

Text Readability within Video Retrieval Applications: A Study On CCTV Analysis

Neil Newbold

University of Surrey, Guildford, UK

Email: n.newbold@surrey.ac.uk

Lee Gillam

University of Surrey, Guildford, UK

Email: l.gillam@surrey.ac.uk

Abstract— The indexing and retrieval of video footage requires appropriate annotation of the video for search queries to be able to provide useful results. This paper discusses an approach to automating video annotation based on an expanded consideration of readability that covers both text factors and cognitive factors. The eventual aim is the selection of ontological elements that support wider ranges of user queries through limited sets of annotations derived automatically through the analysis of expert annotations of prior content. We describe how considerations of readability influence the approach taken to ontology extraction components of the system in development, and the automatic population of a CCTV ontology from analysis of expert transcripts of video footage. Considerations are made of the semantic content of the expert transcripts through theories on readability analysis and terminology extraction to provide knowledge-based video retrieval. Using readability studies to improve the text, we suggest that the semantic content can be made more accessible and improves the terminology extraction process which highlights the key concepts. This information can be used to determine relationships in the text, as a proxy for relationships between video objects with strong potential for interlinkage.

Index Terms—video retrieval, inter-annotation, readability, automatic annotation

I. INTRODUCTION

Multimedia retrieval is of increasing importance in the ever-expanding web, particularly given the increased prevalence of multimedia data. The increasingly available volumes of multimedia place a substantial burden on end users in understanding and filtering such data. Much current multimedia retrieval is text, or metadata-driven, with but a few extant examples of retrieval using low-level features of images or videos. Unlike text, where relevant segments can be readily highlighted and summaries are relatively easily produced, segments and summaries of videos are not so readily or rapidly produced. Semantic Web technologies, which some refer to as Web 3.0 or the Executable Web, are promised that analyze and automatically annotate and relate many different kinds of multimedia documents, including video, image, music and sound, text, and various

fragments of these. This process of auto-annotation involves various vocabularies and classifications, often referred to as ontologies and folksonomies. Auto-annotation is fundamental to the success of the semantic web in general, and multimedia indexing and retrieval in particular, since the cost of manual annotation is substantial. Given the current limits of reliability of metadata on standard web pages, where it is often not even be possible to trust the metadata that indicates the language of the page, automatic provision of (standardized) metadata of various kinds would be beneficial to the wider community.

Auto-annotation could be particularly beneficial where large volumes of multimedia data are produced on an ongoing basis and variously retrieved, and where it is currently only possible to assign a limited amount of metadata. One such application is *crime fighting*, involving extensive deployment of Closed Circuit Television (CCTV) systems. CCTV is an increasingly popular way of inexpensively enabling remote monitoring and policing of specific (visual) scenes such as ATMs, retail outlets, football (soccer) matches, in airports, in railway stations, in car parks, and so on. The expected continuation in the growth of multimedia data, and especially in CCTV that is allied to increased image resolution and incorporation of audio capture, suggests potential challenges will arise for those wishing to identify participants, key scenes, activities and events within several thousands of hours of footage taken using multiple cameras from different angles and perspectives. While it is increasingly possible to identify specific sets of information directly from video frames in certain application contexts, and to use such information for queries against video collections, pure video processing to date has only achieved specific successes in recognizing particular kinds of objects in very specific scenes. Bridging much of what researchers refer to as the Semantic Gap - understanding the relationships between identified objects and things in the real world - is still a way off. Such consideration of bridging has been a cornerstone of the DARPA-sponsored TRECVID initiative [1]. It has also led to auto-annotation that makes use of extant textual descriptions in which the analysis of the descriptions provides the basis for annotating objects

and events within unseen video footage. The derivation of key concepts and their terms from these texts and other related, or collateral [2], texts is occasionally used in combination with whatever information can be extracted from the visual scenes. Elsewhere, free text descriptions, such as image captions, are considered as valuable sources of annotation information, not least in applications such as Google Images where they are used as the only searchable element.

A review of work in the semantic gap, relating to multimedia, can be found in [3] that covers, for example, using a co-occurrence model for keywords and low-level features of rectangular image regions, approaches that segment images into regions and so on [4]. Another approach, described in [5], involves processing text available with multimedia to extract terms by combining typical information retrieval approaches of word frequency, TFIDF and entropy, and using stemming. In this approach, keywords above a threshold are used to manually construct an ontology. Relationships in the ontology are discovered using association rules, with relevant relationships also manually selected and incorporated. Reference [5] claimed that constructing ontologies using purely textual information is not adequate: it is unclear where they consider the inadequacy to lie, though they note that issues such as different correct specifications for the same domain may be a factor. Perhaps the problems arise from the quality of the text being processed. Our approach aims at producing features suitable for a co-occurrence model, and attempting to build on [5] by automatically populating the ontology. Current research into video annotation, and particularly this kind of auto-annotation, is sometimes undertaken within the rubric of so-called “multimedia ontologies”, and it is with the construction of a multimedia ontology for CCTV that we are principally concerned.

A brief consideration of some of the work on multimedia retrieval is presented in Table 1. Typically, multimedia ontologies are built for only one purpose: identifying activities against a largely green background in football matches, or identifying specific shots of known people, explosions, and other such features of broadcast news. Some transition across applications may be possible, e.g. for sports reports in broadcast news, but generally the adaptation to other purposes outside the scope of the original application requires significant retraining of various systems. Additionally, there are obvious differences amongst these application areas: in football matches and broadcast news, the participants are largely known; in CCTV, they are largely unknown. While broadcast news will have a specific running order, and football matches are scheduled, important activities in CCTV may occur at any time and need to be detected – indeed, there may be extended periods of time during which nothing of interest is happening.

In this paper, we discuss how recent considerations of readability [9] that account for both text factors and cognitive factors may assist our construction of the

multimedia ontology. This relates to two aspects of our work:

1. Accounting for inter-annotator variability in keyword use and grammatical construction to lead towards more “machine-readable” text.
2. Considering terminology extraction techniques as an integral part of the assessment of readability.

To automatically and reliably extract the significant concepts from expert transcripts of video footage, we would ideally have clear, concise, and consistent text. Convolved, unnecessarily verbose, and inconsistent text can produce difficulties for identification of semantic content, and many systems require extensive manual training and/or “eyeballing” of texts to circumvent this. The effectiveness of these systems often depends on the training, and the cost of adaptation often limits the use of such systems across domains. Part of the motivation for our work, then, is to be able to improve the likelihood of success of the extraction by making consideration of the machine-readability of the text. In Section 2, we discuss readability research and how it can be applied generally to information retrieval; Section 3 describes our system for automatically capturing semantic content using readability techniques; Section 4 outlines results on parallel annotations; Section 5 concludes the paper and makes some considerations for future work.

TABLE I.
RECENT WORK ON MULTIMEDIA ONTOLOGIES INVOLVING VIDEOS AND TEXT

	Soccer Ontology	TRECVID	REVEAL
Domain	Football	Broadcast News	CCTV
Analysis	Video	Video	Video + Text
Algorithms employed	Fuzzy c-means clustering algorithm, sum of all Needleman-Wunch distances, Bertini’s annotation algorithm	Gabor texture (image properties), Grid Color Movement (colour distribution), Edge Direction Histogram (geometric cues)	Video: Motion detection, motion tracking, colour classification, geometric classification and object classification. Text: Linguistic, Statistical, Synonym and Known Terminology Analysis
Resources used	Videos of 3 football matches (World Cup & UEFA), MPEG-7, OWL	Kodak Video Collection (manually annotated), YouTube videos (manually annotated), MPEG-7, OWL	CCTV footage, transcripts of police descriptions, MPEG-7, OWL
Publication	Reference [6]	Reference [7]	Reference [8]

II. BACKGROUND

Reference [10] defined readability as “the ease of understanding or comprehension due to the style of writing”. According to this definition, the style of writing is paramount, and the specific abilities of the reader are not important. Historically, readability research has focused primarily on producing a numeric evaluation of style of writing. Work on readability stems largely from an initial study by [11] who demonstrated differences in sentence lengths and word lengths, measured in syllables, between two newspapers and two magazines. Reference [11] suggested that differences in lengths accounted for differences in readership.

This work led variously on to the development of other metrics for readability, many of which are popular in modern office software applications. A variety of readability metrics now exist [12], [13], [14] and [15]. These are based on different considerations of sentence length and word length, and in some cases as a function of the number of syllables. The results of applying these formulae relates to some attempts on the one hand to indicate the level (age) of education, and on the other to provide a difficulty score on a scale of 1-100. These formulae and the elements on which they rely are presented in Table 2 – further discussion of these can be found in [16] and [17]. There are two main types of application for readability metrics:

1. for educators in selecting appropriate material for the target audience’s reading ability or to determine whether feedback comments provided to students will improve learning outcomes [18];
2. for authors in improving and/or simplifying texts when, used to indicate whether they have appropriately targeted their intended audience.

Readability measures are supposed to enable anyone, without special knowledge or training, to determine the proportion of people who could comfortably read a piece of text. This assumes that those using the readability measures are aware of the ideal values they are attempting to obtain. Authors using these metrics may attempt to iteratively simplify technical and scientific documents to ensure they can be understood by wider audiences. However, many readability researchers advise against attempting to influence readability formulae in this way as modifying small amounts of text does not guarantee that texts are any easier to understand. They suggest that readability formulae should be used only for iterative feedback on the entire document [19], though it should be possible to evaluate the impact on the readability score of specific changes to the text. For Natural Language Processing (NLP) systems, their abilities as machine “readers” of text are clearly relevant to how effectively they can process text, and indications of, for example, ambiguities within the text are beneficial to both human and machine readers alike.

For us, the most considered view of readability to date is presented by [9]. However, the authors have only elaborated this view to a limited extent. This view entails

TABLE II.
FEATURES OF THE TRADITIONAL READABILITY METRICS

	Flesch	Kincaid	Fog	SMOG	ARI
Sentence length	✓	✓	✓	✓	✓
Characters/word					✓
Syllables/word	✓	✓			
Complex words count (more than three syllables)			✓	✓	
Scale	0-100	US Grade level	US Grade level	US Grade level	US Grade level
Ideal outcome	100	7-8 (Ages 13-14)	7-8 (Ages 13-14)	7-8 (Ages 13-14)	7-8 (Ages 13-14)

decomposing readability into two initial considerations: reader factors, which consider the ability of the reader, and text factors, which consider the formulation of the text. These factors, and the components considered against each, are shown in Fig 1. Reader factors include the person’s ability to read fluently, whether they have sufficient background knowledge in the subject, their lexical knowledge or familiarity with the language, and whether they are suitably motivated and engaged in the subject matter. Text factors appear in part to account for metrics considered as “readability” today, but also cover considerations of syntax, lexical selection, idea density, and cognitive load - the effort required by the reader to correctly interpret the text. This view of readability would imply that an overall measure of “text difficulty” can be produced that depends only on text factors, but this measure is going to vary depending on each individual reader. In the remainder of this section, we elaborate this view, identifying some difficulties in interpreting the factors, and attempt to relate these various factors to the text-based (knowledge-based) retrieval of video content by consideration of machine readability.

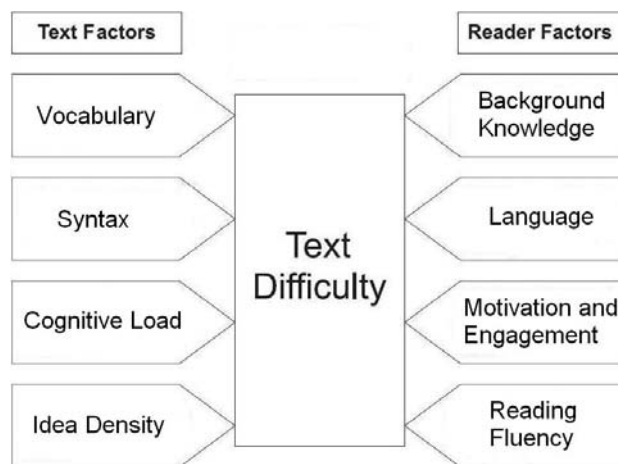


Figure 1. Framework for readability, encompassing text factors and reader factors described in [9].

A. Text Factors

As discussed, readability measures can be used as objective measures of text difficulty, but they may be somewhat artificial constructs predicated only on everyday use of language. When dealing with specialized, i.e. domain-specific/subject-specific use of language, a wider variety of considerations is certainly needed to account, at least, for the inherent specialized terminology, and that implies a deeper analysis. The existing measures of readability do not account for any kind of “conceptual” difficulty: by some of the formulae, Einstein’s theory of relativity reads for ages 10-11. In addition, they cannot check the text is syntactically or semantically acceptable; “a man walks across the street” will score identically to “a street walks across the man”. Indeed, both will score identically to “walks across a the man street”.

The readability formulae are also demonstrably inconsistent if used to consider changes to a specific piece of text. A particular modification to the text may improve readability according to some measures, and make it worse according to others. This is a consequence of the differences in weighting on the input factors. For example, the sentence, ‘Further funding comprises an element of additional financing for those institutions which have a high historic unit cost’ contains the word ‘comprises’. If we substitute this word for the phrase ‘is made up of’, all the readability formulae will register more readable text, apart from the FOG index which will suggest this change made the sentence slightly harder.

1) Vocabulary

Reference [9] cite factors such as simplicity or familiarity as more effective means of measuring the difficulty of a word than by counting characters or syllables. They suggest that word difficulty can be determined by examining whether a word is challenging, unusual or technical, and suggest **vocabulary** as a text factor contributing to text difficulty. If the text is populated with difficult words then it becomes harder to read causing readers to complain about ‘jargon’. Here, the word ‘jargon’ has a negative connotation implying unnecessary overuse of complex terms where the same scientific concepts could be expressed in non-technical terms without loss of understanding. The problem with unfamiliar words for a novice is that they may become part of the jargon they use as an expert and forget how to write for novice audiences.

In scientific domains, terminology collections provide definitions for a large number of terms, each of which represent a specific concept. The meaning of these specialized semantic units may be difficult to deduce accurately for a novice reader. In some cases, a terminology collection may be provided alongside some documents to help in understanding the specialized documentation. This is particularly true for international standards (ISO) where the construction of the terminology should be a vital consideration for avoiding ambiguity and accurate application. A terminology collection may be essential both for novice readers hoping to get to grips with a domain and for those applying standards in order to do so accurately.

The European Association of Aerospace Industries (formerly AECMA, now ASD) developed Simplified Technical English [20] (formerly AECMA Simplified English), a specification for aircraft maintenance, to ensure non-native speakers of English did not create potentially dangerous situations through misinterpreting documentation. The specification includes a dictionary providing a limited vocabulary for use in their documents. Each word in the dictionary has only one meaning, for example the word ‘drive’ is always used in the mechanical sense such as ‘the drive was faulty’. The word can never be used to describe a journey such as ‘the drive was boring’. Having a predefined description for a word can help avoid confusion

The terminological nature of such documents means that the terms are used disproportionately frequently throughout the documents, and more often than one would expect to encounter in everyday language. While, this means that anyone unfamiliar with the terminology would find the vocabulary used in the document hard to understand, it also means that terminology can be identified as words used frequently in the document but are unfamiliar in general use. From this perspective, vocabulary could be measured in at least two ways: (i) as a metric relating to the relative simplicity of the words used, in relation to Simplified English; (ii) a measure of the frequency of the words within the document in accordance to their familiarity in general language.

2) Syntax

The vocabulary, however, does not tend to exist in isolation, and to understand why text can easily become difficult we may also consider grammar. The vocabulary may be well-defined, yet included in overly verbose sentences. Furthermore, the relationships between the terms may hinder understanding. Reference [9] suggest that long and complex sentences, and unfamiliar syntactic structures, can confuse the reader. They include **syntax** as a text factor contributing to text difficulty.

Words tend not to combine randomly or freely, rather they are used with preferred ‘friends’. According to [21] “you shall know a word by the company it keeps!” and this company may be evident in combinations as certain kinds of ‘collocations’. Collocations demonstrate the preference for friends, and importantly have “distant” friends, i.e. there is significance to the distance and order between the collocating words: for example, the collocation between “bread” and “butter” is rather more frequently encountered as “bread and butter” than as “butter and bread”. In addition, expected synonyms may be largely excluded as friends, so while we have ‘strong tea’ we appear to have rather less by way of ‘powerful tea’. However, a reader unfamiliar with such constructions might not understand the precise meanings or variations. Put another way, individual words may not be particularly difficult but their combination can produce different meanings to those that component words might suggest. This is readily demonstrable in general language, where a ‘tyre’ is thought of as an object of a circular nature but ‘flat tyre’ does not refer to a flat circle, rather it indicates a lack of air pressure.

Collocations have been widely discussed in the linguistic community, though there is little consensus as to their exact nature. Some researchers refer to them as n-grams, depending on the size of 'n' (e.g. bigrams, trigrams and so on), or multiword expressions (MWE). Most researchers agree that collocations are sequences of words that co-occur more often than by chance, assuming randomness, which can be found using statistical measures of association. Some linguists consider collocations are the building blocks of language, with the whole collocation being stronger than the sum of its parts. They describe collocations as lexical items that represent uniquely identifiable concepts or semantic units. Reference [22] elaborated on the criteria for a collocation, describing them as recurring and cohesive domain-dependent lexical structures. They used examples of 'stock market' and 'interest rate', and suggested how components can imply collocations, for example 'United' produces an expectation of 'Kingdom', 'Nations', or 'States'.

In scientific writing, collocations can become complex, providing a more specified meaning, but as a result become harder to understand. Consider, for example, 'glass crack growth rate'. Each word will achieve a good readability score and, on a word-by-word basis this, should be easy to understand. However there are several possible interpretations, due to bracketing (see, e.g. [23]) that might lead us even to consider a 'crack growth rate' made of 'glass'. This is more readily apparent in examples demonstrated using documents such as draft international standards [24]. Reference [25] described such kinds of difficulty as 'syntactic ambiguity', one of a series of categories of common problems in writing scientific English. The cause of this syntactic ambiguity is the passive voice, which has evolved through scientific discourse despite it opening text up to ambiguity. The use of passive voice can create groups of words orientated on a noun resulting in long collocations. ASD Simplified Technical English recommends writers avoid lengthy collocations by breaking up these noun-centered expressions. For example, instead of 'runway light connection resistance calibration', you should write 'calibration of the resistance on a runway light connection'. This unpacking of semantics helps to remove ambiguity arising from bracketing and may also be beneficial for translation to certain other languages such as French. In attempting to derive an approach to semantic unpacking, considerations of adjacency and dependency [26] may be of interest to determine the roots of the packing. The Plain English Campaign [27] offers rules and techniques that are intended to reduce ambiguity, improve understanding, increase reading speed and ease translation. In particular, the campaign provides an A-to-Z guide of over 1300 plain English substitutions for supposedly difficult words and phrases. According to this list, and depending on sentence structure, "essential" could be replaced with either "important" or "necessary" and "according to our records" could be substituted for "our records show".

Misuse of syntax can result in long noun-orientated expressions within the text, and in overly verbose text fragments which it may be easily possible to simplify. It could be possible to consider syntax, in this case syntactic complexity, as measurable by considering the extent of collocations, the complexity of phrasing, and the use of passive voice in the text.

3) *Cognitive Load*

Reference [25] also identified 'lexical density' as another category demonstrating problems with scientific English. This category applies when a significant amount of information is conveyed in a relatively small amount of text and is demonstrated in the following example with the words in bold being considered as lexical,

"My **father** used to **tell** me about a **singer** in his **village**."

The four lexical words represent the semantic content in the text. Lexical density can vary from one clause to the next in all types of writing. Informal spoken language has a lexical density around two words per clause; with more formal and planned written language having a lexical density of four to six lexical words per clause and scientific language, reaching 10 -13 words per clause. The use of sentence length and word length by [11] tends to be relatively successful since longer words tend to be lexical and longer sentences are more likely to more lexical items. Although it is easy to assume, high lexical density isn't necessarily an indicator of poor readability. Long collocations that form semantic units reduce conceptual complexity. Problems occur when numerous semantic units are described within a short space of each other causing the reader to make numerous inferences. The semantic units do not need to be limited to just one clause, it is the overall analysis and reasoning required that can cause confusion. Reference [9] describe this as **cognitive load**.

This problem often occurs when writers assume that adding further defining clauses into their text makes it text easier to understand, when in fact the opposite is true. If a reader fails to understand one of the first concepts, subsequent definitions are also unlikely make sense. The writer is also intimidating the reader by referring to them in such short space of each other. This implies that these terms are simple and should be quickly understood. It is this fast succession of defining clauses which can cause problems. The amount of ideas expressed in the text contributes to cognitive load by increasing the work demanded of the reader by authors to interpret their text correctly. Perhaps cognitive load is measurable by examining (i) the quantity of defined and undefined terms within short distances of each other; (ii) indications of defining clauses within the text. This will determine the workload required by a reader to process or interpret the text correctly.

4) *Idea Density*

So-called **idea density** manifests when writers present new information to the reader without making clear its relationship to previous information: the writer assumes

that they have provided enough information to allow readers to follow their arguments logically. While this poses no problem for specialists, it can often be intimidating for novices. Writers often expect the reader to make ‘semantic leaps’ [25] from existing understanding to understand particularly abstract ideas and conclusions, and this may lead to incorrect inferences. Idea density is linked to vocabulary (terminology) in that an expert will find it easier to associate the content of the text. It is easy to confuse idea density with cognitive load, but where cognitive load is concerned with the number of ideas in the text, idea density refers to ‘strength’ or ‘abstractness’ of the ideas.

Idea density may be related to considerations such as lexical cohesion [28], where repetition of a lexeme and its synonyms provides a structure for the reader to connect with. Repetitious patterns help readers form an understanding throughout the text. Sentence links and bonds enable summarization and allow consideration of the overall characterization of the text. If a large number of new, seemingly unrelated ideas are being introduced, this should be evident in low cohesion. Perhaps lexical cohesion can provide us with a measure of idea density.

B. Reader Factors

Text factors presume that difficulty is an artifact of text. However, different readers will have different views of the same piece of text. Reader characteristics may amplify or negate problems with difficult text. For a variety of reader factors identified [9], we consider that it would be necessary somehow to capture and analyze the user’s experience with prior documents as a proxy for reader knowledge, and that the capture and analysis would lead towards a personalized assessment for the document.

1) Background Knowledge

Although many readability metrics determine a grade level of an audience capable of understanding the text, they make no distinctions according to the background knowledge of the reader. As discussed in **vocabulary**, word familiarity gives a much better indication of word difficulty than word length. A longer word may only be difficult for a particular reader and certain shorter words may be more difficult to understand for wider audiences. Consider a general reader confronted in text discussing a ‘muon’. The term is short and would be rated as simple by the current readability formulae. However most people would be unfamiliar with this term and only particle physicists are likely to know the term, its definition, and related items. A reader well-versed in a particular subject field should find the words rather more familiar and [9] suggest that **background knowledge** contributes to text difficulty.

Reference [29] described a series of studies conducted by the U.S. military showing how prior knowledge affected readability. In a manner similar to the Plain English Campaign, they simplified and changed the style of technical documents while experts ensured that all the technical terms were kept and that the intended message was not changed. The simplified versions resulted in faster reading speeds and greater retention of information,

but differences were only noted in readers who were naïve in the subject. There was little observed benefit for the experts. Reference [30] followed up these experiments to demonstrate that more readable text is beneficial for those with less knowledge and interest. The problems of difficult text are effectively ‘drowned out’ by knowledge of a subject. However, it is not easy to measure the amount of background knowledge required to differentiate between levels of knowledge; [16] queries the measurement used in the conclusions from results of reading tests – is this a reflection on comprehension, prior knowledge, memory, or just the difficulty of the question used in the reading test? More generally, the reader’s background knowledge needs to be captured and measured somehow.

Perhaps a terminology collection and its definitions can in some way, reflect the background knowledge required by a reader to interpret the text correctly. Knowing the precise meaning of certain words can help distinguish some of the ambiguity of the surrounding words. One way to measure background knowledge would be through the extent of use of terminology in the text with consideration of previous documents within the reader’s experience.

2) Motivation and Engagement

Reference [30] showed that more readable text is beneficial for those with less knowledge and interest. In another part of their study, students were presented with written material below their reading level. When the reader’s interest was high, text below their grade level did not improve comprehension. However, when the reader’s interest was low their comprehension was improved by simpler text. This suggests that more readable text improves comprehension for those less interested in the subject matter. Reference [9] characterize this as **motivation and engagement**. A study by [31] showed that experiments using readability formulae to simplify texts can be skewed by the interests and motivations of the reader: readability is more important when interest is low. Some researchers argue that the link between text comprehension and motivation is due to the extent of reading performed by the reader. Reference [32] provide evidence that reading motivation predicts quantity of material read, and this in turn predicts text comprehension. In some ways, then, a readability system would need to ascertain whether a reader had demonstrated an interest in similar or related previous material. One way to measure motivation would be to examine a reader’s history for similar documents, building on the measurement for **background knowledge**.

3) Language

Reference [9] identify **language** as another reader factor contributing to text difficulty. The process by which readers develop sufficient knowledge of a language is referred to as language acquisition and concerns familiarity with words and the development of the language capability. First language (or L1) acquisition concerns the development of language in children, while second language (or L2) acquisition focuses on language

development in adults. To measure language development in children, [33] introduced the mean length of utterance (MLU). This measure of language growth stems from observations that most advances of morphological and syntactic skills result in longer utterances by children. Subsequent studies have shown MLU is highly correlated with age for normal children ([34], [35]), and [36] correlated MLU with age for children with specific language impairments. MLU follows similar principles to readability formulae, which all use average sentence length either in words, syllables or characters. The amount of information contained in a statement is considered a measure of complexity by various fields of research.

Researchers have shown that frequency is one of the strongest determiners in acquiring language, but have yet to explain how humans acquire the more abstract forms of linguistic knowledge. References [37] and [38] showed that frequency has an impact on comprehension and the development of language categories, and although it is widely assumed that grammar cannot be learned from experience alone, researchers working on collocations and distributional lexical semantics may produce interesting future insights. Reference [39] stated that frequency is a necessary component of theories of language acquisition but is not a sufficient explanation – otherwise, we would never get beyond the definite article in our speech. Distributional cues are useful for categorizing high frequency items encountered in the identical contexts, and considerations of distributional lexical semantics pay strong heed to this, but these cues are less useful when considering lower frequency words. Reference [40] notes that human language acquisition must involve more than high-frequency input. Frequency, along with other factors, needs to be considered in the framework of a comprehensive theory of the representation, processing and acquisition of linguistic knowledge.

As we discussed in **syntax**, collocations can cause confusion due to syntactic ambiguity. However, research has shown that frequency is indispensable for dealing with these ambiguities. Reference [41] showed how subjects used innate statistical information to determine how sentences should be interpreted. More recently, [42] showed that language users expect a particular word or word category to appear with a particular linguistic expression. These linguistic expressions are stored in memory and are reinforced by frequency to help comprehension. In addition, frequently combined linguistic expressions may develop into a processing unit so that many of the linguistic elements are ignored and the whole chunk is compressed and treated as one piece. This work relates back to collocations, with multiword units representing singular concepts and developing into terminology. Reference [42] concluded that several psychological mechanisms such as information processing and analogy interact with frequency based mechanisms to develop linguistic structure. A person's grammar is an emergent linguistic structure developed from their use of language.

The familiarity with words in language acquisition relates to the factors of **vocabulary** and **background knowledge**: while it is possible to ascribe an overall score for a word, perhaps as a measure of its rarity in discourse, words will have different familiarity for different readers. A difficult word for a novice is not always the same as a difficult word for an expert; a word that is difficult for L2 audiences may not be difficult for L1 audiences. However, beyond this, it is difficult to make a clear distinction between the reader factors of **background knowledge** and **language** unless we make consideration for the non-terminological elements of the text – measuring the reader's familiarity with similar grammatical structures and general language in prior documents.

4) *Reading Fluency*

Reference [9] identify **reading fluency** as the final reader factor contributing to text difficulty. As discussed in relation to language acquisition, the more text a person reads, the stronger their experience-based grammar becomes. This in turn results in a more fluent reader. Research has shown the importance of reading fluency in developing reading proficiency and differences in reading fluency can distinguish between good and poor readers. Reference [43] showed how a lack of reading proficiency is a reliable predictor of reading comprehension problems. There is a strong correlation between reading fluency and reading comprehension with each aspect of reading fluency having a connection to text comprehension. Reference [44] showed that fluent reading consists of three important elements, "accurate reading of connected text at a conversational rate with appropriate prosody or expression." Inaccurate word reading can cause readers to misinterpret the text and limit their access to the author's intended meaning. Reference [45] showed that factors such as knowledge of a large bank of high frequency words are needed for accurate word reading. Words are only analyzed when they cannot be read from memory as sight words. This relates back to **language**: the reader's lack of knowledge of words will affect their reading fluency in that readers are likely to dwell over unfamiliar words or grammatical constructions. This impedes the reader's ability to construct an ongoing interpretation of the text. Experiments in language acquisition have shown that the categorization of word classes can be improved by incorporating phrasal boundaries. The correct placing of pauses around phrase boundaries contributes significantly to their meaning. For example, [46] used the following example to show how an ambiguity introduced into a string of words can produce interpretations that are either meaningful or nonsensical,

"The young man the jungle gym."

The majority of readers pause at 'man', rendering the phrase meaningless. However, if the reader pauses at 'young', they can construct the meaning and interpret the sentence. Reference [47] suggested that fluent readers use morphemic, syntactic, semantic and pragmatic cues

present in the text to organize it into meaningful phrases. This work relates to collocations and the text factor of **syntax** with frequent collocations used to decipher text. In the example, without any punctuation, the frequent collocation 'young man' is used to try and construct meaning from the phrase. In this instance, the collocation leads us to an invalid interpretation rendering the phrase meaningless. It is only with a phrasal boundary dividing the collocation that we can begin to interpret the phrase as it was intended.

Difficult text can have a negative effect on reading fluency, [48] showed how the accuracy, speed and expressiveness of poor readers are more affected by text difficulty than average readers. Poor readers find difficult text harder to understand and as we have already discussed, experts in a particular field do not notice difficult text. **Background knowledge** makes difficult text less noticeable and poor **reading fluency** can make the effects of difficult text more prominent. Perhaps reading fluency can be addressed through the reader's familiarity with general language and the consideration of collocations and phrasal boundaries. The analysis of general language would relate to the factor of **language**.

C. Commentary on the Framework

The framework described by [9] has potential benefit in considering the formulation of applications that require automatic access to semantic content expressed in text. By improving the readability of text and incorporating factors which help human readers understand text, we hopefully increase the likelihood of effective automatic processing – improved machine-readability. Reference [9] discussed, but did not elaborate, a framework for readability, but this needs to be implemented and evaluated. In doing so, comparability to other approaches is vital - it must be possible to derive a meaningful value. It is with being able to derive such a value that our work is currently concerned, however we make considerations here for the kinds of applications that readability supports. Much of our previous work has focused on terminology extraction techniques that can be used to identify the important concepts specific to one or more documents: terms denote the key concepts. The important (frequent) relationships between these key concepts can also be identified. Such analysis, more generally, is usually referred to as ontology learning, and can result in the production of domain terminologies (ontologies). Work on terminology extraction for terminology enhancement has already been undertaken in relation to controlled authoring of international (ISO) standards [24]. Here, extracted terms were considered as likely additions to a standard terminology.

We consider that the terminological component is key to measuring the vocabulary, with the number of terms which the reader has previously encountered relevant for the background knowledge. Both idea density and cognitive load appear to relate to the introduction and packing of terms within the text. Furthermore, while syntax deals with the structuring amongst these, language and reading fluency address the reader's familiarity with the syntax. Finally, motivation and engagement seems to

be consistent with the frequency of previous encounters with the background knowledge. We elaborate these further in relation to CCTV analysis. In this application, we begin to consider the machine-as-reader, in contrast to considering what a human being might know.

III. CCTV ANALYSIS

Closed Circuit Television (CCTV) surveillance has both its supporters and its critics. It is used with increasing frequency to monitor large proportions of the public, to try to detect the rather smaller proportion undertaking or complicit in criminal, or at least antisocial, activities. CCTV systems may be used in shopping centres, on roads and in car parks, in railway stations and airports, and so on. It is estimated that the UK has over 4 million CCTV cameras operating, capturing the movements of its citizens continuously in certain locations throughout the year. The quantities of data being collected are substantial: a set of a few hundred cameras monitoring car parks in a major UK city requires a terascale (around 100TB) data storage facility to cope with retention of about a year's worth of data, just in case there becomes a need for it. Subsequently, it is archived to tape and all these archive tapes also require housing. As capabilities for increased image resolution and magnification improve, and as very small CCTV systems can be cheaply and rapidly deployed, the challenges associated with finding any specific happening within such extensive video collections becomes ever greater. For one event, this could require trawling through footage captured over several days from numerous cameras, with a need to piece together the movements of an individual from a variety of differently positioned and differently located cameras. A multimedia ontology for CCTV, then, would be of particular value. A CCTV ontology would document the semantics of video scenes that cannot easily be captured through video analysis techniques alone. A framework has been constructed specifically for the CCTV domain where high level semantics extracted from video are intended to be associated with concepts extracted from expert descriptions of the video [8]. Visual semantics are extracted using motion detection and tracking, and classification according to geometric attributes of the blobs and the dynamics of their trajectories. This approach can distinguish between people and vehicles, and observed motion is used to confirm the constraints on objects to locations.

We use various text analysis techniques, centered around terminological analysis as described in the previous section, in order to generate the ontology from expert descriptions of video samples. The eventual aim is to use this ontology for auto-annotation of unseen CCTV footage. The CCTV ontology is initially composed of the **vocabulary** representing the objects (O) and actions (A), and this vocabulary is automatically extracted from the expert descriptions. This level of abstraction is insufficiently accurate and produces a large set for consideration. The text is further analyzed for the existence of triples of the form O-A-O, for example 'man-drive-car' or 'person-crosses-street'. Inferences are

made over whether the object is an agent or a recipient. Fig 2 shows, in white, the factors contributing to text difficulty which are directly addressed in this subsystem. At present, the blacked out factors are largely ignored, however there is future scope for their inclusion.

The first task in creating the CCTV ontology is to find all the objects and actions described by the experts in their commentary. We address **background knowledge** by accessing the expert understanding inherent in the text. Key objects are found by annotating terms found in a police ontology for describing crime scenes. The ontology was provided by the National Policing Improvement Agency (NPIA), formerly known as the Police Information Technology Organisation (PITO). Additional objects and the actions performed by those objects are found using terminology extraction techniques. The text factor of **vocabulary** is addressed by using statistical methods to find frequent words which are unusual in everyday language. The text factor of **syntax** is then considered in finding collocations by examining the neighboring words. Both single and multiword expressions are considered as potential objects. The analysis of everyday language incorporates the factors of **language** and **reading fluency**.

Linguistic methods are used to find further objects and potential actions using part of speech (PoS) tagging. Frequent nouns and collocations are considered as objects with recurrent verbs used to determine actions. The analysis of frequent nouns and verbs throughout the text addresses the text factor of **vocabulary** and the consideration of collocations addresses the text factor of **syntax**. Identified concepts are verified using thesauri to ensure the linguistic expressions found are likely valid concepts. This process is modeled on the reader factor of **language**, by using entries in the thesauri to model the reader's syntactic development or familiarity with general language. WordNet [49] was chosen as it is currently the most comprehensive lexical database.

Once all the objects and actions have been identified, their proximity to each other in the text is used to determine when important events have occurred in the video. This process was modeled on the text factor of **cognitive load**. An object, an action and an object occurring within a short space of each other were used to form triplet relationships. These relationships represent an object performing an action on another object such as 'man drives car'. The text factor of **vocabulary** was considered to reduce complexity of the text surrounding identified concepts to help find triples. However, initial experiments using Plain English and ASD Simplified Technical English substitutions found there was little benefit when analyzing transcripts of spoken commentary. The replacements are best suited for verbose written text and there was little benefit in finding additional concepts or relationships using the substitutions in this application. It should be noted that the Plain English Campaign advises that many of their alternatives won't work in every situation. However, the substitutions have proved beneficial in previous work in other domains in relation to international standards (ISO)

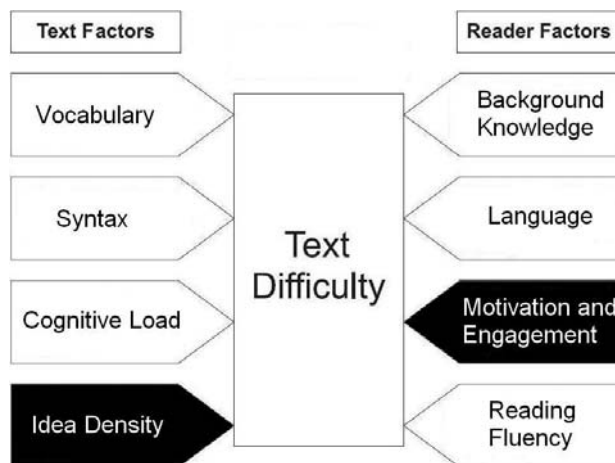


Figure 2. The factors addressed by the ontology extraction subsystem.

[24] and we believe further experimentation in other fields would demonstrate further benefit.

Additional consideration using synonyms and hyponyms of words which are present in WordNet and associated terms in the police ontology are used to merge similar triplet relationships. These processes again model the reader factors of **language** and **background knowledge**. The triples can then be used to generate the CCTV ontology detailing the particular actions an object can perform and which particular objects can receive the action. We have yet to address **idea density** in this application, and since we are considering a machine-as-reader, are not assuming a lack of **motivation and engagement**.

The methods for extracting semantic content from text, incorporating consideration from [9], are detailed as follows. Components are numbered and described in detail later with some addressing more than one factor.

- Text Factors
 - Vocabulary
 - (A.) Linguistic Concept Identification
 - (B.) Statistical Concept Identification
 - Syntax
 - (A.) Linguistic Concept Identification
 - (B.) Statistical Concept Identification
 - Cognitive Load
 - (E.) Triplet Finder
 - *Idea Density*
- Reader Factors
 - Background Knowledge
 - (C.) PITO Terminology Analysis
 - (F.) Triplet Merger
 - Language
 - (D.) WordNet Verification
 - (B.) Statistical Concept Identification
 - (F.) Triplet Merger
 - Reading Fluency
 - (B.) Statistical Concept Identification
 - *Motivation and Engagement*

These techniques for text analysis have been devised and integrated into components for use with the commonly available NLP development framework, GATE (**G**eneral **A**rchitecture for **T**ext **E**ngineering) [50]. These components build on existing GATE plug-ins from ANNIE, for preliminary NLP tasks of such as PoS tagging and sentence splitting. OWL is used for representing the ontology which captures the weight of the co-occurrence between objects and actions. The weight is calculated using frequency of triples to determine the strongest relationships between objects and actions. This information is used to associate video objects with concepts from the ontology for potential use in video annotation and keyword related search expansion. The pipeline for these resources is shown in Fig 3, followed by brief descriptions of each component in relation to the factors contributing to text difficulty.

A. Linguistic Concept Identification

Compound nouns are often labeled as technical terms by readers. Unfamiliar and technical words were identified by [9] as the text factor of **vocabulary**. The recognition of long compound nouns also addresses the text factor of **syntax**. By using linguistic techniques to identify compound nouns, we can detect the terminology used by experts to describe the content of video scenes.

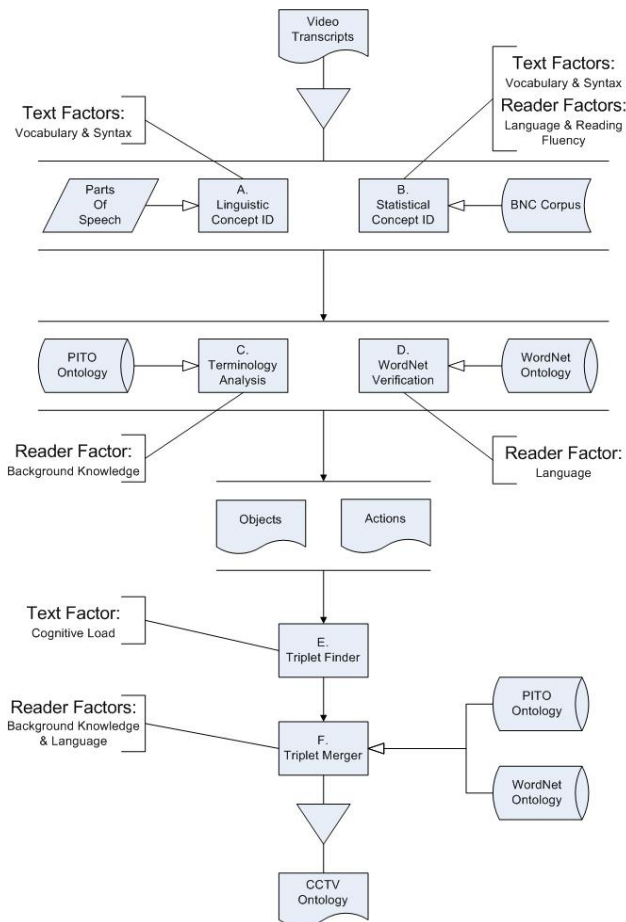


Figure 3. Pipeline for the automatic generation of the term-based CCTV ontology

This terminology along with single noun words can be used to categorize the objects in the video scenes. In addition, any words annotated as verbs in the text can be considered to be performed by an object and therefore categorized as actions. These annotated object and actions are used to populate the CCTV ontology.

We use the ANNIE tokeniser and PoS tagger to identify all the nouns and verbs in the transcripts, and to provide a basis for the identification of compound nouns according to specified patterns of part of speech annotations, extended from [51] with the formula

$$N \rightarrow \{N, A\} N. \tag{1}$$

For the initial assessment of potential object and action concepts some basic stemming does occur to detect plurals. The stemming consists of a set of rules for handling nouns and verbs.

B. Statistical Concept Identification

The initial linguistic identification of concepts is supplemented by a statistical approach both to validate the linguistic extraction and to identify other potential concepts that may have been incorrectly classified by the part of speech tagger. This involves examining expressions or collocations which occur frequently in the document, but less commonly or not at all in the English language. These words or phrases can be used by experts when describing a video scene and therefore can be considered as entries in the CCTV ontology. This process relates to the text factors of **vocabulary** and **syntax**.

To provide statistical evidence for concepts and to act as confirmation for the linguistic extraction, the transcripts are analyzed for salient keywords using frequency and weirdness information as outlined by [52], and with the British National Corpus (BNC) used as a reference corpus for this analysis. Words that are used disproportionately frequently in the expert descriptions are identified. The use of the BNC as a means to represent knowledge of linguistic expressions provides a machine-oriented interpretation of the reader factors of **language** and **reading fluency**. Neighbors of these words, within a given size of neighborhood, are also examined to identify collocation patterns that satisfy a given threshold. A value of weirdness for a word w is given by the following equation.

$$\tau(w) = \frac{N_{GL} f_{SL}(w)}{(1 + f_{GL}(w)) N_{SL}} \tag{2}$$

Where:

- SL is the collection of specialist language – here, the expert descriptions
- GL is the collection of general language – here, BNC
- N is total number of tokens - in the respective corpus
- $f(w)$ is the frequency of the word – in the respective corpus

We adopt the approach described by [22] for detecting collocations by analyzing a neighborhood of five words surrounding the keyword identified previously. The frequency of occurrences of each word at each position around the nucleate is recorded. If the nucleate and another token consistently appear together in the same positions with respect to each other, there will be a high frequency at the position of the collocating token. The variance across positions is then calculated, and [22] suggests that a variance greater than 10 is significant though we have found a variance greater than 5 more suited to the scale of analysis being undertaken here. Results are classified as either objects or actions via WordNet Verification (D.).

C. Terminology Lookup

Terminology lookup is used for the reader factor of **background knowledge** by using information and procedures defined by the National Policing Improvement Agency (NPIA), formerly known as the Police Information Technology Organisation (PITO). PITO/NPIA specified a terminology for describing events or incidents, providing data definitions which help subdivide information into simpler data elements. The PITO/NPIA elements were transformed into an ontology by re-interpreting the category and sub-category information as classes and sub-classes.

Each relationship in PITO was compared with WordNet's hyponyms. Each class and subclass was located in WordNet and then each possible path to the associated word was explored. If the semantic distance between the class and subclass was significantly large, the relationship was removed. This process removed relationships such as 'padlock' and 'removed' ('padlock' and 'door' being subclasses of 'removed' since this relates to descriptions of burglaries) and age range categories presented as subclasses for 'male' and 'female' (due to age range questioning of suspects).

The revised PITO/NPIA ontology is used for initial terminology lookup to help classify objects and actions for the CCTV ontology. Basic stemming is used to detect plurals in the PITO/NPIA ontology. In addition, the ontology is used to identify superclasses, subclasses and siblings of detected terms. These associated classes are used for merging triples at a later stage (F.). For example, 'road' including subtypes 'street' and 'motorway' are present in this ontology so the triplet 'man crosses street' can be assumed to have some degree of equivalence to 'man crosses road.'

D. WordNet Verification

Wordnet is used to implement the reader factor of **language**, as a proxy for general lexical knowledge. We use the Java JWordNet API to assess results of PoS tagging and concept identification through both linguistic (A.) and statistical methods (B.). Each single word tagged as a noun is checked in the WordNet noun dictionary. Nouns not found are considered to be erroneous and are

rejected. Basic stemming is used to prevent rejection of plurals. WordNet is also used to verify compound nouns: if the compound noun is not found in WordNet, left-to-right removal is applied - the first word is removed and the compound noun is then re-tested. For example, 'first storey shopping centre' is not known to WordNet, so 'storey shopping centre' is subsequently tested. Eventually, 'shopping centre' is tested and found in WordNet. This ensures that the longest multiword expressions that represent semantic units are used for object identification. Each single word tagged as a verb is also checked in the WordNet verb dictionary and labeled as an action if present. These verified objects and actions are used to construct the CCTV ontology accordingly.

E. Triplet Finder

With objects and actions identified, we analyze connections between the objects and actions to produce triples. The triples should represent 'agent-action-recipient' events which describe the relationship between two objects and the directed action that connects them. We calculate collocation distances between words connected in events to determine dominant patterns. The process of finding events through a high quantity of concepts within a relatively short space of each other is derived from the text factor of **cognitive load**.

F. Triplet Merger

Once triples have been identified, the WordNet lexical database and the PITO/NPIA terminology are used to merge similar relationships. Using these resources elaborates the reader factors of **language** and **background knowledge** respectively. WordNet is used to compare each agent, action and recipient in a triplet with all the other triples. If an item matches within a user-selected semantic distance, then the frequency of the triples is enhanced by the frequency from an associated triplet, weighted according to the semantic distance of the matching item - similar to the use of PageRank. For example, the triplet 'car-go-road' will make reference to the triplet 'vehicle-go-road' as car and vehicle have a small semantic distance in WordNet. The semantic distance between car and vehicle is 9, so the frequency of the triplet will be divided by this number and added into the frequency of the initial triplet to improve its rank. The words 'car' and 'motorcar' are synonymous so the semantic distance is 1. In this instance the frequencies can simply be added together.

PITO/NPIA information is used similarly to Wordnet, though preferentially given the specificity. Frequency information is calculated using semantic distance information from the ontology with super and sub classes assigned a semantic distance of 2 and sibling classes given 3 to represent the additional traversal through the ontology. Once all the triples have been identified, the agent-action-recipient information can be written to the CCTV ontology.

IV. RESULTS

To demonstrate the approach, we consider 6 expert commentaries on 12 video clips, each of somewhere between 60 and 180 seconds, showing traffic and pedestrians moving around a single lane carriageway. An image from one of the video clips can be seen in Fig 4 and two examples from the commentaries in reference to the scene are shown in Fig 5.



Figure 4. Scene from video

“On the left hand side is a parked white van with its rear doors open with a small dark saloon parked just to the rear of it with a gap of several feet between. There is a car parked on the right hand side of the near side and pedestrians walking towards the camera.”

“The white van on the far left of the screen, the door is now open with the lift platform by itself, or the van sorry left by itself with the doors open. Pedestrians walking up and down, cars going down, there does not seem to be too much happening.”

Figure 5. Two examples from six transcripts describing the same scene as featured in Fig 4

A. Analysis

The transcripts demonstrate limited agreement between the experts: the same object was described in different ways: a ‘white van’, ‘kind of van’, ‘stationary vehicle’ and ‘ambulance’ were all used to describe the same object in the scene. The experts had agreed that there was something to annotate, but differed in their selections. Despite such inconsistencies in the descriptions, there were some frequent objects and actions identified through application of linguistic (A.) and statistical (B.) processing. The most frequent objects and actions in the transcripts are detailed in Table 3. The verbs ‘be’, ‘is’, and ‘are’ are joined as they are considered the same within WordNet.

PITO/NPIA lookup was used to correct inconsistencies in the PoS tagging, resulting in higher counts. Table 4 shows the five most frequent objects found in the document. It should be noted that the erroneous object of ‘side’ is not augmented using PITO/NPIA.

TABLE III.

TOP 10 IDENTIFIED OBJECTS AND ACTIONS FROM THE TRANSCRIPTS

Rank	Object	Count	Action	Count
1	van	410	be/is/are	283
2	road	396	park	159
3	car	350	pull	140
4	side	310	come	139
5	pedestrian	188	walk	126
6	vehicle	149	get	125
7	people	96	cross	105
8	street	61	have	100
9	camera	51	go	99
10	person	48	see	88
Totals		3456		2138

TABLE IV.

TOP 5 IDENTIFIED OBJECTS CORRECTED USING PITO/NPIA TERMS

Rank	Term	Count		Variation
		Original	Revised	
1	van	410	426	16
2	road	396	402	6
3	car	350	363	13
4	side	310	310	0
5	pedestrian	188	239	51
Totals		3519	3605	86

The relationships between concepts determined through the triplet finder (E.) are detailed in Table 5, showing the 583 triples found before any merging consideration has occurred. The ‘Key’ refers to the ranks of object and action seen in Table 3; the inter-annotator agreement column shows the percentage of expert transcripts featuring the relationship, though it doesn’t imply that annotators would necessarily disagree about using these descriptions. The name of the triplet is derived from the agent, action and recipient used to construct the relationship in the original sentence. For example, ‘car-come-road’ is derived from phrases such as ‘there is a car coming down the road’.

The triples are related to the video via a timestamp associated with the comments in the transcripts. The timestamp refers to the offset from the start of the video being described. Fig 6 shows the initial comment made by one expert, from which two triples are derived (shown, with their keys, in Table 5). The objects and actions in the triples are then available for relating to the extracted visual semantics. In Fig. 6, the ideal relationship between transcript and video is demonstrated.

TABLE V.
TOP 10 IDENTIFIED TRIPLES FROM THE TRANSCRIPTS

Key	Triplet	Frequency	Inter- Annotator Agreement %
3-4-2	Car-Come-Road	20	67
10-7-2	Person-Cross-Road	16	67
5-7-2	Pedestrian-Cross-Road	13	50
7-7-2	People-Cross-Road	10	50
3-9-2	Car-Go-Road	9	33
7-8-39	People-Have-Discussion	8	17
5-5-4	Pedestrian-Walk-Side	6	50
3-4-4	Car-Come-Side	5	33
7-5-8	People-Walk-Street	4	50
19-7-8	Somebody-Cross-Street	4	17

Similar collocations are next used to strengthen frequencies via semantic distance. The triples ‘Person-Cross-Road’ and ‘Somebody-Cross-Street’ describe the same scene with the objects ‘person’ and ‘somebody’ being synonymous. This means that the frequencies of the triples can be added together for ranking purposes. At greater semantic distances, further variations can be merged but resulting in lowered frequency contributions. For example, ‘pedestrian’ has a semantic distance of 2 from ‘person’ within WordNet.

The path between the two objects is described in Table 6 with Table 8 showing the most frequent of the 328 triples after merging using WordNet synonymy and semantic relationships. The original and revised frequencies are shown in the table. A breakdown of the frequency calculation showing how the semantic distance is used to limit the frequency information is shown in Table 7. The table shows the semantic distance for each agent action and recipient. The total incorporates frequency from Table 5.



3-4-2	Car-Come-Road
10-7-2	Person-Cross-Road

00.30	See a few pedestrians on the pedestrian bit. I see a big white van and
00.40	two people just chatting about something at the back of the van. The road is quiet.
00.50	I see some cars come down the road. I can see a cyclist, he looks like a cyclist.
01.00	There's a person crossing the road.
01.10	The road is quiet, and the pedestrian bit.
01.20	I can see some people walking.

Figure 6. Two examples of triples taken from an expert which are then associated with a video scene

TABLE VI.
SEMANTIC DISTANCE BETWEEN ‘PERSON’ AND ‘PEDESTRIAN’ WITHIN WORDNET

Semantic Distance from “person”	Synonyms
0	Individual, Someone, Somebody, Mortal, Human, Soul
1	Traveler, Traveller
2	Pedestrian, Walker, Footer

TABLE VII.
FREQUENCY CALCULATION FOR ‘CAR-COME-ROAD’ MERGED TRIPLET VIA WORDNET

Variant	Semantic Distance			Frequency	
	Agent	Action	Rpt	Orig	Amend
<u>Car-Come-Road</u>	<u>0</u>	<u>0</u>	<u>0</u>	<u>20</u>	<u>20.0</u>
Car-Come-Street	0	0	2	1	0.3
Car-Go-Road	0	1	0	9	4.5
Car-Move-Road	0	1	0	2	1.0
Car-Travel-Street	0	1	2	2	0.3
Taxi-Come-Street	3	0	2	1	0.1
Vehicle-Move-Road	2	1	0	2	0.3
Van-Come-Road	3	0	0	1	0.3
Van-Go-Road	3	1	0	3	0.4
Total					27.2

TABLE VIII.
TOP 10 MERGED TRIPLES USING WORDNET AND ACCOUNTING FOR SEMANTIC DISTANCE IN FREQUENCY

Key	Triplet	Frequency		Inter- Annotator Agreement %
		Orig	Rev	
3-4-2	Car-Come-Road	20	27.2	100
3-9-2	Car-Go-Road	9	25.4	100
3-14-2	Car-Move-Road	2	25.4	100
11-7-2	Person-Cross-Road	16	24.9	100
19-7-2	Somebody-Cross-Road	1	24.9	100
55-7-2	Someone-Cross-Road	1	24.9	100
5-7-2	Pedestrian-Cross-Road	13	20.2	100
19-7-8	Somebody-Cross-Street	4	13.7	100
55-7-8	Someone-Cross-Street	1	13.7	100
3-29-8	Car-Travel-Street	2	11.7	100

As an alternative to WordNet, the PITO/NPIA ontology is used to merge triples using more specifically specified relationships for this domain. In PITO/NPIA, the objects ‘car’ and ‘van’ have a semantic distance of 2 (the parent has a distance of 1). In Wordnet, the objects have a semantic distance of 3 with ‘minivan’ and ‘passenger van’ present on the intervening levels. The PITO/NPIA information can be used to merge triples using specialized information. The most frequent of the resulting 142 triples merged using the PITO/NPIA ontology are shown in Table 9 showing the original and revised frequencies.

TABLE IX.
TOP 10 MERGED TRIPLES USING THE PITO/NPIA ONTOLOGY

Key	Triplet	Frequency		Inter- Annotator Agreement %
		Orig	Rev	
3-4-2	Car-Come-Road	20	20.7	67
5-7-2	Pedestrian-Cross-Road	13	14.0	67
7-7-2	People-Cross-Road	10	11.0	67
3-9-2	Car-Go-Road	9	10.3	50
5-7-8	Pedestrian-Cross-Street	2	8.5	67
1-4-2	Van-Come-Road	1	8.0	67
7-7-8	People-Cross-Street	2	7.0	67
1-9-2	Van-Go-Road	3	6.3	50
3-4-4	Car-Come-Side	5	6.0	50
42-9-2	Bus-Go-Road	1	5.0	50

The PITO/NPIA ontology alone is insufficient to significantly improve the degree of agreement, suggesting that the language used is more akin to general language. There is, however, some impact evident. The PITO/NPIA ontology and WordNet were used together to assess this impact. Triples were merged by preferring the expert knowledge of the PITO/NPIA ontology and subsequently using general knowledge represented by WordNet. There were 342 merged triples from the combined process with the 10 most frequent shown in Table 10, with raised values for frequency.

With unmerged triples, erroneous relationships exist including ‘Pedestrian-Cross-Van’, derived from the sentence, ‘Pedestrian crossing between the van and the other car’. These relationships are filtered out as frequency of events increases through merging of triples and analysis of further transcripts.

TABLE X.
TOP 10 TOP 10 MERGED TRIPLES USING BOTH THE PITO/NPIA ONTOLOGY AND WORDNET

Key	Triplet	Frequency		Inter- Annotator Agreement %
		Orig	Rev	
3-4-2	Car-Come-Road	20	28	100
3-9-2	Car-Go-Road	9	26.8	100
3-14-2	Car-Move-Road	2	26.8	100
11-7-2	Person-Cross-Road	16	25.9	100
19-7-2	Somebody-Cross- Road	1	25.9	100
55-7-2	Someone-Cross- Road	1	25.9	100
5-7-2	Pedestrian-Cross-Road	13	20.8	100
19-7-8	Somebody-Cross-Street	4	17.5	100
55-7-8	Someone-Cross-Street	1	17.5	100
3-29-8	Car-Travel-Street	2	16.1	100

B. Evaluation

Although CCTV produces significant quantities of data, there are few existing benchmark collections, and none that appear to be provided with associated text though some expect the use of automatic speech recognition. Three key video corpora considered were TRECVID, CAVIAR and VideoCLEF (see Table 11). TRECVID’s video surveillance task provides for about 100 hours of uncommented video data, with a number of kinds of events to be detected, and a further “freestyle” task involving defining the important events in the domain [55]. In a sense, the transcripts analyzed in this paper are part of the definition of the important events in this domain. CAVIAR data may be useful for training detectors and associating through co-occurrence models to high-level descriptions of video such as “Two persons cross paths at the entrance of a store”. These descriptions already have a restricted syntax, with minor variations, but are of limited value in discerning what might be of importance in the video for constructing more specific descriptions. While not directly related, it is claimed that future iterations of VideoCLEF may involve the use of subtitle text, offering potential for evaluating the readability issues as might emerge within. However, this would be a single source of text in contrast to the multiple similar descriptions analyzed here, and there is an inherent brevity required in subtitling. We have considered, further, use of text corpora as used for polarity assessment in, for example, sentiment analysis of movie reviews [56] and political debate [57]. In contrast to considerations with such sentiment corpora, it is the variability not consistency that is of interest in such collections.

In the absence of a benchmark collection that involves multiple descriptions of the same video fragments, we demonstrate the impact of “readability” in two ways: (i) in relation to video retrieval, we use typical information retrieval measures of precision and recall (sub-section 1);

TABLE XI.
TOP 10 TOP 10 MERGED TRIPLES USING BOTH THE PITO/NPIA ONTOLOGY AND WORDNET

	TRECVID	CAVIAR	VideoCLEF
Data	100-hours of video surveillance from Gatwick Airport provided by UK Home Office.	Footage from lobby at INRIA and shopping centre in Lisbon	Documentaries and talk shows, involving Dutch and English spoken language.
Analysis	Detection and tracking: Events including PersonRuns, PeopleMeet, PeopleSplitUp.	Detection and tracking.	Classification, classes including: Dance, Music and Scientific Research
Associated text	None	None	Subtitle text may subsequently be made available (not in 2008)
See also	Reference [1]	Reference [53]	Reference [54]

(ii) in relation to inter-annotator agreement, we use cosine distance to demonstrate increased similarity in descriptions (sub-section 2). Additionally, we consider that annotators will increase their outputs depending on a cognitive load that stems from the video itself: a significant amount of information occurs in the video needs to be described rapidly, and this is likely to lead to greater frequency density of objects, events, and emergent triples (sub-section 3).

1) *Recall and Precision*

To find if the triples correctly associated with events in the CCTV footage, the video was manually analyzed as it might be for annotation purposes. Each 10 second segment containing a person crossing the road was identified. There were 191 such segments, with 92 of those having been commented on by the experts as containing a person crossing the road. The 16 ‘person-cross-road’ triples were used to locate the timestamps in the original transcripts. Precision and recall was used to find the instances of video segments returned for the triplet. A semantic distance of 5 was used to merge the ‘person-cross-road’ triplet to similar occurrences. The results of precision and recall for each triplet are shown in Table 12.

The sentence used to derive the false positive for ‘person-cross-road’ contained the phrase, ‘the person who has just crossed the road’ where the expert was referring to the same person who had performed the action previously. The false positive for ‘somebody-cross-street’ just missed a video segment containing a person crossing and we assume the expert was slightly late in their commentary. The false positive for ‘driver-go-road’ is clearly wrong.

Although the number of correct results was high, with few undesired results, there was a significant amount of video segments not returned using the triples. There were sections of video including people crossing roads which had not been indicated by the experts. Even so, the results show good effects for query expansion, and we now have segments of video that may act as a useful test bed for automatic annotation and an annotated dataset for testing automatic annotations systems.

2) *Description similarity*

Further analysis on the inter-annotator agreement was performed on the sample of the transcripts detailed in Fig 5. The top 500 most frequent words from BNC were removed from the samples to produce bags of words, which were used to calculate the cosine distance of the two expert samples. The bags of words were then revised so that ‘car’, ‘van’ and ‘saloon’ are consolidated to determine whether there is an increase in the similarity between these. Table 13 shows the vector size and cosine distance of each of the raw and revised bags of words with the value for identical vectors being 1. The results demonstrate that accounting for these variations, produces greater similarity amongst the descriptions.

TABLE XII.
THE PRECISION AND RECALL OF VIDEO RETRIEVAL USING THE ‘PERSON-CROSS-ROAD’ TRIPLET AND ITS ASSOCIATES SEPARATELY, AND COMBINED VALUES (TOTAL)

Triplet	True Pos	True Neg	False Pos	False Neg	P	R
Person-Cross-Road	15	98	1	77	0.94	0.16
Pedestrian-Cross-Road	13	99	0	79	1	0.14
Pedestrian-Cross-Street	2	99	0	90	1	0.02
Somebody-Cross-Street	3	98	1	89	0.75	0.03
Somebody-Cross-Road	1	99	0	91	1	0.01
Someone-Cross-Street	1	99	0	91	1	0.01
Someone-Cross-Road	1	99	0	91	1	0.01
Lady-Cross-Road	1	99	0	91	1	0.01
Child-Cross-Road	1	99	0	91	1	0.01
Teenager-Cross-Street	1	99	0	91	1	0.01
Driver-Go-Road	0	98	1	92	0	0
Total	39	96	3	53	0.93	0.42

TABLE XIII.
RAW AND REVISED COSINE DISTANCE OF THE TWO SAMPLE DESCRIPTIONS USING ‘VEHICLE’ WORD CONSOLIDATION

Raw		Revised	
Vector Size	Cosine Distance	Vector Size	Cosine Distance
23	0.43	21	0.56

3) *Video-borne cognitive load*

We attempt to address cognitive load as the number and similarity in resulting triples between four annotators describing the same events. The numbers of objects and actions described by all annotators for each timestamp was analyzed to determine whether cognitive load would be indicative of increased numbers of descriptions in relation to clusters of video events. Table 14 shows the number of objects and actions described by the annotators, along with the total number of words for each timestamp. The percentage of objects and actions (terms) for the total words was calculated with the results compared to the number of triples over the course on one of the videos.

All timestamps with a triplet count of 3 or more occur when the proportion of terms is greater than 33.3%. With triplet counts of less than 3, the results become less consistent covering a wider range of percentages. At timestamp 2:10 there is an increase of activity due to a congestion of pedestrians creating a large number of descriptions. It would appear, then, that, on the one hand, this could be a measure of text difficulty, but on the other could be an indication of important events increasing the cognitive load of the annotators.

TABLE XIV.
COGNITIVE LOAD FOR EACH TIMESTAMP CORRELATED WITH
IDENTIFIED TRIPLES

Time stamp	Objects	Actions	Words	Terms%	Triples
00:00	15	9	80	30	0
00:10	15	7	66	33.3	2
00:20	14	8	58	37.9	3
00:30	16	11	68	39.7	1
00:40	11	8	69	27.5	1
00:50	11	10	74	28.4	1
01:00	10	7	66	25.8	0
01:10	14	10	60	40	4
01:20	12	4	36	44.4	2
01:30	14	10	61	39.3	1
01:40	13	10	58	39.7	3
01:50	13	3	48	33.3	2
02:00	14	3	43	39.5	3
02:10	13	11	58	41.4	7
02:20	6	5	34	32.4	1

Further work needs to be done to investigate, amongst other things, use of active and passive voice (syntax). Currently distinct triples will be produced for the active and passive voice for 'the driver parks the car' and 'the car is parked by the driver'. Further analysis is required involving a more granular consideration of inter-annotator variability and how query similarity may be accounted for using this system.

V. CONCLUSION

In this paper we have demonstrated how considerations of readability relate to the formulation of suitable annotations for video retrieval. There are substantial opportunities for the extended evaluation of the techniques described, and this is the subject of ongoing efforts. In Section 2, we related a range of research to the framework for readability described by [9], covering eight different elements that are generally characterized as text factors and reader factors. These may be broadly characterized as making any kind of text appropriate for potential readership by simultaneously considering the text and the reader. Here, we may need to associate prior literature familiar to the reader to the text at hand. If the reader has seen a somewhat similar text, the text at hand may be deemed more easily readable. We consider an analogous situation for automatic video annotation: if a system has a database of previous videos and well-constructed annotations, it might be more easily possible to 'read' similar videos. In Section 3 we focused on using a selection of the eight elements of readability in describing a system for automatically capturing semantic content of text. These eight elements of readability may be applied with varying degrees of success to other collections of text, as demonstrated in prior work [24]. In Section 4 we demonstrated the impact of our selection of

readability elements on parallel expert annotations of CCTV footage and showed how such annotations may be more deeply related (concepts) than is apparent on the surface (terms). Considering the structuring of these concepts and deriving the relationships amongst them leads us to an enhanced form of automatic ontology learning, with considerations derived from earlier work [58], [59]. Automatic ontology learning here considers 'ontology pruning' during the learning process: terms may attain higher importance through consideration of other contributory semantically-related terms. With due respect to a familiar approach to scaling importance of web links, we might refer to such contributions as 'termrank'. The semantic content of the expert transcripts, then, is represented in the CCTV ontology. It should be noted that the expert commentary had a high lexical density for spoken language with more than the usual two words per clause. Further work needs to be done exploring the advantages of using Plain English Campaign and ASD Specified Technical English substitutions in other applications where verbose language is notorious. It would be interesting to see additional triples which may be found using these techniques.

It is not only the text factors that demonstrate potential; reader factors can potentially play a significant role in multimedia information retrieval. As we have seen the reader factor of language can be targeted using Wordnet to reflect syntactic development by considering synonymy to find alternate expressions for the same concepts. These synonyms are used to gather higher frequencies for similar semantic relationships. In addition, the reader factor of background knowledge is used to find related or hierarchical terms for identified concepts. These approaches can be amended for information retrieval so that search keywords are expanded into hyponyms and associated terms according to the knowledge of the user. Information stored on the user's PC may identify the extent of their knowledge of particular domains and recall appropriate information accordingly.

The readability framework devised by [9] can be applied to the automatic annotation of multimedia by ensuring consistency of description for metadata. Their methods can be deployed to ensure concepts or items represented within the media are consistently identified and annotated. This approach will address some of the problems currently facing multimedia information retrieval such as the inadequacy of uniform textual descriptions for describing color, shape etc and keywords being limited to particular domains as the original keywords cannot anticipate future applications.

In addition, the ontology might be extended by assigning multimedia objects to relevant concepts. Reference [6] showed how multimedia ontologies can be used to perform automatic annotation on unknown sequences of video. Using visual descriptors and temporal cues, a representative set of sequences containing concepts described in the linguistic ontology can be used to create a multimedia ontology. This

ontology is a form of visual vocabulary for the concepts present in the video and just as [9] showed how a text vocabulary affects text difficulty; a visual vocabulary is required to interpret visual media correctly. Reference [7] believed that video retrieval systems should integrate all available media such as audio, video and captions and devised a system of using animated visual sketches to formulate search queries. In these sketches motion and temporal duration are the key attributes attached to an object in the sketch. Similarly, image retrieval systems can use a sketch based query to calculate the correlation between the sketch and the edge map of images within the search field. These sketches allow retrieval of content similar to the presented query in manner which can not be achieved through text-based query search alone. The sketches provide a vocabulary for image/video retrieval and provide a means to connect the visual media to semantics. By checking the similarity of visual descriptors of unseen video with the representative sequences present in the vocabulary, we can begin to automatically recognize visual features in the video.

We hope to bridge the semantic gap by matching the extracted terminology from text to the vocabulary of sketches used in image/video analysis. The sketch queries devised by [7] could form the visual descriptors needed to build this visual vocabulary. However, for textual labels to be applied to these features we need both visual and textual labels for the same concepts. It is the integration of visual and textual vocabularies which can lead to the multimedia ontologies. Fig 7 shows the relationships to events for scenes using a multimedia ontology defined for CCTV. Here the visual representation of a parked estate is categorized as a subclass of 'Park Car'. This is a starting point for automatically creating annotations for video and other forms of multimedia information in the semantic web. The resulting process could demonstrate aspects of the human cognitive ability to associate visual experiences with semantic information.

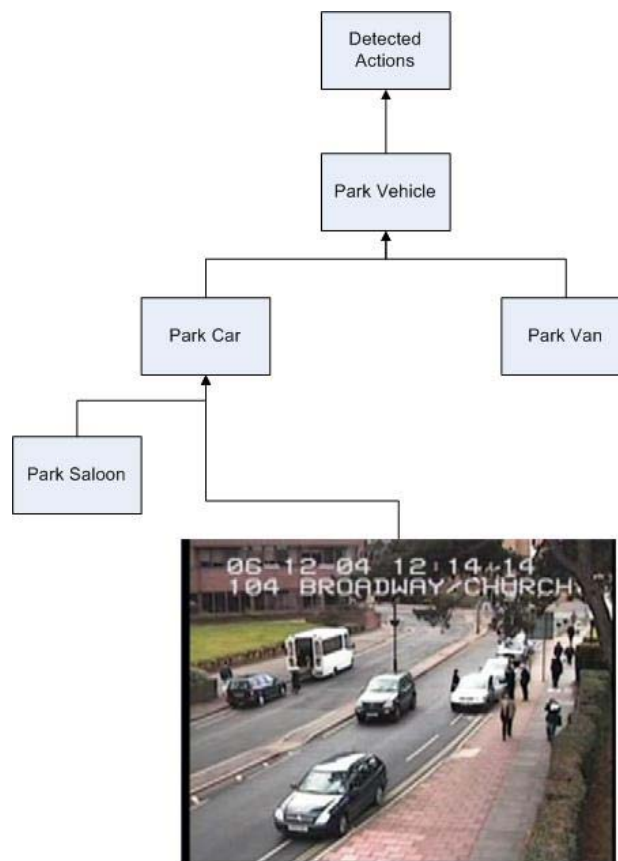


Figure 7. Schema used to annotate CCTV video clip

ACKNOWLEDGMENT

This work has been supported, in part, by EPSRC-sponsored REVEAL project (GR/S98450/01) and by the EU eContent project LIRICS (22236).

REFERENCES

- [1] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVID," in *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, Santa Barbara, California, USA, October 26 - 27, MIR '06. ACM Press, New York, NY, pp. 321-330, 2006.
- [2] R. K. Srihari, "Use of Collateral Text in Understanding Photos," *Artificial Intelligence Review*, special issue on Integrating Language and Vision, vol. 8, pp. 409-430, 1995. [* Also reprinted as book chapter in Paul McKevitt (ed), Kluwer, 1995.]
- [3] J. S. Hare, P. A. S. Sinclair, P. H. Lewis, K. Martinez, P. G. B. Enser, and C. J. Sandom, "Bridging the semantic gap in multimedia information retrieval - top-down and bottom-up approaches," in *Proceedings of the 3rd European Semantic Web Conference*, Budva, 2006.
- [4] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *Proceedings of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99)*, 1999.
- [5] A. Jaimes, and J. R. Smith, "Semi-automatic, data-driven construction of multimedia ontologies," in *Proceedings of IEEE Int'l Conference on Multimedia & Expo*, vol. 1, pp. 781-4, 6-9 July 2003.

- [6] M. Bertini, A. Del Bimbo, and C. Torniai, "Soccer Video Annotation Using Ontologies Extended with Visual Prototypes," in *Proceedings of the International Workshop on Content-Based Multimedia Indexing*, CBMI '07, vol. 25-27, pp. 212-218, 2007.
- [7] S. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo, "Large-Scale Multimodal Semantic Concept Detection for Consumer Video," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*, Augsburg, Bavaria, Germany, pp. 255-264, 2007.
- [8] B. Vrusias, D. Makris, J. P. Renno, N. Newbold, K. Ahmad, and G. Jones, "A Framework for Ontology Enriched Semantic Annotation of CCTV Video," *International Workshop on Image Analysis for Multimedia Interactive Services*, Santorini, Greece, June 2007.
- [9] T. Oakland, and H. B. Lane, "Language, Reading, and Readability Formulas: Implications for Developing and Adapting Tests," *International Journal of Testing*, vol. 4(3), pp. 239-252, 2004.
- [10] G. R. Klare, *The Measurement of Readability*. Ames, Iowa: Iowa State University Press, 1963.
- [11] H. D. Kitson, *The Mind of the Buyer*. New York: Macmillan, 1921.
- [12] R. Flesch, "A new readability yardstick," *Journal of Applied Psychology*, vol. 32, pp. 221-223, 1948.
- [13] H. McLaughlin, "SMOG grading - a new readability formula," *Journal of Reading*, vol. 22, pp. 639-646, 1969.
- [14] R. J. Senter, and E. A. Smith, "Automated readability index," *AMRL-TR*, 66-22, Wright-Patterson AFB, OH: Aerospace Medical Division, 1967.
- [15] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, & B. S. Chissom, "Derivation of new readability formulas (automated readability index, Fog count and Flesch reading ease formula) for navy enlisted personnel," *Research Branch Report 8-75*, Naval Air Station, Memphis, TN, 1975.
- [16] W. H. DuBay, *The Principles of Readability*. Costa Mesa, CA: Impact Information, 2004.
- [17] L. Gillam, and N. Newbold, "Quality Assessment," Deliverable 1.3 of EU eContent project LIRICS, 2007, URL: http://lirics.loria.fr/doc_pub/T1.3Deliverable.final.2.pdf, last accessed 29 July 2009.
- [18] S. Willaims, and E. Reiter, "Generating basic skills reports for low-skilled readers," *Journal of Natural Language Engineering*, vol. 14(4), 2008.
- [19] G. R. Klare, "Readability" in *Handbook of Reading Research*, P. D. Pearson, ed. New York: Longman, 1984, pp. 681-744.
- [20] ASD Simplified Technical English, URL: <http://www.simplifiedenglish-aecma.org/>, last accessed 29 July 2009.
- [21] J. R. Firth, *Papers in Linguistics: 1934-1951*. London: Oxford University Press, 1957.
- [22] F. Smadja, "Retrieving collocations from text: Xtract," *Computational Linguistics*, vol. 19(1) pp. 143-178, 1993.
- [23] J. Pustejovsky, S. Bergler, and P. Anick, "Lexical Semantic Techniques for Corpus Analysis," *Computational Linguistics*, vol. 19(2), pp. 331-358, 1994.
- [24] N. Newbold, and L. Gillam, "Automatic Document Quality Control," in *Proceedings of the Sixth Language Resources and Evaluation Conference (LREC)*, Marrakech, May 2008.
- [25] M. A. K. Halliday, and J. R. Martin, *Writing Science: Literacy and Discursive Power*. London: Falmer Press, 1993.
- [26] M. Lauer, "Designing Statistical Language Learners: Experiments on Noun Compounds," unpublished PhD thesis, Macquarie University, Sydney, Australia, 1995.
- [27] The Plain English Campaign, URL: <http://www.plainenglish.co.uk/>, last accessed 29 July 2009.
- [28] M. Hoey, *Patterns of Lexis in Text*. Oxford: OUP, 1991.
- [29] G. R. Klare, J. E. Mabry, and L. M. Gustafson, "The relationship of style difficulty to immediate retention and to acceptability of technical material," *Journal of Educational Psychology*, vol. 46, pp. 287-295, 1955.
- [30] E. B. Entin, and G. R. Klare, "Relationships of measures of interest, prior knowledge, and readability comprehension of expository passages," *Advances in reading/language research*, vol. 3, pp. 9-38, 1985.
- [31] G. R. Klare, "A second look at the validity of the readability formulas," *Journal of Reading Behaviour*, vol. 8, pp. 159-152, 1976.
- [32] K. E. Cox, and J. T. Guthrie, "Motivational and cognitive contributions to student' amount of reading," *Contemporary Educational Psychology*, vol. 26, pp. 116-131, 2001.
- [33] R. Brown, *A first language: The early stages*. Cambridge, MA: Harvard University Press, 1973.
- [34] S. Conant, "The relationship between age and MLU in young children: A second look at Klee and Fitzgerald's data," *Journal of Child Language*, vol. 14, pp. 169-173, 1987.
- [35] J. Miller, and R. Chapman, "The relationship between age and mean length of utterance in morphemes," *Journal of Speech and Hearing Research*, vol. 24, pp. 154-161, 1981.
- [36] T. Klee, M. Schaffer, S. May, I. Membrino, and K. Mougey, "A comparison of the age-MLU relation in normal and specifically language-impaired preschool children," *Journal of Speech and Hearing Disorders*, vol. 54, pp. 226-233, 1989.
- [37] R. Bod, J. Hay, and S. Jannedy, Eds. *Probabilistic linguistics*. Cambridge, MA: MIT Press, 2003.
- [38] J. Bybee, and P. Hooper, (Eds.) *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, 2001.
- [39] N. Ellis, "Implicit and explicit language learning: An overview," in *Implicit and explicit learning of languages*, N. C. Ellis, Ed. San Diego, CA: Academic Press, 1994, pp. 1-32.
- [40] J. H. Hulstijn, "What does the impact of frequency tell us about the language acquisition device?" *Studies in Second language Acquisition*, vol. 24(2), pp. 269-73, 2002.
- [41] M. Ford, J. Bresan, and R. M. Kaplan, "A competence-based theory of syntactic change," in *Mental representations and grammatical relations*, J. Bresan, Ed., Cambridge, MA: MIT Press, 1982, pp. 727-796.
- [42] H. Diessel, "Frequency Effects in Language Acquisition, Language Use, and Diachronic Change," *New Ideas in Psychology*, vol. 25(2), pp. 108-127, 2007.
- [43] K. E. Stanovich, "Word recognition: Changing perspectives," in *Handbook of Reading Research*, vol. 2, R. Barr, M. L. Kamill, P. Mosenthal and P. D. Pearson Eds. New York: Longman, 1991, pp. 418-452.
- [44] R. Hudson, C. Mercer, and H. Lane, "Exploring reading fluency: A paradigmatic overview," unpublished manuscript, University of Florida, Gainesville, 2000.
- [45] L. C. Ehri, and S. McCormick, "Phases of word learning: Implications for instruction with delayed and disabled readers," *Reading and Writing Quarterly: Overcoming Learning Difficulties*, vol. 14(2), pp. 135-164, 1998.

- [46] T. V. Rasinski, *The Fluent Reader: Oral Reading Strategies for Building Word Recognition, Fluency, and Comprehension*. New York: Scholastic, 2003.
- [47] P. A. Schreiber, "On the acquisition of reading fluency," *Journal of Reading Behaviour*, vol. 7, pp. 177–186, 1980.
- [48] A. Young, and P. G. Bowers, "Individual difference and text difficulty determinants of reading fluency and expressiveness," *Journal of Experimental Child Psychology*, vol. 60, pp. 428–454, 1995.
- [49] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.
- [50] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications," in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, July 2002.
- [51] C. Jacquemin, *Spotting and Discovering Terms through Natural Language Processing*. The MIT Press, 2001.
- [52] L. Gillam, "Systems of concepts and their extraction from text," unpublished PhD thesis, University of Surrey, UK, 2004.
- [53] R. B. Fisher, "PETS04 Surveillance Ground Truth Data Set," in *Proceedings of 6th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS04)*, pp. 1–5, May 2004.
- [54] M. Larson, E. Newman, G. Jones, "Overview of VideoCLEF 2008: Automatic generation of topic-based feeds for dual language audio-visual content," in *CLEF 2008 workshop notes*, F. Borri, A. Nardi, and C. Peters, Eds. 2008, URL: http://www.clef-campaign.org/2008/working_notes/Larson_overviewCLEF_VideoCLEF.pdf, last accessed 29 July 2009.
- [55] NIST, "Straw Man Proposal for the TRECvid 2008 Evaluation," in National Institute of Standards and Technology, 2008 URL: <http://www.nist.gov/speech/tests/trecvid/2008/doc/20080114-trecvid-strawman-proposal.pdf>, last accessed 29 July 2009.
- [56] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," in *Proceedings of EMNLP*, pp. 79–86, 2002.
- [57] M. Thomas, B. Pang, and L. Lee, "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts," in *Proceedings of EMNLP*, pp. 327–335, 2006.
- [58] L. Gillam, M. Tariq, and K. Ahmad, "Terminology and the construction of ontology," in *Application-Driven Terminology Engineering*, Anne Condamines and M. Teresa Cabré Castellví, Eds. Ibekwe-SanJuan, Fidelia, 2007, pp. 49–73.
- [59] L. Gillam, and K. Ahmad, "Pattern mining across domain-specific text collections," *LNAI 3587*, pp. 570–579, 2005.