L., Haug, Y., Staiger, J. and Steinmetz, H. (1995) Increased corpus callosum size in musicians, *Neuropsychologia*, **33**(8), pp. 1047–1055.
- Silveira, N. (2008) Towards a Framework for Question Generation, In *Workshop on the Question Generation Shared Task and Evaluation Challenge*, [online] Available from: <http://www.cs.memphis.edu/~vrus/questiongeneration/4-Silveira-QG08.pdf> (Accessed 27 September 2013).
- Sirotnik, K. A. (1983) What you see is what you get— consistency, persistency, and mediocrity in classrooms. 53, *Harvard Educational Review*, **53**(5), pp. 16–31.
- Skalban, Y. (2009) Improving the output of a multiple-choice test generator: analysis and proposals, University of Wolverhampton.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2006) A study of translation edit rate with targeted human annotation, In *Proceedings of Association for Machine Translation in the Americas*, pp. 223–231, [online] Available from:

<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.129.4369> (Accessed 6 October 2013).

Stevens, R. (1912) *The question as a measure of efficiency in instruction: A critical study of class-room practice.*, New York, Teachers College Press.

Taylor, B. M., Pearson, P. D., P. D. S. and Rodriguez, M. C. (2003) Reading growth in high-poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. 104, *Elementary School Journal*, pp. 3–28.

Taylor, B. M., Pearson, P. D., P. D. S. and Rodriguez, M. C. (2005) The CIERA school change framework: An evidence-based approach to professional development and school reading improvement., *Reading Research Quarterly*, **40**, pp. 40–69.

Vanderwende, L. (2008) The importance of being important, In *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*.

Voorhees, E. M. and Harman, D. K. (2000) NIST Special Publication 500-246: The Eighth Text Retrieval Conference (TREC-8), In *Proceedings of the 8th Text Retrieval Conference (TREC-8). NIST Special Publication*.

Weaver, W. (1949) Translation, In *Machine translation of languages: fourteen essays*, Booth, W. N. and Locke, D. A. (eds.), (Technology Press of the Massachusetts Institute of Technology, Cambridge, Mass., and John Wiley & Sons, Inc., New York, 1955, pp. 15–23, [online] Available from: <http://wrap.warwick.ac.uk/id/eprint/42415> (Accessed 20 September 2013).

Winne, P. H. (1979) Experiments relating teachers' use of higher cognitive questions to student achievement., *Review of Educational Research*, **49**, pp. 13–49.

APPENDIX A: PREVIOUSLY PUBLISHED WORK

Some of the work described in this thesis has been previously published in the proceedings of peer-reviewed international conferences. The work has been extended or modified to adapt to the context of this thesis. This appendix provides the list of the previously published work as well as a brief explanation of their contribution to this thesis:

- Skalban, Y., Ha, L. A., Specia, L., Mitkov, R. (2012). Automatic Question Generation in multimedia-based learning. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING2012)*, Mumbai, India. pp. 1151–1160

This paper presents the user study described in Chapter 5, in which the feasibility of system-generated questions as ‘pre-questions’ (text-based and image-based) was analysed. The experiment also served to compare psychometric parameters of questions generated by the framework developed as part of this thesis, a state-of-the-art system and questions generated by human experts.

APPENDIX B: COH-METRIX OUTPUT FOR SUBTITLES AND WIKIPEDIA ARTICLES FOR TWO GENRES

Coh-Metrix Index	History Subtitles	Science Subtitles	Biography Subtitles	Food Subtitles	Linguistics Subtitles	History Wikipedia	Science Wikipedia	Biography Wikipedia	Food Wikipedia	Linguistics Wikipedia
DESPC ‘Paragraph count, number of paragraphs’	1	1	1	1	1	15	16	16	39	24
DESSC ‘Sentence count, number of sentences’	129	94	126	163	77	82	62	78	79	75
DESWC ‘Word count, number of words’	1598	1549	1604	1628	1553	1579	1547	1643	1379	1716
DESPL ‘Paragraph length, number of sentences in a paragraph, mean’	1598	1549	1604	1628	1553	105.267	96.688	102.688	35.359	71.5
DESPLd ‘Paragraph length, number of sentences in a paragraph, standard deviation’	0	0	0	0	0	2.85	1.746	1.893	1.597	2.309
DESSL ‘Sentence length, number of words, mean’	12.388	16.479	12.73	9.988	20.169	19.256	24.952	21.064	17.456	22.88
DESSLd ‘Sentence length, number of words, standard deviation’	7.911	9.216	7.234	6.854	10.544	10.623	13.13	14.445	14.48	15.669
DESWLsy ‘Word length, number of syllables, mean’	1.437	1.478	1.433	1.302	1.579	1.548	1.707	1.579	1.808	1.729
DESWLsyd ‘Word length, number of syllables, standard deviation’	0.771	0.812	0.784	0.645	0.847	0.867	0.967	0.93	1.002	0.981
DESWLlt ‘Word length, number of letters, mean’	4.385	4.555	4.38	3.993	4.807	4.76	5.073	4.684	5.465	5.154
DESWLltd ‘Word length, number of letters, standard deviation’	2.399	2.466	2.347	2.112	2.465	2.596	2.752	2.663	2.837	2.783

PCNARz 'Text Easability PC Narrativity, z score'	0.345	-0.318	0.272	0.46	-0.961	-0.39	-1.138	-0.407	-1.675	-1.186
PCNARp 'Text Easability PC Narrativity, percentile'	63.31	37.83	60.64	67.72	16.85	34.83	12.92	34.46	4.75	11.9
PCSYNz 'Text Easability PC Syntactic simplicity, z score'	0.675	0.296	0.605	0.853	-0.104	-0.31	-0.529	-0.262	0.237	-0.552
PCSYNp 'Text Easability PC Syntactic simplicity, percentile'	74.86	61.41	72.57	80.23	46.02	37.83	30.15	39.74	59.1	29.12
PCCNCz 'Text Easability PC Word concreteness, z score'	-0.331	-0.54	-0.155	-0.734	0.065	1.041	-1.242	1.027	0.358	-0.472
PCCNCp 'Text Easability PC Word concreteness, percentile'	37.07	29.46	44.04	23.27	52.39	85.08	10.75	84.61	63.68	31.92
PCREFz 'Text Easability PC Referential cohesion, z score'	-0.846	-0.999	-1.235	-1.065	-0.985	-1.125	0.466	-0.914	-0.514	-0.267
PCREFp 'Text Easability PC Referential cohesion, percentile'	20.05	15.87	10.93	14.46	16.35	13.14	67.72	18.14	30.5	39.74
PCDCz 'Text Easability PC Deep cohesion, z score'	0.363	0.691	0.359	-0.233	-0.468	-0.464	1.642	0.366	0.079	-0.371
PCDCp 'Text Easability PC Deep cohesion, percentile'	64.06	75.49	63.68	40.9	32.28	32.28	94.95	64.06	52.79	35.57
PCVERBz 'Text Easability PC Verb cohesion, z score'	0.868	0.584	0.291	0.388	0.218	-0.557	-0.273	-0.204	-1.585	0.263
PCVERBp 'Text Easability PC Verb cohesion, percentile'	80.51	71.9	61.41	64.8	58.32	29.12	39.36	42.07	5.71	60.26
PCCONNz 'Text Easability PC Connectivity, z score'	-1.63	-1.098	-0.933	-1.963	-3.951	-1.537	-1.241	-1.828	-2.613	-2.231
PCCONNp 'Text Easability PC Connectivity, percentile'	5.16	13.79	17.62	2.5	0	6.3	10.75	3.44	0.45	1.29
PCTEMPz 'Text Easability PC Temporality, z score'	0.008	-1.483	-0.244	0.991	-0.099	0.566	0.915	0.633	0.203	0.346
PCTEMPp 'Text Easability PC Temporality, percentile'	50	6.94	40.52	83.89	46.41	71.23	81.86	73.57	57.93	63.31

CRFNO1 'Noun overlap, adjacent sentences, binary, mean'	0.141	0.247	0.152	0.117	0.289	0.222	0.689	0.416	0.487	0.486
CRFAO1 'Argument overlap, adjacent sentences, binary, mean'	0.391	0.376	0.328	0.272	0.342	0.42	0.721	0.506	0.5	0.554
CRFSO1 'Stem overlap, adjacent sentences, binary, mean'	0.18	0.323	0.176	0.142	0.474	0.309	0.77	0.455	0.59	0.622
CRFNOa 'Noun overlap, all sentences, binary, mean'	0.079	0.186	0.122	0.046	0.262	0.195	0.455	0.323	0.361	0.404
CRFAOa 'Argument overlap, all sentences, binary, mean'	0.257	0.271	0.249	0.156	0.313	0.293	0.519	0.407	0.386	0.458
CRFSOa 'Stem overlap, all sentences, binary, mean'	0.109	0.245	0.139	0.065	0.404	0.284	0.547	0.346	0.483	0.514
CRFCWO1 'Content word overlap, adjacent sentences, proportional, mean'	0.108	0.07	0.076	0.099	0.067	0.055	0.134	0.088	0.139	0.087
CRFCWO1d 'Content word overlap, adjacent sentences, proportional, standard deviation'	0.145	0.102	0.122	0.186	0.073	0.082	0.108	0.113	0.188	0.104
CRFCWOa 'Content word overlap, all sentences, proportional, mean'	0.061	0.045	0.046	0.043	0.063	0.038	0.081	0.05	0.088	0.061
CRFCWOad 'Content word overlap, all sentences, proportional, standard deviation'	0.105	0.08	0.086	0.106	0.082	0.067	0.093	0.075	0.138	0.08
CRFANP1 'Anaphor overlap, adjacent sentences'	0.453	0.333	0.328	0.272	0.25	0.346	0.098	0.351	0.051	0.068
CRFANPa 'Anaphor overlap, all sentences'	0.155	0.07	0.115	0.105	0.035	0.073	0.016	0.072	0.007	0.01
LSASS1 'LSA overlap, adjacent sentences, mean'	0.171	0.162	0.101	0.133	0.207	0.155	0.436	0.182	0.338	0.28
LSASS1d 'LSA overlap, adjacent sentences, standard deviation'	0.185	0.149	0.152	0.199	0.134	0.159	0.242	0.188	0.244	0.206
LSASSp 'LSA overlap, all sentences'	0.053	0.102	0.039	0.041	0.12	0.146	0.435	0.18	0.262	0.267

in paragraph, mean'										
LSASSpd 'LSA overlap, all sentences in paragraph, standard deviation'	0.116	0.139	0.095	0.103	0.131	0.142	0.218	0.179	0.159	0.191
LSAPP1 'LSA overlap, adjacent paragraphs, mean'	0	0	0	0	0	0.279	0.441	0.252	0.502	0.407
LSAPP1d 'LSA overlap, adjacent paragraphs, standard deviation'	0	0	0	0	0	0.246	0.283	0.158	0.231	0.247
LSAGN 'LSA given/new, sentences, mean'	0.308	0.306	0.261	0.294	0.327	0.312	0.439	0.31	0.415	0.367
LSAGNd 'LSA given/new, sentences, standard deviation'	0.111	0.101	0.098	0.13	0.092	0.103	0.132	0.103	0.162	0.117
LDTTRe 'Lexical diversity, type-token ratio, content word lemmas'	0.641	0.573	0.628	0.612	0.634	0.668	0.509	0.675	0.602	0.563
LDTTRa 'Lexical diversity, type-token ratio, all words'	0.391	0.345	0.377	0.344	0.407	0.409	0.338	0.442	0.418	0.374
LDMTLD 'Lexical diversity, MTLTLD, all words'	81.645	97	96.216	71.712	86.94	97.975	72.913	101.669	93.33	68.784
LDVOCD 'Lexical diversity, VOCD, all words'	114.298	105.05	125.34	104.701	96.244	107.084	92.603	129.804	108.128	77.065
CNCAll 'All connectives incidence'	81.352	78.76	71.696	73.096	103.026	82.331	85.326	93.122	97.897	90.909
CNCCaus 'Causal connectives incidence'	25.031	32.924	28.055	21.499	19.961	15.833	38.785	17.651	26.106	16.9
CNCLogic 'Logical connectives incidence'	35.044	36.152	28.678	33.17	28.976	21.533	44.602	29.215	23.93	26.224
CNCADC 'Adversative and contrastive connectives incidence'	13.141	9.684	9.352	13.514	16.742	8.866	15.514	9.738	9.427	11.655
CNCTemp 'Temporal connectives incidence'	20.651	12.266	18.08	11.671	11.59	20.899	16.16	29.215	15.228	23.893
CNCTempx 'Expanded temporal connectives incidence'	16.896	18.722	17.456	16.585	28.332	16.466	16.807	15.216	20.305	18.648

CNCAdd ‘Additive connectives incidence’	41.302	34.216	31.172	39.926	72.762	43.699	36.846	45.648	55.838	55.944
CNCPos ‘Positive connectives incidence’	0	0	0	0	0	0	0	0	0	0
CNCNeg ‘Negative connectives incidence’	0	0	0	0	0	0	0	0	0	0
SMCAUSv ‘Causal verb incidence’	31.289	25.178	34.289	40.541	25.757	26.599	21.978	20.085	28.281	18.648
SMCAUSvp ‘Causal verbs and causal particles incidence’	44.431	43.254	44.888	52.826	33.484	29.132	40.724	26.78	34.808	25.641
SMINTEp ‘Intentional verbs incidence’	33.166	16.785	24.314	28.87	15.454	22.799	4.525	17.042	7.977	6.993
SMCAUSr ‘Ratio of casual particles to causal verbs’	0.412	0.7	0.304	0.299	0.293	0.093	0.829	0.324	0.225	0.364
SMINTER ‘Ratio of intentional particles to intentional verbs’	0.648	1.407	0.95	0.646	1	0.568	4.375	0.724	2.583	1.308
SMCAUSlsa ‘LSA verb overlap’	0.103	0.103	0.069	0.079	0.092	0.046	0.115	0.062	0.052	0.067
SMCAUSwn ‘WordNet verb overlap’	0.598	0.531	0.512	0.456	0.599	0.395	0.55	0.539	0.225	0.452
SMTEMP ‘Temporal cohesion, tense and aspect repetition, mean’	0.844	0.688	0.836	0.935	0.822	0.92	0.902	0.935	0.859	0.885
SYNLE ‘Left embeddedness, words before main verb, mean’	2.938	4.734	2.833	1.853	6.039	4.537	7.984	3.231	4.114	4.773
SYNNP ‘Number of modifiers per noun phrase, mean’	0.794	0.794	0.685	0.695	1.137	0.91	1.048	0.916	1.316	1.192
SYNMEDpos ‘Minimal Edit Distance, part of speech’	0.656	0.68	0.657	0.678	0.683	0.671	0.638	0.706	0.638	0.666
SYNMEDwrd ‘Minimal Edit Distance, all words’	0.895	0.915	0.916	0.897	0.907	0.908	0.869	0.899	0.893	0.866
SYNMEDlem ‘Minimal Edit Distance, lemmas’	0.872	0.891	0.892	0.873	0.891	0.895	0.845	0.891	0.879	0.859
SYNSTRUTa ‘Sentence syntax	0.185	0.108	0.133	0.126	0.091	0.075	0.066	0.066	0.128	0.064

similarity, adjacent sentences, mean'										
SYNSTRUt 'Sentence syntax similarity, all combinations, across paragraphs, mean'	0.143	0.104	0.119	0.104	0.09	0.077	0.066	0.086	0.075	0.059
DRNP 'Noun phrase density, incidence'	387.985	363.46	396.509	358.108	383.773	409.12	361.991	390.749	385.787	409.091
DRVP 'Verb phrase density, incidence'	181.477	249.193	217.581	209.459	137.798	170.994	166.128	164.942	123.278	114.219
DRAP 'Adverbial phrase density, incidence'	34.418	30.342	31.172	42.383	35.415	15.833	30.381	23.737	26.106	28.555
DRPP 'Preposition phrase density, incidence'	109.512	101.356	107.232	77.396	133.29	141.229	146.089	129.032	131.255	155.012
DRPVAL 'Agentless passive voice density, incidence'	6.258	10.975	4.988	5.528	3.863	8.866	14.221	8.521	10.152	8.741
DRNEG 'Negation density, incidence'	8.761	9.684	11.222	9.214	5.151	1.9	5.171	3.652	2.175	2.331
DRGERUND 'Gerund density, incidence'	5.632	12.912	13.716	15.971	14.81	12.033	15.514	12.781	12.328	9.907
DRINF 'Infinitive density, incidence'	9.387	29.697	9.975	12.285	11.59	14.566	10.989	9.13	5.801	8.741
WRDNOUN 'Noun incidence'	260.326	240.155	260.598	196.561	311.656	310.956	281.189	332.928	372.009	351.399
WRDVERB 'Verb incidence'	127.033	128.471	137.158	105.038	99.806	126.663	107.304	116.251	92.821	82.751
WRDADJ 'Adjective incidence'	70.088	83.28	69.202	89.681	98.519	68.398	130.575	57.212	146.483	99.651
WRDADV 'Adverb incidence'	58.824	54.874	54.863	84.153	51.514	22.8	51.713	37.127	44.959	41.376
WRDPRO 'Pronoun incidence'	98.248	49.709	97.257	106.88	41.211	62.698	10.343	51.735	10.152	9.907
WRDPR1s 'First person singular pronoun incidence'	18.148	2.582	12.469	19.656	7.727	0.633	0	1.826	0	0
WRDPR1p 'First person plural pronoun incidence'	10.638	13.557	13.092	14.128	9.659	0	1.293	0.609	0	0.583
WRDPR2 'Second person pronoun	6.258	4.519	13.092	26.413	2.576	0	0	0	0	0

incidence'										
WRDPRP3s 'Third person singular pronoun incidence'	45.682	2.582	24.938	6.757	3.22	36.099	0	34.084	0	0
WRDPRP3p 'Third person plural pronoun incidence'	8.761	7.747	10.599	14.742	7.083	22.166	3.232	10.347	2.175	1.166
WRDFRQc 'CELEX word frequency for content words, mean'	2.467	2.238	2.317	2.336	2.075	2.087	1.996	2.116	1.958	2.068
WRDFRQa 'CELEX Log frequency for all words, mean'	3.172	3.002	3.065	3.063	2.991	3.083	2.846	2.987	2.849	3.079
WRDFRQmc 'CELEX Log minimum frequency for content words, mean'	1.686	1.061	1.382	1.675	0.874	0.97	0.779	1.38	1.143	1.248
WRDAOAc 'Age of acquisition for content words, mean'	324.521	355.207	336.13	315.008	369.862	348.431	387.174	359.476	382.012	396.341
WRDFAMc 'Familiarity for content words, mean'	577.646	573.384	577.387	570.186	560.094	567.01	565.39	566.425	536.547	561.883
WRDCNCc 'Concreteness for content words, mean'	371.235	369.693	377.368	383.381	373.491	407.664	348.405	405.055	408.009	360.952
WRDIMGc 'Imagability for content words, mean'	411.217	400.762	411.887	408.824	404.256	440.229	385.342	432.681	425.498	395.45
WRDMEAc 'Meaningfulness, Colorado norms, content words, mean'	431.025	426.511	437.847	421.27	427.98	443.563	401.253	440.917	434.782	420.706
WRDPOLc 'Polysemy for content words, mean'	3.655	4.34	4.087	4.049	3.343	3.077	4.164	3.614	3.15	3.198
WRDHYPn 'Hypernymy for nouns, mean'	5.251	6.579	6.226	6.037	6.369	5.047	6.509	5.508	6.091	5.754
WRDHYPv 'Hypernymy for verbs, mean'	1.496	1.509	1.538	1.249	1.593	1.648	1.536	1.766	1.481	1.409
WRDHYPnv 'Hypernymy for nouns and verbs, mean'	1.453	1.744	1.73	1.256	1.933	1.662	1.937	1.9	2.27	2.035

RDFRE 'Flesch Reading Ease'	72.691	65.07	72.682	86.548	52.78	56.329	37.097	51.872	36.16	37.338
RDFKGL 'Flesch-Kincaid Grade level'	6.198	8.277	6.284	3.669	10.908	10.186	14.284	11.257	12.552	13.735
RDL2 'Coh-Metrix L2 Readability'	26.724	14.939	18.542	19.73	10.111	8.772	10.324	10.601	13.553	9.361

APPENDIX C: QUESTIONS GIVEN TO THE HUMAN EXPERT EVALUATORS

Questions for “Making of Modern Britain”

Please assign one of the following scores:

- 1 - Unusable**
- 2 – Usable with major revision**
- 3 – Usable with minor revision**
- 4 - Usable without revision**

1 - Rule1

Candidate: On 15th July 1906, a young woman called Adela Pankhurst made her way to a park in Manchester called Boggart Hole Clough.

Question: Who called made her way to a park in Manchester called Boggart Hole Clough a young woman on 15th July 1906?

Answer: Adela Pankhurst

2 - Rule1

Candidate: Adela Pankhurst and her fellow speakers were standing here at the bottom of the hillside.

Question: Who were standing here at the bottom of the hillside and her fellow speakers?

Answer: Adela Pankhurst

3 - Rule1

Candidate: Sir Henry Campbell-Bannerman had a radical streak but looked and sounded like an elderly sea lion.

Question: Who had a radical streak but looked and sounded like an elderly sea lion?

Answer: Sir Henry Campbell-Bannerman

4 - Rule1

Candidate: Sir Henry once declared, “Personally, I am a great believer in bed, “in constantly keeping horizontal.

Question: Who declared, “Personally, I am a great believer in bed, “in constantly keeping horizontal once?

Answer: Sir Henry

5 - Rule1

Candidate: Now, Sir Henry wasn’t a well man, as well as being rather an idle one.

Question: Who wasn’t a well man, as well as being rather an idle one now?

Answer: Sir Henry

6 - Rule1

Candidate: Herbert Asquith was ridiculously clever, a self-made statesman whose sternly sober face hid a wildly romantic heart.

Question: Who was ridiculously clever, a self-made statesman whose sternly sober face hid a wildly romantic heart?

Answer: Herbert Asquith

7 - Rule1

Candidate: Asquith took over.

Question: Who took over?

Answer: Asquith

8 - Rule1

Candidate: And a motor car salesmen called Claude Grahame-White was one of thousands of visitors who came to gape.

Question: Who called was one of thousands of visitors who came to gape a motor car salesmen?

Answer: Claude Grahame-White

9 - Rule1

Candidate: By the time Grahame-White was woken up, He was already an hour behind .

Question: Who was already an hour behind He?

Answer: Grahame-White

10 - Rule1

Candidate: Soon after the Liberal landslide, a quizzical, long-faced aristocrat, the 19th Lord Willoughby de Broke, was clip-clopping around his estate, reflecting on the joys of fox-hunting.

Question: Who was clip-clopping around his estate, reflecting on the joys of fox-hunting a quizzical, long-faced aristocrat, the 19th de Broke soon after the Liberal landslide?

Answer: Lord Willoughby

11 - Rule1

Candidate: David Lloyd George was the most radical Chancellor this country had ever known.

Question: Who was the most radical Chancellor this country had ever known David Lloyd?

Answer: George

12 - Rule1

Candidate: Lloyd George wanted to pay for welfare reforms by making massive cuts in defence spending.

Question: Who wanted to pay for welfare reforms by making massive cuts in defence spending?

Answer: Lloyd George

13 - Rule1

Candidate: To pay for both dreadnoughts and welfare, Lloyd George announced an increase in estate duties - a huge blow to the wealthy.

Question: Who announced an increase in estate duties - a huge blow to the wealthy?

Answer: Lloyd George

14 - Rule1

Candidate: Lord Rosebery, an immensely wealthy landowner who'd also been Liberal prime minister, described the People's Budget as pure socialism...

Question: Who described the People's Budget as pure socialism.. , an immensely wealthy landowner who'd also been Liberal prime minister?

Answer: Lord Rosebery

15 - Rule1

Candidate: Many who heard Lloyd George in his prime said he was the greatest orator British politics ever produced.

Question: Who said he was the greatest orator British politics ever produced Many who heard in his prime?

Answer: Lloyd George

16 - Rule1

Candidate: Lloyd George was merciless.

Question: Who was merciless?

Answer: Lloyd George

17 - Rule1

Candidate: To his fervent admirers, Lloyd George was the Welsh wizard.

Question: Who was the Welsh wizard to his fervent admirers?

Answer: Lloyd George

18 - Rule1

Candidate: He was the Merlin of radical politics.

Question: Who was the of radical politics He?

Answer: Merlin

19 - Rule1

Candidate: While Lloyd George was wowing them at the Edinburgh Castle, another rising star was doing the same in the West End.

Question: Who was doing the same in the West End another rising star?

Answer: Lloyd George

20 - Rule2

Candidate: A lost Eden of innocence and imagination, immortalised in classic children's stories - The Railway Children, The Wind In The Willows, and Peter Pan.

Question: A lost whom of innocence and imagination, immortalised in classic children's stories - The Railway Children, The Wind In The Willows, and Peter Pan?

Answer: Eden

21 - Rule2

Candidate: As The meeting began, a group of thugs began mingling with the Pankhurst supporters.

Question: As The meeting began, a group of thugs began mingling with the whom supporters?

Answer: Pankhurst

22 - Rule2

Candidate: The answer was a terrifying Northern roar, and down the hill poured the men, many of them carrying sticks, and coming straight for Adela.

Question: The answer was a terrifying Northern roar, and down the hill poured the men, many of them carrying sticks, and coming straight for whom?

Answer: Adela

23 - Rule2

Candidate: Things didn't start well for Claude.

Question: Things didn't start well for whom?

Answer: Claude

24 - Rule2

Candidate: For Willoughby and his kind, the modern world was a most unpleasant rumour.

Question: the modern world was a most unpleasant rumour for Willoughby and his kind?

Answer: Willoughby

25 - Rule2

Candidate: But to the rich, Lloyd George himself was a great deal more dangerous than any wolf.

Question: But himself was a great deal more dangerous than any wolf to the rich, Lloyd George?

Answer: Lloyd George

26 - Rule3

Candidate: But just at that moment, Campbell-Bannerman's clever lying-down-in-bed cure failed him.

Question: But , whose clever lying-down-in-bed cure failed him just at that moment?

Answer: Campbell-Bannerman

27 - Rule3

Candidate: Lloyd George's problem was that dreadnoughts were ruinously expensive.

Question: whose problem was that dreadnoughts were ruinously expensive?

Answer: Lloyd George

28 - Rule5

Candidate: March 1906.

Question: When did 1906?

Answer: March 1906

29 - Rule5

Candidate: It was a publicity stunt for the Daily Mail, who were serialising the latest thriller to shock the Edwardian British - The Invasion Of 1910.

Question: null?

Answer: 1910

30 - Rule5

Candidate: On 15th July 1906, a young woman called Adela Pankhurst made her way to a park in Manchester called Boggart Hole Clough.

Question: When did a young woman call Adela Pankhurst made her way to a park in Manchester called Boggart Hole Clough?

Answer: On 15th July 1906

31 - Rule5

Candidate: What happened here one long ago Edwardian Sunday was only the very beginning, because the battle for democracy, women's votes, would prove to be extraordinarily violent.

Question: When be only the very beginning , because the battle for democracy , women 's votes , would prove to extraordinarily violent?

Answer: Sunday

32 - Rule5

Candidate: In January 1906, the Liberals were swept into power, promising to tackle inequality and reform politics.

Question: When were the Liberals swept into power , promising to tackle inequality and reform politics?

Answer: In January 1906

33 - Rule5

Candidate: In April 1908, he became the first and only prime minister to die in Downing Street.

Question: When did he become the first and only prime minister to die in Downing Street?

Answer: In April 1908

34 - Rule5

Candidate: But in 1908, Herbert Asquith was reshaping the administration into the greatest Liberal government of modern times.

Question: When was reshaping the administration into the greatest Liberal government of modern times?

Answer: In 1908

35 - Rule5

Candidate: A Frenchman - Louis Bleriot - was the first man to fly across the Channel in July 1909.

Question: When was A Frenchman - Louis Bleriot - the first man to fly across the Channel in July 1909?

Answer: July 1909

36 - Rule5

Candidate: The Daily Mail, always ready for a sharp stunt, offered a prize of 10,000 - huge money, more than 750,000 today...

Question: When did The Daily Mail offer a prize of 10 , 000 - huge money , more than 750 , 000 today?

Answer: Today

37 - Rule5

Candidate: In 1909, recession was looming, unemployment was rising.

Question: When was unemployment rising?

Answer: In 1909

38 - Rule5

Candidate: And so, on the evening of July 30th 1909, Lloyd George decided he had no choice but to take his People's Budget directly to the people.

Question: When had decided he no choice but to take his People 's Budget directly to the people so?

Answer: On the evening of July 30th 1909

39 - Rule5

Candidate: Music hall, or vaudeville, mattered in 1909.

Question: When did Music hall, or vaudeville, mattered in 1909?

Answer: 1909

40 - Rule6

Candidate: On 15th July 1906, a young woman called Adela Pankhurst made her way to a park in Manchester called Boggart Hole Clough.

Question: Where did a young woman call Adela Pankhurst made her way to a park on 15th July 1906?

Answer: In Manchester called Boggart Hole Clough

41 - Rule6

Candidate: In April 1908, he became the first and only prime minister to die in Downing Street.

Question: Where did he become the first and only prime minister to die in April 1908?

Answer: In Downing Street

42 - Rule6

Candidate: Within 24 hours, his plane was on display at Selfridges in London.

Question: Where was his plane on display at Selfridges within 24 hours?

Answer: In London

43 - Rule6

Candidate: Paulhan had already arrived in Manchester and claimed the prize, thanks to such dastardly tactics as not having a sleep before he started.

Question: Where had Paulhan already arrived and claimed the prize, thanks to such dastardly tactics as not having a sleep before he started?

Answer: In Manchester

Questions for The History of Christianity

Please assign one of the following scores:

1 – Unusable (question is ungrammatical, does not make sense, cannot be answered)

2 – Usable with major revision (e.g. word needs to be rearranged or combination of several errors)

3 – Usable with minor revision (e.g. punctuation error)

4 - Usable without revision (question can be used the way it is)

1 - Rule1

Candidate: Jesus, the wandering Jewish teacher, crucified by the Romans.

Question: Who crucified by the Romans , the wandering Jewish teacher?

Answer: Jesus

2 - Rule1

Candidate: Paul, who had hunted down Christians until on the road to Damascus, he experienced a blinding vision of Jesus Christ resurrected from the dead.

Question: Who experienced a blinding vision of Jesus Christ resurrected from the dead he?

Answer: Paul

3 - Rule1

Candidate: The Church is said to have been built where Jesus was crucified and buried.

Question: Who is said to have been built where was crucified and buried The Church?

Answer: Jesus

4 - Rule1

Candidate: Somehow the followers of Jesus became convinced that he rose from here to new life.

Question: Who became convinced that he rose from here to new life the followers of somehow?

Answer: Jesus

5 - Rule1

Candidate: The belief that Jesus can overcome death is the most difficult and troubling affirmation of the Christian faith.

Question: Who is the most difficult and troubling affirmation of the Christian faith The belief that can overcome death?

Answer: Jesus

6 - Rule1

Candidate: Its core is the unprecedented idea that God became human, not in a pharaoh, a king or even an emperor, but in a humble peasant from Galilee.

Question: Who is the unprecedented idea that became human, not in a pharaoh, a king or even an emperor, but in a humble peasant from Galilee Its core?

Answer: God

7 - Rule1

Candidate: And the conviction that you can meet Jesus, the son of God, and transform your life is a compelling message.

Question: Who is a compelling message the conviction that you can meet , the son of who, and transform your life?

Answer: God and Jesus

8 - Rule1

Candidate: Many versions of Christian history would make this unorthodox too .

Question: Who would make this unorthodox too Many versions of history?

Answer: Christian

9 - Rule1

Candidate: Well, you might think obviously west to Rome, because that's where Paul had gone.

Question: Who might think obviously west to Rome, because that's where had gone you?

Answer: Paul

10 - Rule1

Candidate: Paul had been killed in Rome.

Question: Who had been killed in Rome?

Answer: Paul

11 - Rule1

Candidate: Edessa is special, because its ruler King Abgar set an important precedent here.

Question: Who is special, because its ruler King set an important precedent here who?

Answer: Edessa and Abgar

12 - Rule1

Candidate: And Edessa pioneered something else that has become inseparable from Christianity...

Question: Who pioneered something else that has become inseparable from Christianity?

Answer: Edessa

13 - Rule1

Candidate: Christian Edessa has long since disappeared.

Question: Who has long since disappeared?

Answer: Christian Edessa

14 - Rule1

Candidate: It was more than a 100 years after the King of Edessa had made Christianity his official religion.

Question: Who was more than a 100 years after the King of had made Christianity his official religion It?

Answer: Edessa

15 - Rule1

Candidate: Jesus had told people to abandon wealth, not to ally with the rich and powerful.

Question: Who had told people to abandon wealth, not to ally with the rich and powerful?

Answer: Jesus

16 - Rule1

Candidate: For almost 40 years a holy man called St Simeon lived on top of a stone column.

Question: Who called lived on top of a stone column a holy man for almost 40 years?

Answer: St Simeon

17 - Rule1

Candidate: St Simeon is the most famous of many Syrian hermits who tried to come closer to God by punishing their bodies.

Question: Who is the most famous of many Syrian hermits who tried to come closer to God by punishing their bodies?

Answer: St Simeon

18 - Rule1

Candidate: Further south was Alexandria in Egypt.

Question: Who was in Egypt Further south?

Answer: Alexandria

19 - Rule1

Candidate: According to a thoughtful but maverick Egyptian priest, Jesus was not the same as God.

Question: Who was not the same as God according to a thoughtful but maverick Egyptian priest?

Answer: Jesus

20 - Rule1

Candidate: If Jesus Christ is not fully God, then is his death on the cross enough to save you from your sins and get you to Heaven?

Question: Who is his death on the cross enough to save you from your sins and get you to Heaven then?

Answer: Jesus Christ

21 - Rule1

Candidate: Christ died to give us the chance to have an infinitely better life.

Question: Who died to give us the chance to have an infinitely better life?

Answer: Christ

22 - Rule1

Candidate: The phrase was that Jesus was “of one substance” with the Father.

Question: Who was that was “of one substance” with the Father The phrase?

Answer: Jesus

23 - Rule1

Candidate: It states that God is equally the Father, Jesus the Son and the Holy Spirit.

Question: Who states that is equally the Father, Jesus the Son and the Holy Spirit It?

Answer: God

24 - Rule2

Candidate: It is an epic story starring a cast of extraordinary people, from Jesus himself and the first apostles to emperors, kings and popes.

Question: It is an epic story starring a cast of extraordinary people, from whom himself and the first apostles to emperors, kings and popes?

Answer: Jesus

25 - Rule2

Candidate: Deep down the Christian faith boasts a shared core

Question: Deep down the whom faith boasts a shared core?

Answer: Christian

26 - Rule2

Candidate: better to start than in the city which first knew Jesus the Christ

Question: better to start than in the city which first knew whom the Christ?

Answer: Jesus

27 - Rule2

Candidate: We've all heard something of the Christian story.

Question: We've all heard something of the whom story?

Answer: Christian

28 - Rule2

Candidate: Paul, who had hunted down Christians until on the road to Damascus, he experienced a blinding vision of Jesus Christ resurrected from the dead.

Question: Paul, who had hunted down Christians until on the road to Damascus, he experienced a blinding vision of whom resurrected from the dead?

Answer: Jesus Christ

29 - Rule2

Candidate: It reshaped not just the faith of Christ but in the end, all eastern civilisation.

Question: It reshaped not just the faith of whom but in the end, all eastern civilisation?

Answer: Christ

30 - Rule2

Candidate: The belief that Jesus can overcome death is the most difficult and troubling affirmation of the Christian faith.

Question: The belief that Jesus can overcome death is the most difficult and troubling affirmation of the whom faith?

Answer: Christian

31 - Rule2

Candidate: But the Church built around the tomb of Jesus is also the starting point for a forgotten story, a story that may overturn your preconceptions about early Christianity.

Question: But the Church built around the tomb of whom is also the starting point for a forgotten story, a story that may overturn your preconceptions about early Christianity?

Answer: Jesus

32 - Rule2

Candidate: Orthodoxy is a large part of the Christian story.

Question: Orthodoxy is a large part of the whom story?

Answer: Christian

33 - Rule2

Candidate: Because the origins of the Christian faith are not in the West, but here in these ancient Churches of the East.

Question: faith are not in the West, but here in these ancient Churches of the East because the origins of the Christian?

Answer: Christian

34 - Rule2

Candidate: I'm not giving you a history of Christian theology, though I won't be afraid to plunge you into many ancient arguments about Christian faith.

Question: I'm not giving you a history of whom theology, though I won't be afraid to plunge you into many ancient arguments about whom faith?

Answer: Christian

35 - Rule2

Candidate: I'm not giving you a history of Christian theology, though I won't be afraid to plunge you into many ancient arguments about Christian faith.

Question: I'm not giving you a history of whom theology, though I won't be afraid to plunge you into many ancient arguments about whom faith?

Answer: Christian

36 - Rule2

Candidate: It is in fact the Church, the institution of Christian faith that has fought its way through history.

Question: It is in fact the Church, the institution of whom faith that has fought its way through history?

Answer: Christian

37 - Rule2

Candidate: It all started here in Jerusalem, when the first followers of Jesus formed a Jewish Christian Church.

Question: It all started here in Jerusalem, when the first followers of whom formed a Jewish Christian Church?

Answer: Jesus

38 - Rule2

Candidate: It was led by James, whom the gospels call the brother of Jesus.

Question: It was led by James, whom the gospels call the brother of whom?

Answer: Jesus

39 - Rule2

Candidate: It was led by James, whom the gospels call the brother of Jesus.

Question: It was led by whom, whom the gospels call the brother of Jesus?

Answer: James

40 - Rule2

Candidate: So had the Apostle Peter.

Question: So had the Apostle whom?

Answer: Peter

41 - Rule2

Candidate: In the first century it was called Edessa, capital of a small kingdom, and wealthy because it controlled part of the main trade route east.

Question: it was called whom, capital of a small kingdom, and wealthy because it controlled part of the main trade route east in the first century?

Answer: Edessa

42 - Rule2

Candidate: He chose to show his personal devotion to Jesus by adopting Christianity as the Kingdom's official state religion, at least 100 years before the Romans did.

Question: He chose to show his personal devotion to whom by adopting Christianity as the Kingdom's official state religion, at least 100 years before the Romans did?

Answer: Jesus

43 - Rule2

Candidate: And this is where it all started - in the ancient Eastern Christian kingdom of Edessa.

Question: this is where it all started - in the ancient Eastern Christian kingdom of whom?

Answer: Edessa

44 - Rule2

Candidate: But its liturgical chant is still based on the distinctive tradition of Edessa.

Question: But its liturgical chant is still based on the distinctive tradition of whom?

Answer: Edessa

45 - Rule2

Candidate: These hymns are derived from the poetry of the great 4th century Syrian theologian St Ephrem.

Question: These hymns are derived from the poetry of the great 4th century Syrian theologian whom Ephrem?

Answer: St

46 - Rule2

Candidate: HORN BLARES In the West, most Christians wouldn't be singing the public praises of God because it was too dangerous.

Question: most Christians wouldn't be singing the public praises of whom because it was too dangerous hORN BLARES In the West?

Answer: God

47 - Rule2

Candidate: It was a turning point in the history of the Christian faith.

Question: It was a turning point in the history of the whom faith?

Answer: Christian

48 - Rule2

Candidate: It gave the Christian faith the chance of becoming a universal religion.

Question: It gave the whom faith the chance of becoming a universal religion?

Answer: Christian

49 - Rule2

Candidate: Some Christians actually listened to what Jesus had said.

Question: Some Christians listened to what whom had said actually?

Answer: Jesus

50 - Rule2

Candidate: Crowds came to see St Simeon sitting on his pillar.

Question: Crowds came to see whom sitting on his pillar?

Answer: St Simeon

51 - Rule2

Candidate: St Simeon is the most famous of many Syrian hermits who tried to come closer to God by punishing their bodies.

Question: St Simeon is the most famous of many Syrian hermits who tried to come closer to whom by punishing their bodies?

Answer: God

52 - Rule2

Candidate: The day will be in the next life where we will see God.

Question: The day will be in the next life where we will see whom?

Answer: God

53 - Rule2

Candidate: Some started to gather in communities where they could follow God in purity and simplicity.

Question: Some started to gather in communities where they could follow whom in purity and simplicity?

Answer: God

54 - Rule2

Candidate: And their fight gets mixed up with what they believe about God.

Question: their fight gets mixed up with what they believe about whom?

Answer: God

55 - Rule2

Candidate: Constantine presided over four rival centres of Christian authority.

Question: Constantine presided over four rival centres of whom authority?

Answer: Christian

56 - Rule2

Candidate: The Bishop of Rome was the Pope, honoured in the West as successor to the Apostle Peter.

Question: The Bishop of Rome was the Pope, honoured in the West as successor to the Apostle whom?

Answer: Peter

57 - Rule2

Candidate: Matters came to a head over a question at the heart of the Christian faith.

Question: Matters came to a head over a question at the heart of the whom faith?

Answer: Christian

58 - Rule2

Candidate: According to a thoughtful but maverick Egyptian priest, Jesus was not the same as God.

Question: Jesus was not the same as whom according to a thoughtful but maverick Egyptian priest?

Answer: God

59 - Rule2

Candidate: He claimed that it was impossible for God, who is perfect and indivisible, to have created the human being Jesus out of himself.

Question: He claimed that it was impossible for whom, who is perfect and indivisible, to have created the human being Jesus out of himself?

Answer: God

60 - Rule2

Candidate: He claimed that it was impossible for God, who is perfect and indivisible, to have created the human being Jesus out of himself.

Question: He claimed that it was impossible for God, who is perfect and indivisible, to have created the human being whom out of himself?

Answer: Jesus

61 - Rule2

Candidate: The power of Christian belief lay in its claim to wipe away all the misery that humans feel about sin and death, the guilt and shame.

Question: The power of whom belief lay in its claim to wipe away all the misery that humans feel about sin and death, the guilt and shame?

Answer: Christian

62 - Rule2

Candidate: After many more arguments over the next half century, this phrase stayed at the heart of one of the most important Christian texts of all time.

Question: this phrase stayed at the heart of one of the most important whom texts of all time after many more arguments over the next half century?

Answer: Christian

63 - Rule2

Candidate: It's still recited in everyday worship throughout the Christian world.

Question: It's still recited in everyday worship throughout the whom world?

Answer: Christian

64 - Rule2

Candidate: Bishop Nestorius wasted little time in plunging the Church into a fresh quarrel about the nature of Jesus.

Question: Bishop Nestorius wasted little time in plunging the Church into a fresh quarrel about the nature of whom?

Answer: Jesus

65 - Rule3

Candidate: Paul's new-found zeal focused on people beyond the Jews - Gentiles.

Question: whose new-found zeal focused on people beyond the Jews - Gentiles?

Answer: Paul

66 - Rule3

Candidate: According to the Syrian enthusiast for St Simeon's Church I met, this approach set Eastern Christians apart from the West.

Question: this approach set Eastern Christians apart from the West according to the Syrian enthusiast for St Simeon's Church I met?

Answer: St Simeon

67 - Rule5

Candidate: For the last 17 centuries, Christianity has been repeatedly linked with the state, so In the United Kingdom, the monarch is still Supreme Governor of the Church of England.

Question: null?

Answer: For the last 17 centuries

68 - Rule5

Candidate: Yet the fact was many Christians had said the same over the previous three centuries, here on the shores of the Bosphorus as much as anywhere else.

Question: When was the fact many Christians had said the same over the previous three centuries, here on the shores of the Bosphorus as much as anywhere else?

Answer: Previous three centuries

69 - Rule5

Candidate: After many more arguments over the next half century, this phrase stayed at the heart of one of the most important Christian texts of all time.

Question: When did this phrase stay at the heart of one of the most important Christian texts of all time?

Answer: After many more arguments over the next half century

70 - Rule6

Candidate: I'm in Jerusalem for a very good reason, .

Question: Where 'm i?

Answer: In Jerusalem for a very good reason,

71 - Rule6

Candidate: But I'm here in Jerusalem because I want to look for something else.

Question: Where 'm i here because I want to look for something else?

Answer: In Jerusalem

72 - Rule6

Candidate: HE INTONES Orthodoxy and Catholicism dominated Christianity in Europe, in the West, for its first 15,000 years.

Question: Where did hE dominate INTONES Orthodoxy and Catholicism Christianity, in the West, for its first 15, 000 years?

Answer: In Europe

73 - Rule6

Candidate: And indeed at one point it was poised to triumph in Asia, maybe even in China.

Question: Where was it poised to triumph indeed at one point?

Answer: In Asia, maybe even in China

74 - Rule6

Candidate: And indeed at one point it was poised to triumph in Asia, maybe even in China.

Question: Where was it poised to triumph in Asia, maybe even indeed at one point?

Answer: In China

75 - Rule6

Candidate: It all started here in Jerusalem, when the first followers of Jesus formed a Jewish Christian Church.

Question: Where did it start here, when the first followers of Jesus formed a Jewish Christian Church?

Answer: In Jerusalem

76 - Rule6

Candidate: Paul had been killed in Rome.

Question: Where had Paul been killed?

Answer: In Rome

77 - Rule6

Candidate: It was Eastern Christians here in Syria who led the way...

Question: Where were Eastern Christians here?

Answer: In Syria who led the way.

78 - Rule6

Candidate: In the north of Syria there is one of the oddest souvenirs of the new religious movement in Eastern Christianity.

Question: Where is one of the oddest souvenirs of the new religious movement in the north of Syria?

Answer: In Eastern Christianity

79 - Rule6

Candidate: That's interesting because the stereotype in Europe of the hermit is someone who goes away from the world, .

Question: Where is that interesting because the stereotype someone who goes away from the world?

Answer: In Europe of the hermit

166 - Rule6

Candidate: Further south was Alexandria in Egypt.

Question: Where was further south Alexandria?

Answer: In Egypt

Questions for Britain from Above

Please assign one of the following scores:

- 1 – Unusable (question is ungrammatical, does not make sense, cannot be answered)
- 2 – Usable with major revision (e.g. word needs to be rearranged or combination of several errors)
- 3 – Usable with minor revision (e.g. punctuation error)
- 4 - Usable without revision (question can be used the way it is)

1 - Rule1

Candidate: Paul Finch was one of its early inhabitants, now returning after 40 years to the estate he lived in as a child.

Question: Who lived in as a child he?

Answer: Paul Finch

2 - Rule1

Candidate: Archaeologist Chris Going has been documenting the changing face of London from the air for the last five years.

Question: Who has been documenting the changing face of London from the air for the last five years Archaeologist?

Answer: Chris Going

3 - Rule1

Candidate: By flying exactly the same route the RAF did 60 years ago to create the first aerial surveys, Chris hopes to create an identical modern survey of his own.

Question: Who hopes to create an identical modern survey of his own by flying exactly the same route the RAF did 60 years ago to create the first aerial surveys?

Answer: Chris

4 - Rule1

Candidate: By lining up the two complete sets of images, Chris is able to switch between the past and the present.

Question: Who is able to switch between the past and the present by lining up the two complete sets of images?

Answer: Chris

5 - Rule1

Candidate: Years earlier, after the great fire, Christopher Wren came up with a grand new vision for London.

Question: Who came up with a grand new vision for London after the great fire?

Answer: Christopher Wren

6 - Rule1

Candidate: So when You're looking, specifically, from the west side of London, you would block St Paul's if the building was straight up.

Question: Who would block 's if the building was straight up you?

Answer: St Paul

7 - Rule2

Candidate: A formal European capital that would radiate out from the glorious centrepiece of St Paul's Cathedral.

Question: A formal European capital that would radiate out from the glorious centrepiece of whom's Cathedral?

Answer: St Paul

8 - Rule3

Candidate: There are dozens of ancient buildings, none greater than the looming presence of St Paul's Cathedral.

Question: There are dozens of ancient buildings, none greater than the looming presence of whose Cathedral?

Answer: St Paul

9 - Rule3

Candidate: This picture, and thousands like it, form part of a giant 3-D graphic model showing the whole city, with St Paul's at its heart.

Question: This picture, and thousands like it, form part of a giant 3-D graphic model showing the whole city, with whose at its heart?

Answer: St Paul

10 - Rule5

Candidate: Looking down in the capital today, what's obvious is the sheer scale and complexity of this sprawling metropolis.

Question: When is the sheer scale and complexity of this sprawling metropolis?

Answer: Today

11 - Rule5

Candidate: London's transformation began on September 7th, 1940.

Question: When did london's transformation begin on September 7th , 1940?

Answer: September 7th, 1940

12 - Rule5

Candidate: They arrived here at 6.45 in the evening, and looked down on their target, the heart of London's docks.

Question: null?

Answer: 6.45

13 - Rule5

Candidate: What You're looking at here is probably the most devastating change to London since the fire of London in 1666.

Question: When is probably the most devastating change to London since the fire of London in 1666?

Answer: 1666

14 - Rule5

Candidate: Between 1945 and 1949, the RAF flew more than 200 missions over London, shooting 50,000 individual frames, recording every square inch of the capital.

Question: When did the RAF fly more than 200 missions over London, shooting 50,000 individual frames, recording every square inch of the capital?

Answer: Between 1945 and 1949

15 - Rule5

Candidate: The late 1940s was a radical time, when Britain first turned old ideas of a National Health Service and a full welfare state into reality.

Question: When was a radical time, when Britain first turned old ideas of a National Health Service and a full welfare state into reality?

Answer: 1940

16 - Rule5

Candidate: This is Churchill Gardens today.

Question: When is this Churchill Gardens today?

Answer: Today

17 - Rule5

Candidate: Archaeologist Chris Going has been documenting the changing face of London from the air for the last five years.

Question: When has archaeologist Chris Going documented the changing face of London from the air for the last five years?

Answer: Last five years

18 - Rule5

Candidate: By flying exactly the same route the RAF did 60 years ago to create the first aerial surveys, Chris hopes to create an identical modern survey of his own.

Question: When did Chris hope to create an identical modern survey of his own?

Answer: By flying exactly the same route the RAF did 60 years ago to create the first aerial surveys

19 - Rule5

Candidate: It does not that like the sort of envisaged city of the planners of the '40s and '50s.

Question: When did it envisage does not that like the sort of city of the planners of the '40s and '50s?

Answer: The '40s

20 - Rule5

Candidate: From the city of the 1940s to the city of today, there's a world of difference.

Question: When 's a world of difference?

Answer: From the city of the 1940s to the city of today

21 - Rule5

Candidate: From the city of the 1940s to the city of today, there's a world of difference.

Question: When 's a world of difference?

Answer:

22 - Rule5

Candidate: Money changed London in ways no-one in the 1940s could ever have imagined, because London changed the way it made money, and nowhere shows this more clearly than here.

Question: When have money changed London in ways no-one in the 1940 s could ever imagined , because London changed the way it made money , and nowhere shows this more clearly than here?

Answer: The 1940s

23 - Rule5

Candidate: Stretching for ten miles along the Thames, by the late 1930s, the port of London had grown to be the largest in the world.

Question: When be the port of London grown to the largest in the world?

Answer: The late 1930s

24 - Rule5

Candidate: As The last dock facilities finally closed at the end of the '70s, the remaining 10,000 jobs went with them, leaving behind a vast, derelict wasteland.

Question: When did the remaining 10,000 jobs go with them , leaving behind a vast , derelict wasteland?

Answer: The '70s

25 - Rule5

Candidate: When The docks became redundant in the early 1970s, there was a great think about what to do with this area.

Question: When was a great think about what to do with this area?

Answer: The early 1970s

26 - Rule5

Candidate: By the end of the '80s, the wasteland had become the biggest building site in the world.

Question: When had the wasteland become the biggest building site in the world?

Answer: By the end of the '80s

27 - Rule5

Candidate: In the late 1940s, People believed in planning, .

Question: When did people believed in planning?

Answer: In the late 1940s

28 - Rule6

Candidate: In fact, almost all open spaces in London were used for cultivating food.

Question: Where were used for cultivating food in fact?

Answer: In London

29 - Rule6

Candidate: Cars and roads would be the way forward, as Abercrombie had seen in America.

Question: Where be cars and roads would the way forward, as Abercrombie had seen?

Answer: In America

30 - Rule6

Candidate: We don't have the classical streets that you see in Paris, .

Question: Where have we do n't the classical streets that you see?

Answer: In Paris

31 - Rule6

Candidate: The grid forms we see in New York.

Question: Where did the grid forms we see?

Answer: In New York

APPENDIX D: EVALUATORS' SCORING SHEETS FOR EXPERIMENT 1

Britain from Above				
Q no	1	2	3	4
1		1		
2			1	
3				1
4				1
5				1
6	1			
7		1		
8				1
9		1		
10		1		
11		1		
12	1			
13	1			
14				1
15				1
16	1			
17		1		
18		1		
19	1			
20		1		
21	1			
22		1		
23				1
24			1	
25				1
26				1
27		1		
28		1		
29		1		
30		1		
31		1		
Sum	6	14	2	9

Making of Modern Britain				
Q no	1	2	3	4
1		1		
2			1	
3				1
4				1
5				1
6			1	
7				1
8			1	
9			1	
10			1	
11			1	
12				1
13				1
14				1
15			1	
16				1
17				1
18			1	
19			1	
20			1	
21			1	
22				1
23				1
24	1			
25	1			
26				1
27				1
28	1			
29	1			
30			1	
31		1		
32				1

History of Christianity				
Q no	1	2	3	4
1		1		
2			1	
3		1		
4	1			
5		1		
6		1		
7	1			
8	1			
9		1		
10				1
11		1		
12				1
13				1
14		1		
15				1
16		1		
17				1
18	1			
19				1
20		1		
21				1
22		1		
23			1	
24		1		
25	1			
26	1			
27		1		
28			1	
29			1	
30		1		
31				1
32		1		

33			1	
34			1	
35			1	
36			1	
37				1
38		1		
39			1	
40			1	
41			1	
42				1
43				1
Sum	5	5	22	21

33	1			
34	1			
35	1			
36	1			
37		1		
38		1		
39		1		
40	1			
41		1		
42			1	
43		1		
44		1		
45	1			
46	1			
47		1		
48		1		
49				1
50				1
51				1
52				1
53				1
54				1
55			1	
56			1	
57			1	
58				1
59			1	
60		1		
61			1	
62		1		
63			1	
64				1
65				1
66	1			
67	1			
68			1	
69				1
70	1			
71	1			
72	1			

73				1
74	1			
75		1		
76		1		
77		1		
78	1			
79	1			
Sum	22	29	15	23

Scoring Sheet Evaluator 2

Britain from Above				
Q no	1	2	3	4
1				1
2		1		
3				1
4				1
5				1
6	1			
7	1			
8	1			
9	1			
10	1			
11		1		
12	1			
13	1			
14				1
15				1
16	1			
17		1		
18	1			
19	1			
20			1	
21		1		
22	1			
23			1	
24		1		
25	1			
26				1
27			1	
28	1			
29				
30	1			
31		1		
Sum	15	8	6	11

Making of modern Britain				
Q no	1	2	3	4
1		1		
2		1		
3				1
4			1	
5				1
6				1
7	1			
8		1		
9			1	
10	1			
11		1		
12				1
13				1
14				1
15	1			
16				1
17				1
18	1			
19	1			
20	1			
21			1	
22				1
23				1
24		1		
25	1			
26				1
27			1	
28	1			
29	1			
30			1	
31	1			
32			1	
33			1	
34			1	
35				1
36	1			

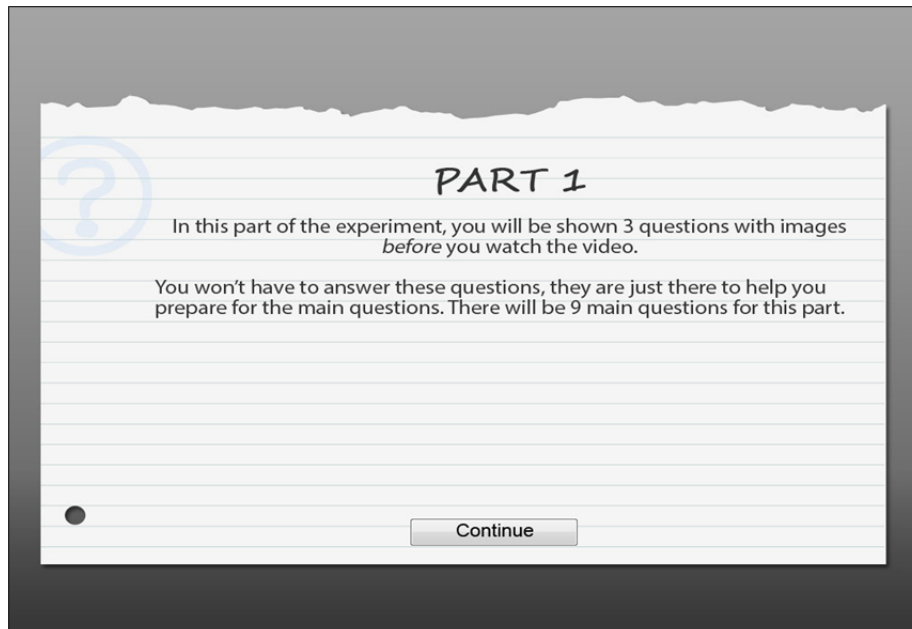
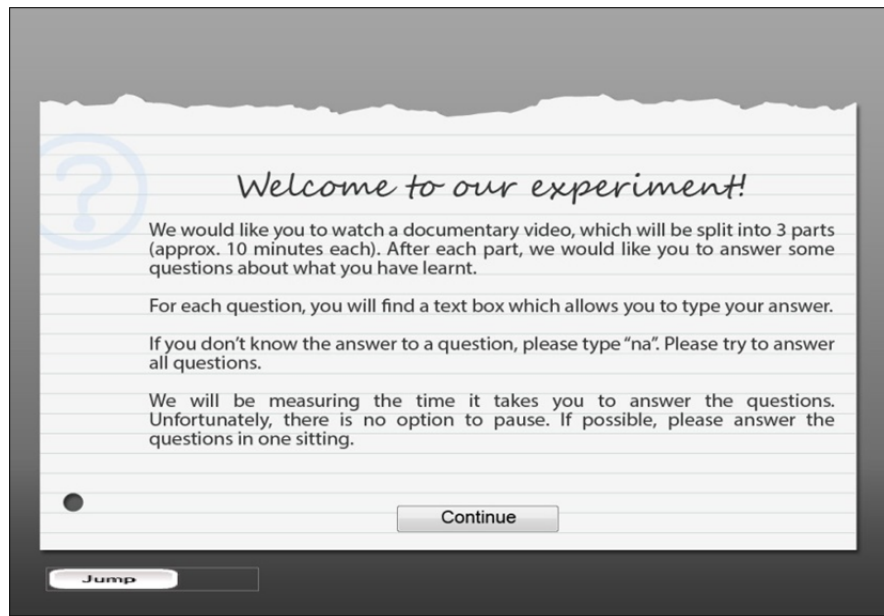
Christianity				
Q no	1	2	3	4
1			1	
2			1	
3	1			
4	1			
5	1			
6	1			
7	1			
8	1			
9	1			
10				1
11	1			
12				1
13				1
14	1			
15				1
16		1		
17				1
18	1			
19				1
20	1			
21				1
22	1			
23	1			
24	1			
25			1	
26	1			
27	1			
28				1
29	1			
30			1	
31			1	
32	1			
33	1			
34	1			
35	1			
36			1	

37				1
38	1			
39			1	
40			1	
41	1			
42	1			
43				1
Sum	15	7	13	8


37				1
38	1			
39	1			
40	1			
41	1			
42				1
43			1	
44			1	
45	1			
46		1		
47	1			
48	1			
49	1			
50				1
51				1
52	1			
53				1
54	1			
55	1			
56	1			
57	1			
58				1
59				1
60	1			
61	1			
62	1			
63	1			
64				1
65				1
66	1			
67	1			
68	1			
69				1
70	1			
71	1			
72	1			
73	1			
74	1			
75	1			
76	1			

77	1			
78	1			
79	1			
80	1			
Sum	53	4	11	22

APPENDIX E: SAMPLE QUESTIONS AND SCREENSHOTS FROM EXPERIMENT 1




What is nuclear fusion?




Continue

What had that Professor Martin Fleischmann and a colleague, Stanley Pons, discovered?

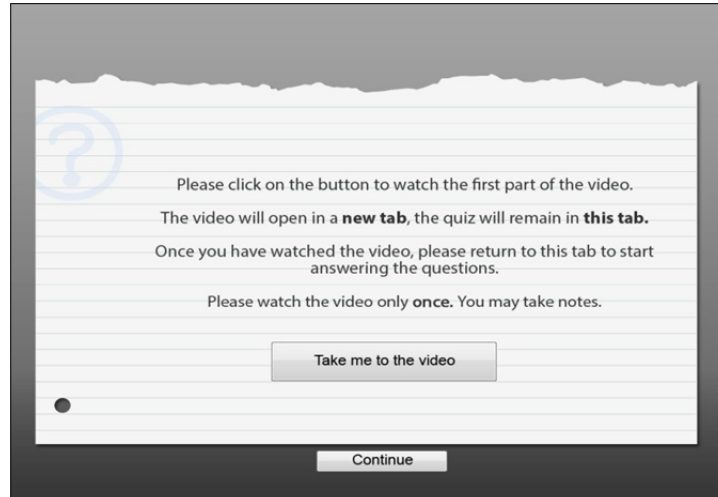


Continue

What would prove that fusion had taken place in theory?



Continue



?

Please click on the button to watch the first part of the video.
The video will open in a **new tab**, the quiz will remain in **this tab**.
Once you have watched the video, please return to this tab to start answering the questions.
Please watch the video only **once**. You may take notes.

Take me to the video

Continue

This card features a light blue question mark icon in the top left corner. The background is a white sheet of lined paper with a torn top edge, set against a dark grey background. A small black dot is located in the bottom left corner of the paper area. A button labeled 'Take me to the video' is centered below the text, and a 'Continue' button is centered at the bottom of the card.

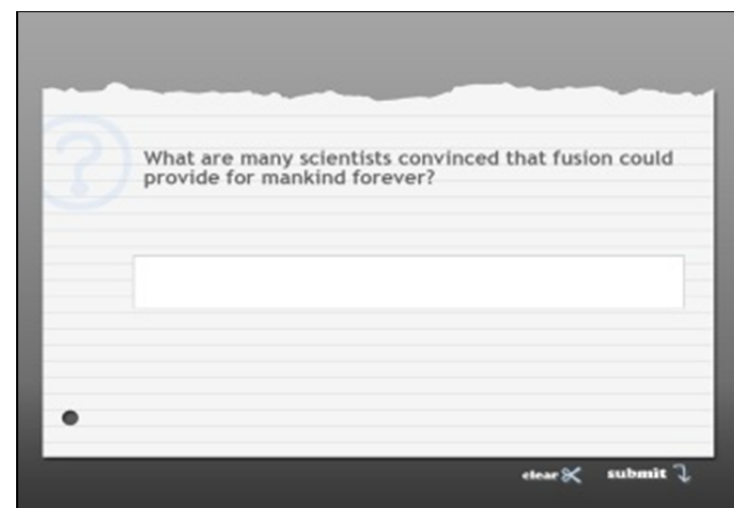


?

What is nuclear fusion?

clear ✕ submit ↵

This card features a light blue question mark icon in the top left corner. The background is a white sheet of lined paper with a torn top edge, set against a dark grey background. A small black dot is located in the bottom left corner of the paper area. A large, empty white rectangular input field is centered below the question. At the bottom right of the card, there are two buttons: 'clear' with a small 'x' icon and 'submit' with a small arrow icon.



?

What are many scientists convinced that fusion could provide for mankind forever?

clear ✕ submit ↵

This card features a light blue question mark icon in the top left corner. The background is a white sheet of lined paper with a torn top edge, set against a dark grey background. A small black dot is located in the bottom left corner of the paper area. A large, empty white rectangular input field is centered below the question. At the bottom right of the card, there are two buttons: 'clear' with a small 'x' icon and 'submit' with a small arrow icon.

APPENDIX F: PET OUTPUT (EXTRACT)

219

#annotator	#edit_time	hter_ins	hter_delete	hter_sub	hter_shift	hter_errors	hter_words	#usab_bef	#usab_aft	#PE_effort	System	Source Sentence	System Question	Post-edited Question
A1	10.11	9	0	2	0	11	5	1	3	3	CMU	Like all nuclear devices, Garwin's exploited an extraordinary property of Einstein's equation, that tiny amounts of matter would contain massive amounts of energy.	What does Garwin's exploited an extraordinary property of like all nuclear devices?	What did Garwin exploit?
A1	20.967	16	0	0	0	16	4	1	4	2	CMU	Photosynthesis is the process by which plants take sunlight, combine it with water and carbon dioxide and create energy.	What is Photosynthesis is by which plants take sunlight, combine it with water and carbon dioxide and create energy?	What is photosynthesis?

A1	13.011	0	0	1	0	1	15	2	3	2	CMU	Former CIA agent, Robert Baer, spent a decade in the Middle East trying to unravel fact from fiction.	What spent a decade in the Middle East trying to unravel fact from fiction?	Who spent a decade in the Middle East trying to unravel fact from fiction?
A1	53.843	0	2	4	0	6	19	2	4	2	CMU	With the support of successive Presidents, 60,000 Americans were sterilised- a practice that continued until the 1970s.	Who were sterilised- a practice that continued until the 1970s with the support of successive Presidents?	How many Americans were sterilised -a practice that continued until the 1970s with the support of successive Presidents?
A1	9.286	0	0	0	0	0	13	3	none	1	WLV	With just five characters, Albert Einstein revealed an extraordinary truth of the world.	Who revealed an extraordinary truth of the world with just five characters?	Who revealed an extraordinary truth of the world with just five characters?
A1	22.311	0	1	0	0	1	14	3	4	2	WLV	Since 1960, there's been another secret organisation a president can turn to.	When has there been another secret organisation a president can turn to?	Since when has there been another secret organisation a president can turn to?

A1	9.948	0	0	0	0	0	19	4	none	1	WLV	James Dewey Watson was one of the team that, in 1953, unravelled one of nature's deepest secrets.	Who was one of the team that, in 1953, unravelled one of nature's deepest secrets?	Who was one of the team that, in 1953, unravelled one of nature's deepest secrets?
A1	14.51	0	0	0	0	0	16	4	none	1	WLV	James Dewey Watson has repeatedly courted controversy with claims of how the genes may determine the destiny, .	Who has repeatedly courted controversy with claims of how the genes may determine the destiny?	Who has repeatedly courted controversy with claims of how the genes may determine the destiny?
A1	6.395	0	0	0	0	0	14	4	none	1	WLV	Once enriched, uranium atoms can be fissioned or split, yielding their energy.	What can be fissioned or split, yielding their energy, once enriched?	What can be fissioned or split, yielding their energy, once enriched?
A1	7.53	0	0	0	0	0	5	5	none	1	WLV	So what enrichment is, it involves separating out the useful, lighter atoms from the less useful heavier	What does enrichment involve?	What does enrichment involve?

													atoms.		
A1	7.895	0	0	0	0	0	8	5	none	1	Manua l	It is the Zipper-type centrifuge, it performs a crucial task in the production of nuclear fuel and weapons, the enrichment of uranium.	What is the Zipper-type centrifuge used for?	What is the Zipper-type centrifuge used for?	
A1	11.641	0	0	0	0	0	21	5	none	1	CMU	Photosynthesis is the process by which plants take sunlight, combine it with water and carbon dioxide and create energy.	What is the process by which plants take sunlight, combine it with water and carbon dioxide and create energy?	What is the process by which plants take sunlight, combine it with water and carbon dioxide and create energy?	
A1	17.527	0	0	0	0	0	28	rejected	rejected	rejected	WLV	In fact, we are only recently beginning to tease apart, nanosecond by nanosecond, the precise process by	What are we only recently beginning to tease apart , nanosecond by nanosecond , the precise process by which nature	What are we only recently beginning to tease apart , nanosecond by nanosecond , the precise process by which nature	

												which nature performs this magic called photosynthesis.	performs this magic called photosynthesis in fact?	performs this magic called photosynthesis in fact?
A1	4.699	0	0	0	0	0	7	rejected	rejected	rejected	CMU	For Watson, the benefits of this moment far outweigh any risk.	What outweigh any risk for Watson?	What outweigh any risk for Watson?