Using a new MCQ Generation methodology (CAREGen) in the UK Electricity Distribution Industry

Robert Michael Foster
Research Institute in Information and Language Processing
University of Wolverhampton, UK

Abstract

The Construed Antonym Realisation Exercise Generation (CAREGen) methodology for generating Multiple Choice Question (MCQ) test items is described and illustrated using two case studies from the UK Electricity Distribution Industry. The first step of the methodology is to create Controlled Specific Learning Objectives (CSLOs) that define the required assessment outcomes for the generated MCQs. The CSLOs are used to identify appropriate source documents. Statistical NLP tools are then applied to the syntactic patterns within the source documents to identify complementary and antonymic pairs of words and phrases. Construal theory is then used to define 'correct'/'incorrect' construal of these pairs in relation to the CSLOs and a template is applied to generate the test items. The component theories of CAREGen are illustrated using the TAR Induction case study. The Cable-Jointing case study is used to illustrate application of the steps of the methodology and the current evaluation method. The conclusion explains why effective evaluation of the methodology has proved so difficult and calls for the development of a suitable evaluation measure for MCQ generation systems such as CAREGen.

1. Introduction

In 1990 the company who host the development of the CAREGen methodology began developing and delivering Multiple Choice Questions (MCQs)[1] to facilitate an annual confirmation exercise which was intended to confirm that each staff member had maintained a working knowledge of documents from the company's Policy Library. In more recent years the range of topics for which there is a dedicated MCQ routine, has increased dramatically. An internal study in 2006 showed that manually creating and updating the 'MCQ test item bank' was an expensive process.

Subsequent research has resulted in various changes to the methodologies that are applied when MCQs are created and delivered [2], [3], [4]. This article describes the latest proposed MCQ generation methodology and illustrates its application with two case studies.

The article is organized as follows: Section 2 describes the background of the research effort that has led to the creation of the CAREGen methodology and then introduces the specific circumstances of each of the featured case studies. Section 3 uses selected examples from the TAR Induction case study to provide illustrated descriptions of the component theories: MAC MCQ formats[1].[5], CRST-compliant CSLOs[6].[7], Causal Coherence Relations[8], Complementary and Antonym relations [9], [10], [11] and Construal operations [12].

Sections 4 and 5 use the Cable-Jointing case study to illustrate how the steps of the methodology are applied and to present the method and results from the most recent CAREGen methodology evaluation. The conclusions listed in section 6, highlight the difficulties experienced by the researchers when comparing the efficacy [13], [14] of the CAREGen MCQ creation methodology with other systems that tackle the MCQ generation task[15],[16]. There does not appear to be an appropriate, quantitative measure of efficacy in the literature that might be used to evaluate the CAREGen methodology. The recommendations in section 7 therefore include a call for the development of a suitable evaluation measure that would allow the performance of different MCQ generation systems to be compared.

2. Background

When MCQ test items were first used in the company who host the development of the CAREGen methodology, the MCQ test items were only used for the highest level of authorisation, to provide field audit pre-tests and training course post-tests for a small group of staff. The process of completing an MCQ routine, reinforced learning and the stored responses to questions demonstrated to students, trainers, managers and external auditors that training had been received in accordance with requirements stated in UK Legislation [18].

After an internal study in 2006 highlighted a development bottleneck in the MCQ creation process, the company took several steps. Research was conducted [5] which led to a change in the format of MCQ test item that was recommended for use within the company. There were also some

experiments investigating the potential of some externally developed software which claimed to generate MCQ test items automatically. The creators of the automatic system [16], [17] expressed an interest in collaborating to improve their system. However, during initial experiments with a particular policy document, most of its clauses were filtered out and so the number of usable MCQ test items produced was very small.

These disappointing results led to a re-examination of the manual MCQ generation processes being applied by the company. This led to the identification of one possible way forward which involved more extensive pre-processing of source documents to avoid the observed filtering out of sentences in the early stages. The long term objective of the research is for the MCQ generation process to be automated. There have been several published refinements to the manual MCQ creation process [2],[3],[4] which might be automated when the Statistical NLP tools become sufficiently powerful. This article describes CAREGen, the latest of the designed methodologies.

2.1. TAR induction case study

In 2004 the company who host the development of the CAREGen methodology took on 30 new Apprentices and the indications were that the number of new recruits would continue to grow in the coming years. UK legislation requires UK companies to provide training for new staff:

"13. Capabilities and training:
(2) Every employer shall ensure that his employees are provided with adequate health and safety training
(a) on their being recruited into the employer's undertaking; …."

Management of Health and Safety at Work Regulations 1989

The upward trend in the count of new employees together with these legal requirements, led to an analysis of the process for delivering and assessing the initial training that is received by new recruits. This review highlighted the need for a formal system of progress monitoring, and a consistent system of assessment that could be applied when trainees had completed the prescribed training courses and development exercises.

A new progress monitoring system was therefore created to address this requirement. The system was named the Technical Achievement Record (TAR) system and has been developing over the past eight years. The TAR support systems are delivered over the company LAN from an intranet website and many of the rules for operating the system have been presented in the form of a FAQ (Frequently Asked Questions) document. An illustration showing the 'no frills' front menu webpage for the TAR system is shown below.

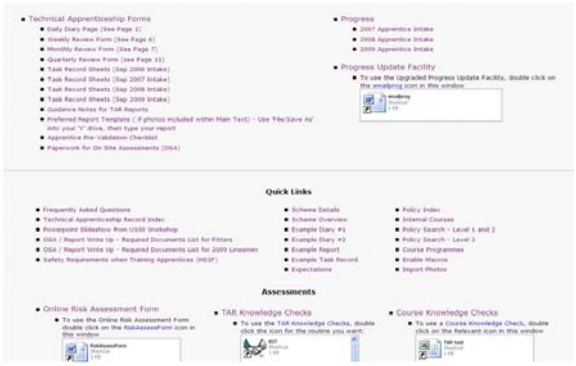

Figure 1. Front Screen for the Technical Achievement Record (TAR) system

The TAR induction is a one week training course which introduces the TAR system to new employees. Some of the training sessions include formative [18] quizzes that have been woven into the learning plan of the training session. The quizzes are also available outside the classroom to allow new staff refresh their knowledge of how the TAR system operates. The MCQ test items within these quizzes are provided in an interactive mode that delivers immediate feedback [25]. The TAR induction quizzes both Traditional 4-option Multiple Choice (4MC) test items and Multiple Alternative Choice (MAC)-formatted MCQ test items [1]

After receiving TAR induction training and working through the formative assessments and development exercises, apprentices take a summative assessment of their TAR system knowledge, which uses a different presentation of the MAC-formatted test items that appeared in the quizzes. An example of these items is provided below. These are 'MAC formatted MCQ test items'. Although the text cannot be read in such a small version, the significant point is that readers appreciate that candidates click FOUR buttons to enter a complete response to this item. Responses to parts 1,2 and 4 were answered correctly, but the response to part 3 was incorrect.

Figure 2. A MAC MCQ that tests knowledge of events following a failure (referral?) at the first attempt at the trade specific 'trade test'

2.2. Cable-jointing case study

On completion of their training which can take up to three years, apprentices must demonstrate that they have achieved a sufficient level of the knowledge and skills for effective conduct of their trade (Overhead Lines, Cable-Jointing, Substation-Fitting). In addition to monitoring progress of apprentices through their formal training courses, the TAR system also ensures that a sufficient breadth of experience has been achieved by new recruits before they are put forward for their trade test and then become fully qualified crafts-persons. The trade test includes a knowledge check and increasingly these knowledge checks are carried out using MCQs.

The MCQ test items used within trade-test knowledge-checks are provided in a NON interactive mode that does NOT deliver immediate feedback. The trade-test knowledge-checks use two different formats of MCQ test item:

- Traditional 4-option Multiple Choice (4MC) test items and
- Multiple Alternative Choice (MAC)-formatted MCQ test items

The MCQ test items in this article that have been used to illustrate the application and evaluation of the methodology were created for use in a Cable-Jointing knowledge check.

3. Academic Context

3.1. The Multiple Alternative Choice (MAC) MCQ format

The Multiple Alternate Choice (MAC) test item format is described in Haladyna's Review of item-writing guidelines [1] as a more general version of the Multiple True False (MTF) test item format in that the two responses available are not restricted to 'True' / 'False'. This format is the preferred format of MCQ for use in this company, following the conclusive results from a previous study [3].

3.2. Controlled Specific Learning Objectives

The 'CRST-compliant CSLO' structure draws from the concept of controlled language, exemplified by the AECMA project in the aircraft industry, to unite Robert Mager's theory of Specific Learning Objectives [6] with Mann and Thompson's Rhetorical Structure Theory [7]. The application of the CSLO standard seeks to enforce clarity when using content words to define the goals of a learning intervention.

The 'SLO' standard [6] requires that learning objectives define the Audience, Behaviour, Context and Degree of the required intervention. The suggested benefit of adding this structure is that successful completion of the SLO can be clearly identified and degrees of failure can be measured against the specified standard.

The 'CSLO' standard requires that CSLOs contain the elements of a SLO and that they are expressed using a controlled language. The concept of 'controlled language' is exemplified by the AECMA project which provides a domain specific lexicon and a series of 'rules' for well-formed text when writing aircraft maintenance manuals. The suggested benefit of adding this restriction is that it provides further protection against un-measurable objectives by:

(a) eliminating multi-sense content words and

(b) protecting against possible mis-understanding of content words through definitions in the domain specific lexicon.

The 'CRST-compliant CSLO' standard requires that the coherence relations used within a CRST-compliant CSLO are taken from a fixed range of coherence relations as defined in the Rhetorical Structure Theory (RST)[7], and that the relation that has been used is clearly identified.

The CSLO for the TAR induction training course is provided as an illustrated example of a CRST-compliant CSLO:

"After Apprentices have attended the TAR Induction training sessions they must be able to recognise correctly the facts that were covered during the TAR induction training sessions".

This complies with the CSLO standard because

(a) the four SLO elements[6] are included:

Table 1. CSLO for the TAR induction case study

Audience	**"After Apprentices have attended the TAR Induction training sessions they must be able to recognise correctly the facts that were covered during the TAR induction training sessions".**
Behaviour	
Context	
Degree	

(b) the domain specific lexicon (consisting of is definitions within documents contained in the company policy library) contains definitions for the domain specific terms: "Apprentices", "TAR

Induction training sessions" and "facts covered during the TAR Induction training sessions"

(c) the RST relations [7] that were used are identified as a time-**conditional** relation (signalled by the presence of 'After').

3.3. Causal Coherence Relation Primitives

When applying CAREGen to the 'TAR induction' case study, the choice of the most appropriate Coherence Relation for the CRST-compliant CSLO was straightforward. However, there are occasions when considering each of the available coherence relations can give benefits. This requires an appropriate system for categorising Causal Coherence relations.

The Sanders et al system [8] for categorising causal coherence relations assumes that all relations are cognitively basic. The proposal is that only four cognitive primitives are required to express the primitive causal coherence relations necessary for communication.

Combination of these four primitives by a writer can then present increasingly sophisticated types of causal coherence relation between a text's information units. The primitives are described in detail in the literature [8] but can be summarized as follows:

(a) Basic operation (causal vs additive)

(b) Source of coherence. (semantic vs pragmatic)

(c) Order of information units (basic vs complex).

(d) Polarity (positive vs negative)

An analysis of existing MCQ generation techniques and a comparison with the steps within the original process [16], [17] identified the importance of coherence relations in source sentences during the item creation process. The response to this discovery was the development of the CREAM technique [3]. However, the possibility that other theories might be able to help further became apparent when the following conclusions were drawn after observing and interviewing item designers who were applying a manual MCQ generation methodology.

This led to the discovery that:

(a)	When manually creating MCQs, Item designers seek to anticipate erroneous reader construal operations [12] and then identify instances of possible mis-construal following course attendance.
(b)	When manually creating MCQs, Item designers seek the salient features of the source documents within potentially very complex Antonym ranges [9], [10], [11] and then rationalise them into relatively simple domain-specific and construal-specific Complementarities [12].

3.4. Construal operations

The discovery that item designers in this domain seek to anticipate erroneous reader construal operations and then identify instances of possible mis-construal following course attendance led to an investigation of theories about construal operations:

William Croft explains in chapter 3 of 'Cognitive Linguistics' [12] that the construal operation classification system proposed in his book combines the observations of previous systems from the Linguistic Semantics and Cognitive Psychology communities.

"A basic premise of Cognitive Linguistics is that 'Language is an instance of general conceptual abilities'. The classification of construal operations …. is not intended to be a reduction of construal operations in just four processes. The various construal operations listed under the four headings are all distinct cognitive processes. The analysis we propose is that the various construal operations are manifestations of the four basic cognitive abilities in different aspects of experience. "

'Cognitive Linguistics' - Cruse and Croft 2004

The two other systems described in the literature for categorising construal operations which were considered for use within CARE are: Imaging systems [23] and Focal adjustments. However, as William Croft explains there are several important construal operations which are inadequately represented in these systems. For example the fundamental considerations of Framing [19] are missing and the more complex cognitive processes of Metaphor [21] and Image schemas [20], [22] are also inadequately covered for our purpose.

Application of this theory within the CAREGen methodology provides a domain specific response to the following questions, which are posed in the conclusion (section 3.6) to Chapter 3 within 'Cognitive Linguistics' [12]:

(a) How do construal operations interact within this domain?

(b) How should we characterise the processes of Language vs Thought vs Experience?

3.5. Complementary and Antonym relations

The discovery that item designers seek the salient features of the source documents within potentially very complex Antonym ranges [9], [10], [11] and

then rationalise them into relatively simple domain-specific, construal-specific Complementarities[12], led to an investigation of theories about antonym ranges:

Research into possible methods for generating the required Alternative Choice component of MAC items identified several contradictory studies [9].[10].[11]. If we take a dynamic construal approach to sense relations then we can adopt the general categories of Complementarities and Antonyms [9] to describe the different forms of semantic opposites in a domain. However, successful communications using these categorisations require both readers and writers of domain defining documents to share a common ground construal of the intended meanings for each set of opposites.

In many other application domains, this might present problems, but the principal objective of the training process is the achievement of a common ground construal of certain key facts and concepts between writers and readers of the company's policy documents. Therefore this requirement for a shared common ground is an acceptable, and perhaps even a desirable, feature of this system.

In the featured domain, when tackling an un-familiar topic, the cognitive response by both learner and trainer involves identifying good examples. Each example is judged according to how 'pure' and 'symmetrical' the opposition is between the two extremes. Assessment of a new concept involves identifying properties that are either present or absent and identifying features of the construal that are not relevant to the opposition. When the reader begins to imagine the effects of more or less of the property upon the degree of oppositeness, then under Cruse and Togia's system [10] the relationship changes from a straight-forward complementarity, into a significantly more complex, Antonymic relationship which is comparatively difficult to learn.

Similarly, when readers pre-suppose the presence of one property over its opposite, then the complementarity's purity decays towards more complex networks of antonymic ranges, scales and oppositions.

Under Cruse and Togia's system [10], Complementarities must also be construed as both mutually exclusive and mutually exhausting. Therefore if a reader begins to imagine a third state which is not included in one or other of the complementarities then the relationship must again be re-categorised as an antonym and will therefore demand considerably greater effort from both teacher and learner before a secure assessment result can be obtained. The chosen system for identifying opposites [10] gives detailed descriptions of antonym pairs.

Our example illustration begins by returning to the first experiment when the output of the MCQ test item generator [16], [17] was compared to the manually created MCQ test items on the same topic. The steps taken by the automatic process were:

1) Identify significant terms within the source document
2) Apply a clause filtering module
3) Transform the filtered clauses into questions
4) Use semantic similarity to select distractors to the correct answer.

One of the differences between the objectives of the automatic system and the manual creation process was that the textual patterns that were applied to source texts by the automatic system during step 3 are applicable to many educational contexts. However, when operating upon the policy documents in this company a greater emphasis usually needs to be placed upon testing for factual knowledge, as opposed to conceptual knowledge. This can be seen from a comparison of MCQ test items that have been created manually by industrial trainers and MCQ test items that were generated by the automatic system during initial experiments:

Source Sentence:
"Make sure you complete all sections of the diary page. In the 'Work Carried Out' section you must give comprehensive details of your day's achievements."

Manually created question:
"A brief description is all that is required in the Work Carried out' section - True or False?"
(Correct: False)

Generated question:
"What kind of details of your day's achievements must you give in the 'Work Carried out' section"
(Correct: Comprehensive)

3.6. Summary of the CAREGen methodology

The next section describes and then illustrates how skilful narrowing of the range of construal operations offered to those answering MAC test items can reduce the level of Bloom Taxonomy [1]. Cognitive learning required to arrive at the correct response. More specifically, the Construed Antonym Realisation Exercise (CARE) creation process allows designers to identify the salient features within potentially very complex Antonym ranges [2] and rationalise them into relatively simple domain-specific and construal-specific Complementarities which can be traced back to exact phrases within the source documents.

The item design processes can then identify alternative construal operations that might be applied

by readers, and prompt decisions about whether each of the construal operations is either erroneous or correct in the context of the specified CSLO [6]. All 'correct' construal operations that a learner might apply to the stem are associated with the 'correct' response option and all incorrect construal operations are associated with the 'incorrect' response option.

The steps of the CAREGen process are as follows:

Table 2. Steps of the CAREGen process

Step 1)	Define Objective of the MCQ routine in a CSLO
Step 2)	Identify the most appropriate source documents
Step 3)	Explicate (and if necessary Add) Coherence Relations for sentences that meet certain criteria and then re-work them into CRST-compliant CSLOs
Step 4)	Extract candidate antonym pairs for each of the identified sentences
Step 5)	Apply construal operations in the context of identified antonym pairs
Step 6)	Generate AC item sets by inserting generated components into a template

There follows an Illustrative example in which the methodology is applied to the Cable jointing case study.

4. Author name(s) and affiliation(s)

The Cable Jointing case study provides a domain that has a rich variety of illustrative work situations and domain specific entities, relationships and boundaries. Members of the target audience are likely to use a combination of the full range of learning styles as they interpret the documented rules for this domain. The quoted examples demonstrate how reader mis-construal of written instructions can sometimes be anticipated. The examples also show the power of MAC formatted MCQ test items when they are used to explicate different construal operations that might be applied by different readers to the same source sentence.

4.1. Step 1 - Define Objective of the MCQ routine in a CSLO

The first step in the application of the CARE creation process is to define the objective of the test routine. For example the CSLO for the cable jointing case study is to:

"Identify candidates who can correctly recognize LV general jointing procedures for live and dead working (ST:CA1C)"

4.2. Step 2 - Identify the most appropriate source documents

The second step is to use the CSLOs to identify appropriate source documents. It might seem that the above CSLO makes this easy because the identifier for a single source document is included. However there is a good chance that this document refers to further subsidiary documents and it is important that those documents are also identified and added to the source document collection.

For the purposes of this case study the source document will be regarded as just the single document referred to in the CSLO.

"Sentences contained within a date-specific version of ST:CA1C 'Relating to General Requirements Low Voltage Jointing"

4.3. Step 3 - Explicate (and if necessary Add) Coherence Relations for sentences that meet certain criteria and then re-work them into CRST-compliant CSLOs

The third step is to apply the Explication and Addition components of the CREAM method [7] to sentences which contain

(a) lexemes that have some semantic relationship to the CSLO
(b) lexemes that indicate coherence relations that can be reliably manipulated into MAC stems
(c) textual patterns which indicate implicit or incomplete coherence

The illustrative example was generated following the identification (within the document whose reference number is quoted in the CSLO) of the lexeme 'if' and 'do not' but missing the lexeme 'then'. These features indicate that the sentence can probably be manipulated into a MAC but first the implicit 'then' needs to be positioned correctly.

"If the flame is accidentally extinguished do not attempt to relight the gas and do not allow any naked flame near, until the accumulated gas has been dispersed by opening ventilators, tent flaps or doors"

ST:CA1C – General Requirement 1

Before this sentence could be put forward to the next step, the 'Condition' Relation had to be Explicated by adding 'then', and the sentence did not make sense without also 'Explicating' using a lexeme from the title of the section from which it was taken. The

result of the 'Explication' action taken in step 3 of the CARE creation process was:

"If the flame of a gas torch is accidentally extinguished then do not attempt to relight the gas and do not allow any naked flame near, until the accumulated gas has been dispersed by opening ventilators, tent flaps or doors"

ST:CA1C – General Requirement 1 (modified)

4.4. Step 4 – Extract candidate antonym pairs for each of the identified sentences

The fourth step is to identify lexical items within the list of extracted sentences that co-locate within syntactic patterns that have been previously identified as likely to co-locate with 'opposite' word pairs. Complementarities [10] are awarded the highest scores while increasingly complex and ambiguous antonyms are awarded progressively lower scores. Complementarities and Antonyms are relations between construal operations, not between lexical items [12] and [10].

4.5. Step 5 – Apply construal operations in the context of identified antonym pairs

The fifth step of the CARE Generation methodology gives more detailed guidance about the Manipulation step than is specified in CREAM [3]. An attempt is made to perceive each of the sentences from the source document through as many of the construal operations as are sensible. Each of the identified antonym pairs is then applied to form an AC item set. One of the created MAC test items includes the result of applying a Spacio-temporal mis-construal to the source sentence:

"The use of the disposable plastic gloves provided should eliminate any contact, but if it occurs, the affected area should be washed immediately with plenty of soap and water."

ST:CA1C – General Requirement 37

.. in the context of the antonym pair: "Correct action vs NOT the correct action"

4.6. Step 6 – Generate AC item sets by inserting generated components into a template

The CARE creation process concludes by using the 'correct' and 'incorrect' statements and associated antonym pairs to generate AC item sets. AC item sets which use the same antonym pair are grouped together in order to produce MAC test items ready for evaluation. The MAC test item that resulted from the illustrative example process application in this section is provided below:

5. Evaluation exercise

This section describes the latest evaluation of the CARE generation process in action. The Corpus boundary was defined as:
"Sentences contained within a date-specific version of ST:CA1C 'Relating to General Requirements Low Voltage Jointing"

5.1. Hypothesis

The hypothesis is that test items created using the CARE creation process are indistinguishable from manually created MCQ test items. This will have been proved if the domain expert selects an equal or greater number of AC item sets that were generated using the CARE creation process as manually created items for inclusion in a test routine designed to:
"Identify candidates who can correctly recognize LV general jointing procedures for live and dead working (ST:CA1C)"

5.2. Method

The application of the CARE generation methodology was achieved within a simulation as opposed to a reprogramming of the question generator in order to ensure careful and thorough application of the steps as described in the CAREGen Methodology section. This simulated run of the MCQ test item generator produced 68 AC item sets which were paired up with 68 manually created item sets covering equivalent content before being presented for evaluation.

5.3. Evaluation

The final selection by the domain expert was to consist of 68 AC item sets, which addressed the Controlled Specific Learning Objective for the routine. The domain expert had no involvement in the creation of either the manually or automatically generated items and had no prior knowledge of which were generated AC item sets, therefore these factors could not have any bearing upon his decision about which item sets to include in the test routine. The following usability scores were used to record the domain expert's assessments of the items:
A= Use the AC item set unchanged
B= Change the AC item set Antonym pair
C= Change the AC item set Statement
D= Do not use the AC items set

5.4. Results

On the day of the evaluation the 136 AC item sets were presented to the domain expert who then compiled a routine, often using a combination of generated and manually created AC item sets to produce the final MAC test item. The number of AC item sets that were changed to make them usable varied considerably as each MAC was constructed, and in several cases one of the four CARE generated items was the 'inspiration' for the new manually created items. These were counted as 'changed', manually created AC item sets (Category B) and a corresponding number of the original manually created AC item sets for this content set were discarded.

Once the usability categories were assigned for each of the 68 generated AC item sets, the following comparison table was produced:

Table 3. Domain expert decisions for CAREGen AC item sets vs Manually created AC item sets

	Generated AC Item sets	**Manually created AC item sets**
A=Use unchanged	**28% (19 sets)**	**48% (32 sets)**
B=Change Antonym Pair	**13% (9 sets)**	**0% (0 sets)**
C= Change Stem	**12% (8 sets)**	**0% (0 sets)**
D= Do not use	**47% (32 sets)**	**52% (36 sets)**

6. Conclusions

The CARE generation methodology has been described, illustrated and applied. Although it was time consuming, the process of generating new questions was an interesting exercise leading to a variety of suggestions for improvement. By contrast the method used for evaluating the output from the system, seemed slow and pedantic.

The design was also vulnerable to corruption of results by domain experts 'recognising' a question they had designed themselves or applying quite subjective criteria in deciding whether or not to include a particular item. Considerable effort and time was required from researchers and domain experts while setting up the evaluation exercise in order to minimise the chance of results being spoiled. The impression remaining at the end of this evaluation process was 'there must be a better way'.

However the most encouraging outcome from the evaluation is that there was a higher number of manually created AC item sets that were excluded from the final routine than the number of excluded CARE generated AC item sets. The fact that a significant number of the generated AC item sets required changes before they could be used is unfortunate, however these item sets will provide useful guidance in the construction of an automated implementation.

This article provides a foundation for future research into the MCQ generation task in this domain. There are many suggestions for future research projects ranging from refinements to the manual MCQ generation process by pre-processing source documents and post processing the generated MAC items, through to the development of the software solutions.

However, the most important conclusion arising from this study is that a new measure of MCQ generation efficacy is required if there are to be effective comparisons of systems that tackle the MCQ generation task in the future.

7. Recommendations

The most important point that must be made following this research is that a new measure of efficacy is required if the CAREGen methodology is to be compared with other systems that tackle the MCQ generation task. This new efficacy measure must allow for domain specific characteristics of the objectives of the routine that might be different to the objectives of efficacy measures that have been designed for other domains. The paper should define the measurements that need to be taken and subsequent calculations that need to be made in order to measure the efficacy of the methodology.

Apart from the possibilities of creating automated systems to implement the process, made possible by the ever increasing power of NLP tools, other significant points that arise from this research are offered by splitting out the parts of the Generation process.

Thus variants of each of the CREAM, antonym extraction and construal operation identification processes might gainfully be applied in preprocessing steps that are applied to ALL source text sentences or to a wider range of sentences selected by declared relation type (eg conditional relation) or to a wider range of declared relation types (once the sub-steps have been defined for them).

There are also unexplored possibilities for applying variants of each of these processes to post-process the generated MAC items by insisting upon appropriate coherence relations between the stem and the alternative responses.

8. Acknowledgements

I wish to thank Steve Loveridge and Steve Davies from my employing company Western Power Distribution and my Phd supervisors Dr Le An Ha and Professor Ruslan Mitkov from Wolverhampton University for their continued guidance and support.

9. References

[1] Haladyna, T.M., Downing, S.M., Rodriguez, M.C., (2002) "A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment APPLIED MEASUREMENT IN EDUCATION,15(3), 309–334

[2] Foster, R.M. - "Improving the Output from Software that Generates Multiple Choice Question (MCQ) Test Items Automatically using Controlled Rhetorical Structure Theory" RANLP 2009, Borovets – Student Conference

[3] Foster, R.M. – "Automatic generation of Multiple Alternate Choice (MAC) test item stems by applying Causal Relation Explication, Addition and Manipulation (CREAM) to pre-processed source documents" – London International Conference on Education 2010

[4] Foster, R.M. - "Creating a High Voltage Cable-Jointing Knowledge check using the CARE Generation Methodology" – London International Conference on Education 2011

[5] Foster, R.M. - "Multiple Alternative Choice test items (MACs) deliver more comprehensive assessment information than traditional 4-option MC test items" – London International Conference on Education 2010

[6] Mager, R., "Preparing Instructional Objectives (2nd Edition)". Belmont, CA: Lake Publishing Co. 1975

[7] Mann, William C. and Sandra A. Thompson (1988). "Rhetorical Structure Theory: Toward a functional theory of text organization."

[8] Sanders, T.J.M., Spooren, W.P.M., & Noordman, L.G.M. (1993) "Coherence relations in a cognitive theory of discourse representation." Cognitive Linguistics, 8, 93-133.

[9] Paradis, C. (2010) "Good, better and superb antonyms: a conceptual construal approach" - "The annual texts by foreign guest professors" - Faculty of Arts, Charles University

[10] Cruse, D.A. & P. Togia. (1995). "Towards a cognitive model of antonymy". Lexicology 1: 113-141.

[11] Vogel, Anna (2009) A cognitive approach to opposites: The case of Swedish levande 'alive' and död 'dead' - "Approaches to Language" - Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki http://www.helsinki.fi/varieng/journal/volumes/03/vogel/

[12] Croft, William & D.A. Cruse. (2004). Cognitive Linguistics. Cambridge: Cambridge University Press. (ISBN-10: 0521667704; ISBN-13: 978- 0521667708).

[13] Schultz K.F., Altman D.G., Moher D.; for the CONSORT Group (2010). "CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials" Br Med J 340:c332. DOI:10.1136/bmj.c332

[14] Gronlund, N., Constructing achievement tests. New York: Prentice-Hall Inc. 1982

[15] Swanson D.B., Holtzman, K.Z.,Allbee K.,Clauser, B.E., "Psychometric Characteristics and Response Times for Content-Parallel Extended Matching and One-Best-Answer Items in Relation to Number of Options." 2006

[16] Mitkov, R., and L. A. Ha. (2003). "Computer-Aided Generation of Multiple-Choice Tests." In Proceedings of HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing, pp. 17-22. Edmonton, Canada.

[17] Mitkov, R., L. A. Ha, and N. Karamanis. (2006). A computer-aided environment for generating multiple-choice test items. Natural Language Engineering 12(2): 177-194.

[18] Crooks, T., "The Validity of Formative Assessments". British Educational Research Association Annual Conference, University of Leeds, September 13-15 2001

[19] Fillmore, Charles J. (1985). "Frames and the semantics of understanding". Quaderni di Semantica 6: 222-254.

[20] Johnson, M. (1987) "The body in the mind" Chicago: University of Chicago Press

[21] Lakoff G., and Johnson, M. (1980) "Metaphors we live by" Chicago: University of Chicago Press

[22] Lakoff, G. (1987) Women, fire and dangerous things: what categories reveal about the mind.

[23] Talmy, L. (2000) "Toward a cognitive Semantics" vol 1 – Concept Structuring Systems

[24] Bloom, B. (1956), Taxonomy of Educational Objectives: Book 1 Cognitive Domain, Longman, 1956.

[25] Butler A.C., and Roediger H.L. III, 2008, Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing Memory & Cognition 36 (3), 604-616, doi: 10.3758/MC.36.3.604

