OPTIMISING CONTENT CLARITY FOR HUMAN-MACHINE SYSTEMS[1]

George R S Weir & Nikolaos K. Anagnostou
Department of computer and Information Sciences
University of Strathclyde, Glasgow G1 1XH
UK

Email: george.weir@cis.strath.ac.uk, nanagnos@cis.strath.ac.uk

1. Introduction

Increasingly, members of the public are exposed to human-machine systems that are many times more complex in functionality and interactivity than previous systems (Weir, 1991). Often, these systems are domestic appliances, such as digital video cameras or recordable DVD players, and the end-users take the role of operators, seeking to assimilate, command and control their newly acquired technologies. Although, unlike commercial process control environments, systems at the domestic level pose little risk to life and limb, there are similarities in the challenges facing both types of operator. In particular, each will be reliant to some degree upon the readability and comprehensiveness of associated documentation. Whether such support is available on-line, in the form of interactive help systems, or off-line, in the form of associated technical manuals, each operator relies upon the clarity and comprehensibility of its content. Inevitably, there are settings in which the adequacy of afforded information proves critical. Dubay (2004) states that, due to poor comprehension of installation manuals, 79 to 94 percent of child safety seats are installed improperly - a fact that has aggravated the rate of infant deaths in traffic accidents. The inadequacy of technical documentation addressing the complexity of child safety seat installation is a stark pointer to the scale of difficulty inherent in documentation design.

In the present paper, we detail issues associated with the production of clearly expressed and comprehensible technical documentation and describe an approach to optimising the clarity of such content. Our aim in this work is to develop support for authors in checking the likely comprehensibility of chosen forms of expression by reference to an external measure of 'likely familiarity'. Our DOcumentation Support Tool (DoST) will assist in identifying words and expression forms that are likely to be unfamiliar to end users. This approach to documentation support and its underlying principles are detailed below.

2. Content clarity

Technical domains introduce specific difficulties for documentation in virtue of their constrained context (usually, representing the system in question). Linguists refer to such contexts as a 'sublanguage', wherein we find a high frequency of specialised terminology and, often, a restricted set of grammatical patterns (Kittredge & Lehrberger, 1982). For highly technical domains, especially in commercial settings, training is often provided as orientation and as a basis for acquaintance with concepts and operations that are locally pertinent but otherwise exclusive to the system being addressed. In settings with lower associated commercial and health risk, operators are often required to learn 'on the job' and may be left to sink or swim in accord with the effectiveness or otherwise of accompanying documentation. In domestic contexts, owners of highly complex interactive systems are often required to comprehend the technical domain solely on the basis of trial and error, technical manuals and user guides. Achieving content clarity in such settings is a difficult objective that is often not secured.

Difficulties may arise in two contexts. Firstly, when creating original content for technical documents, authors must choose forms of expression that are both adequate to capture the intended meaning and at the same time likely to be familiar to the intended audience. Secondly, many recognised difficulties in documentation arise from translation effects, wherein original non-English documents are 'converted' for an English speaking audience. This process, often

notoriously, introduces language artefacts that compromise the coherence of the resultant documentation (Weir & Lepouras, 2001).

When considering the problem of adequate documentation, a natural assumption is that authors should check on the coherence and comprehensibility of their materials. User testing is often advocated as a consideration for system design (e.g., Rubin & Hudson, 1994; Lindgaard, 1994) and is no less appropriate for system documentation. Yet, for reasons of economy and, perhaps, a pervasive presumption of transparency in authored materials, such measures are often neglected.

3. Approaches to documentation support
In keeping with recommended practice for interactive system design (e.g., Mandel, 1997), support documentation should undergo an iterative development process in which review and revision form a repeated cycle. This permits many potential shortcomings to be detected and corrected prior to final release. Yet such steps take place off-line (i.e., on the completed document), and are only viable once draft documentation is available. While desirable, this belated intervention may usefully be supplemented by support that operates at the documentation authoring stage (i.e., on-line, in the context of document creation). In principle, such earlier intervention may reduce the need for later correction and associated delay.

Automated means of assisting documentation authors include familiar word processor facilities, such as spelling and grammar checking. These aids afford a desirable degree of language quality but are no guarantee that the resultant documents are fit for purpose (i.e., provide adequate user support). In this context, we aim to develop support tools that supplement existing on-line facilities. Our Document Support Tool (DoST) is currently in development and the following describes the techniques that we are employing toward optimising documentation, with a view to enhanced concept clarity and the delivery of better content to the end user of the documentation.

3.1 Readability measures
An appealing approach is to apply readability analysis in order to gauge the quality and likely comprehensibility of any documentation as it is produced. This is made feasible through ready access to readability formula that can be applied directly to electronic texts. Readability formulas use quantifiable textual characteristics as a basis for extrapolating the readability of a text. Such characteristics are usually described as 'semantic' if they concern the words used and 'syntactic' if they have to do with the length or structure of sentences. The two factors most commonly used in readability formulas are vocabulary difficulty, measured by either word difficulty or word length, and average sentence length, since a number of studies have proven them to be strongly associated with comprehension (Dave and Chall, 1995, p. 81). Of course, beyond such 'surface features' of texts, there are other variables that affect readability, including content and the reader's abilities, but since these cannot readily be measured they are not represented in readability formulas.

Often, readability formulas return an estimate of a text's difficulty in terms of grade levels, i.e., the years of schooling needed to be able to comprehend the text. The grade-level scale was adopted because it provided a way to "compare reader's ability levels to the difficulty levels of written material" (Klare, 1984, p. 718) and thus guide teachers in the selection of appropriate material for their students. The outputs from such readability formulas tend to be more accurate in lower grade levels, because as the level increases, content, instead of writing style, becomes the deciding factor for readability (Klare, 1984, p. 730).

While readability measures afford a convenient means of gauging texts, they are often discounted or derided. According to Connaster (1999), 'readability formulas fail to predict text difficulty' and he cites an experiment to demonstrate that "text difficulty" is 'a perception of the reader and therefore cannot be objectively calculated by counting syllables, word length, sentence length, and other text characteristics'. While we have not entirely rejected the use of readability measures in developing our support facility, we feel that existing approaches need to be enhanced to reflect further semantic components in the analysed texts. Specifically, we propose to add frequency-

based measures, such as those described below, as additional components within a readability measure. Further detail on the nature of readability measures and their shortcomings is given in Anagnostou & Weir, 2007.

3.2 Other linguistic indicators

The theoretical basis for our documentation support lies with an estimation of 'textual commonality' that is derived from a frequency list of words and word collocations. As a step toward gauging the likelihood that a general reader will be familiar with the specific words and phrases contained in documentation, the words and collocations (word combinations) can be rated against a selected corpus. This approach is based upon our research on text matching, readability measures and collocational frequency (Anagnostou & Weir, 2007; Campbell & Weir, 2007; Weir & Ritchie, 2007). which allows us to apply corpus-derived frequency measures as a comparator for words and expressions.

3.2.1 Word frequency analysis

Any approach to estimating the likelihood that a reader will be acquainted with a specific word must make assumptions about the readership. Broadly speaking, an audience will either be specific or general. For instance, documents with highly specialised content, such as technical reports or medical articles, are not intended for a general audience. Rather, these publications will presume that readers have an existing degree of subject familiarity. Understanding the document content will depend upon this familiarity. Thereby, a specific audience is anticipated. In contrast, user documentation for devices sold to the general public aims to support a general audience. Consequently, the expected prior familiarity must be low or negligible.

The application of word frequency analysis in specific and general contexts will be similar in approach. In each case, we need a reference corpus of texts. This corpus acts as an indicator of typical content for the domain under consideration. The difference between general and specific domains lies in the coverage of the reference corpus. If the domain is specific, such as process control, the reference corpus must adequately represent the content for this domain and will comprise a large sample of texts considered representative of this knowledge area. If the domain is general, the reference corpus must reflect this generality in its coverage. In either case, the reference corpus must be 'representative', i.e., contain samples of all major text types (Leech, 1993) and these should be proportional to their actual usage (Rayson & Garside, 2000, p.2).

Measuring characteristics of any sample document against a reference corpus is essentially a corpus comparison, since the sample document is a corpus of text in its own right. In order to gauge the prominence of words in our sample text, we can first apply a simple frequency count to determine which words in that document are most common (relative to that document). This measure may reflect absolute number of occurrences, or a ratio of number of occurrences to the total number of words in the whole document. In either case, we derive a ranking of words. To glean broader insight on the word content of our sample, we contrast its word frequency results with a similar measure drawn from our reference corpus. This allows us to make a statistical determination of the likelihood that words have random frequency of occurrence in our sample. Those words whose frequency of occurrence in the sample document is shown to be non-random (statistically significant) may be considered worthy of attention, since they play a more statistically significant role in the considered text. We determine the statistical significance by measuring the log-likelihood (LL) of each word's frequency in the sample text against its frequency of occurrence in our reference corpus (Rayson & Garside, 2000, p.2; McEnery & Wilson, 2001, p84).

This approach to analysis of word frequency in existing texts is available in a variety of current software tools, such as Wordsmith Tools (Scott, 2004) and AntConc (Anthony, 2004). Through a comparison of word frequency in the sample document against a reference corpus, and application of the log-likelihood function (or similar), a ranked list of 'keywords' is derived. Plausibly, these are the words that most accurately describe what the text is about, since their appearance in that text has greater statistical significance than other lexical ingredients. Since our objective is on-line

support for documentation, we have prototyped our own software for frequency measurement and keyword ranking. This is presently only operational off-line but permits applicability testing toward our intended on-line context.

We have employed the British National Corpus (BNC) as our reference corpus (Leech, 1993). This resource serves as a general corpus and affords a comparator for documentation that is intended for a general audience. Figure 1 illustrates the result of analysing the frequency significance for an example user manual (only the top ten ranked words are listed). The LL values indicate that the significance of occurrence frequency of these words is greater than mere chance. In turn, this allows us to deduce that for this manual, these words play the major role in delivering content to the reader. Note that this is not simply a measure of word frequency in the sample manual. This is evident from the difference between 'ActiveSync' and 'host'. 'Host' occurs more often in the sample manual than 'ActiveSync', but the significance of the latter is greater than the former.

Rank	Frequency	Log-Likelihood	Word
1	381	7480.344	MobilePro
2	268	3823.955	battery
3	252	3136.659	PC
4	74	1452.875	ActiveSync
5	113	1246.475	host
6	60	1128.747	USB
7	97	1127.956	Microsoft
8	98	1082.166	port
9	86	984.580	Using
10	57	968.236	NET

Figure 1: Sample Word Analysis

The importance of this approach is that it affords immediate insight on the sample document. Specifically, it allows us to identify words that have greatest impact upon the content delivered by the document. In addition, an initial check of the sample document against the reference corpus allows us to identify those words that are least common in general use. Access to such information allows the document author to review the use of 'uncommon' words.

Note from the example in Figure 1 that several of the words with greatest significance are proper nouns. On reflection, this should be expected. Such terms are likely to refer to aspects of the facility being documented that are unique to that system or device. In consequence, newly coined descriptors (neologisms) and product names will figure highly in the frequency count and significance measure for the documentation. This results in low (or non-occurrence) in the reference corpus, and allied to high incidence in the sample document determines a high impact upon the document content.

This insight, combined with the 'familiarity rating' for uncommon words, is the first step in our planned support for the document author. Armed with a ranking of words used in the documentation, the author can better target supplementary information such as clarification of newly coined product-specific terms and elaboration of uncommon words. Additionally, we can offer synonyms that are known to have greater frequency in our reference corpus and are thereby likely to be more familiar to the general reader. Thirdly, inclusion of a thesaurus will also permit us to indicate words with common multiple meanings (such as 'port') and sensitise the author to potential problems with reader interpretation.

3.2.2 Collocational frequency analysis
Our next documentation support measure is based upon frequency of occurrence for collocations, rather than individual words. According to Choueka (1988), a collocation is 'a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact

and unambiguous meaning cannot be derived directly from the meaning or connotation of its components' (p609). This definition assumes adjacency of words, but a phrase may be regarded as a collocation even if its component words are not consecutive. Three identifying criteria for a collocation are noted by Manning and Schütze, (1999, p. 172):

- **Non-compositionality**: The meaning of a collocation cannot be directly derived from the meaning of its parts. Either the meaning is completely different from the free combination (as in "kick the bucket") or there is an added element of meaning to the whole phrase that cannot be predicted from the parts.
- **Non-substitutability**: Components of a collocation cannot be replaced by similar ones and still hold the same meaning. For example, in "white wine", we cannot substitute white with yellow (and say "yellow wine"), even though yellow is a good approximation of the wine's true colour.
- **Non-modifiability**: Collocations cannot be freely modified by addition of words or by grammatical changes. For example, the idiom "apple of your eye" cannot be changed to "apples of your eye" or "apple of your beautiful eye" without affecting the collocation.

In deriving our collocations from the document samples and for the reference corpus, we again apply the log-likelihood measure. This is widely used in computational linguistics for collocation extraction and can be used for the extraction of multiword collocations.

Our use of collocation frequency echoes our application of word frequencies and similarly employs reference frequencies derived from the British National Corpus. We first create a reference frequency list of collocations from the BNC then for any collocation that appears in the sample text, we can check whether this is a common or uncommon usage. This enables us to present an indication of likely familiarity against collocations employed in the documentation and further supports the author in tailoring the text. Our supposition is that this 'familiarity ranking' feedback will allow authors to focus on forms of expression that achieve the desired meaning yet tend toward more common and, thereby, more comprehensible, collocations. For the present, we eliminate word clusters from the sample texts that include highly infrequent keywords, such as trade names and proper nouns. Such clusters are unlikely to appear within the collocational frequency reference list and may not be readily contrasted with other collocations.

4. Summary and Conclusions
The DoST approach to documentation support employs two frequency-based strategies as a basis for author feedback. The first of these relies upon word frequency. The second extends this to collocation frequency. These methods allow us to target support that would not otherwise be available to documentation authors.

Firstly, we derive a ranked frequency list for the terms used in the documentation, in order to highlight the relative commonality (and thereby, likely familiarity) of these words. Secondly, we apply a measure of statistical significance to identify the 'significant' words in the sample that are key to its content. This keyword extraction highlights proper nouns, topic specific terminology and other terms of importance, so that the author can ensure these ingredients are adequately supported and explained in the documentation. Thirdly, we check the likely familiarity for collocations used in the documentation text. This discounts word clusters that employ proper nouns and similarly restricted terms and compares others with a reference collocation frequency list that we have derived from the BNC. This permits us to display a 'content impact' ranking for key collocations and focuses the documentation author's attention on 'mainstream' forms of expression, that are most likely to be understood by a general audience.

By addition of a thesaurus, we can further assist in comparing terms with similar meaning but different 'familiarity values', thereby permitting the author to sample alternative forms of

expression that are likely to be more comprehensible to the end-user. The full DoST support sequence is shown in Figure 2.
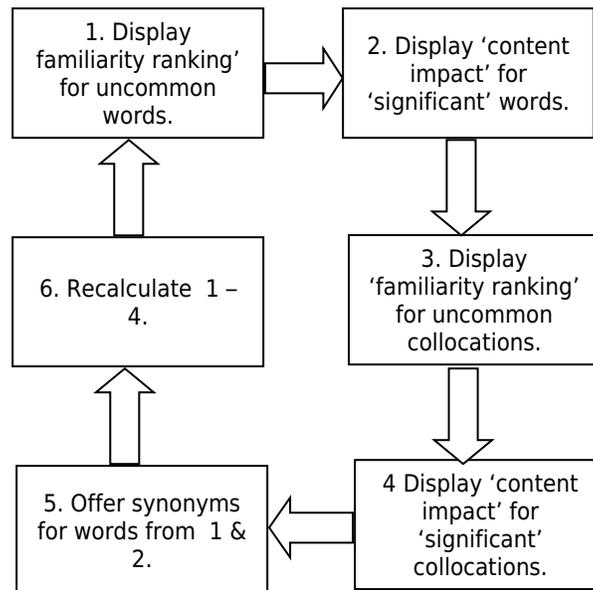

Figure 2: DoST Support Sequence

We have developed components of the DoST facility to evaluate the content of existing documentation by applying textual analysis. This approach will move toward empirical testing and integration in an on-line support facility for documentation authors. Our objective is to generate timely feedback on the likely familiarity of text ingredients, toward optimised content clarity in the authored technical documentation.

References

Anthony, L. (2004) "AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit." Proceedings of IWLeL (An Interactive Workshop on Language E-Learning) 2004: 7–13.

Campbell, G. and G. R. S. Weir, 'Matching Readers to Texts with the Strathclyde Readability Measure', in G. R. S. Weir & T. Ozasa (Eds), *ICT in the Analysis, Teaching and Learning of Languages*, forthcoming, 2007.

Choueka, Yaacov (1988). Looking for needles in a haystack. Proceedings of RIAO '88, pp. 609–623.

Dale, E., and J. S. Chall, *Readability Revisited: The new Dale-Chall readability formula*. Massachusetts: Brookline Books, 1995.

DuBay, W. H., *The Principles of Readability*. Costa Mesa, CA: Impact Information, 2004. URL: http://www.impact-information.com/impactinfo/readability02.pdf (Last Accessed 29 Nov. 2006).

Kittredge, R. and J. Lehrberger, *Sublanguage: Studies on Language in Restricted Semantic Domains*, de Gruyter, Berlin, 1982.

Leech, G. (1993). 100 million words of English: a description of the background, nature and prospects of the British National Corpus project. English Today 33, Vol. 9, No. 1, Cambridge University Press.

Lindgaard, G., *Usability testing and system evaluation*, Chapman & Hall, New York 1994.

Mandel, T. *The Elements of User Interface Design*. John Wiley and Sons, Inc., 1997.

Manning, C. D., Schütze, Hinrich (1999). Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA.

McEnery, T. & A. Wilson, *Corpus Linguistics*, EUP, Edinburgh, 2001.

P. Rayson and R. Garside. Comparing corpora using frequency profiling. In The workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000), 2000.

Rubin, J. & T. Hudson, *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*, John Wiley & Sons, Inc. New York, NY, USA 1994.

Scott, M. (2004) WordSmith Tools 4.0. URL
http://www.lexically.net/downloads/version4/html/index.html

Weir, G. R. S., 'Living with Complex Interactive Systems', in G. R. S. Weir & J. L. Alty (Eds), *Human-Computer Interaction and Complex Systems*, Academic Press, 1991. pp. 1-15.

Weir, G. R. S. and G. Lepouras, 'Localisation and linguistic anomalies', in C. Stephanidis (Ed), *Universal Access in HCI: Towards an Information Society for All*, London, 2001. pp. 535-539.

Weir, G. R. S. and C. Ritchie, 'Estimating Readability with the Strathclyde Readability Measure', in G. R. S. Weir & T. Ozasa (Eds), *ICT in the Analysis, Teaching and Learning of Languages*, forthcoming, 2007.