Automatic Document Quality Control

Neil Newbold and Lee Gillam
University of Surrey,
Surrey GU2 7XH, UK
E-mail: n.newbold@surrey.ac.uk, l.gillam@surrey.ac.uk

Abstract

This paper focuses on automatically improving the readability of documents. We explore mechanisms relating to content control that could be used (i) by authors to improve the quality and consistency of the language used in authoring; and (ii) to find a means to demonstrate this to readers. To achieve this, we implemented and evaluated a number of software components, including those of the University of Surrey Department of Computing's content analysis applications (System Quirk). The software integrates these components within the commonly available GATE software and incorporates language resources considered useful within the standards development process: a Plain English thesaurus; lookup of ISO terminology provided from a terminology management system (TMS) via ISO 16642; automatic terminology discovery using statistical and linguistic techniques; and readability metrics. Results lead us to the development of an assistive tool, initially for authors of standards but not considered to be limited only to such authors, and also to a system that provides automatic annotation of texts to help readers to understand them. We describe the system developed and made freely available under the auspices of the EU eContent project LIRICS.

1. Introduction

The ever-expanding web places a substantial burden on users in filtering, understanding and rapidly processing large volumes of written communication. To achieve the most effective understanding of content in a short time, the need for clear and concise writing gains ever greater importance. Yet the advent of social media, blogs, instant messaging and SMS texts has put emphasis in the opposite direction, on speed of communication at the detriment of clarity. Resources trained on relatively well-controlled uses of language require significant attention to deal with such emergent, and often ephemeral, collections.

How one writes is as important as what one writes, and the use of readily understandable and consistent language is essential for enabling readers to understand written text: it affects their ability to comprehend and assimilate what a writer thinks they are conveying. With extensive work undertaken to increase the accessibility of web pages largely focussing on visual elements and ensuring, for example, that alternative tags have content, such as in the Web Accessibility Initiative, limited attention has been paid to the contribution that simplifying and/or improving the textual content could make (e.g. Boldyreff et al 2001). More generally, this could be considered as a form of textual quality assurance: quality assurance processes aim at removing ambiguity and enabling understandable work, yet quality assurance *per se* appears not to extend directly to the written word. For example, authors of international standards, and here we take ISO as a case in point, demand that written work be precise and comprehensible. However, there is only a small amount of written "guidance" on how to do so, and where it does exist, it is easily ignored: there are no conditions for conformity and compliance. The exploration of readability, and here we consider especially the automation of the quality assurance process, is likely to aid the production of important documents and accessibility of web pages, allow for more accurate machine translation or less-expensive human translation, and provide better source material for agents of the semantic web to find, share and integrate more easily.

We have explored mechanisms relating to content control that could be used in two orientations: the first by authors to improve the quality and consistency of the language used in authoring; the second, a side effect of the first, to find a means to demonstrate this to the reader. To achieve this, we undertook implementation and evaluation of a number of software components, including those of the University of Surrey Department of Computing's content analysis applications (System Quirk). Our work covered the integration and use of supporting resources and components for the standards development process, including a Plain English thesaurus, lookup of ISO terminology provided from a terminology management system (TMS) via ISO 16642, automatic terminology discovery using statistical and linguistic techniques, and comparison of outcomes using five common readability metrics. These components were integrated within an existing framework to demonstrate the potential for controlled authoring based on some of the very standards being used and produced within the EU eContent project LIRICS. In this paper we will describe the system developed and used for assisting in improved readability and demonstrate how improvements can accrue from such analysis. The software can be used within the commonly available GATE software and has been made freely available.

2. Document Readability

Writing standards, in particular, requires a specific approach to style and vocabulary. International standards have specified structures, and content control approaching controlled authoring could be valuable in this arena, and especially where standards involve critical

communications, to ensure that, for example, definitions provided can follow the principle of substitutability (whereby they can be used almost directly in place of words in the text and other definitions). The ability to undertake such a (semantic) task requires the document to be relatively well-written, with consistent terminological use and removal where possible of verbiage and ambiguity. ISO documents transition through a number of stages. Largely, those who are knowledgeable about such things impart knowledge of their construction to others in a relatively ad hoc manner. Furthermore, while this is relatively easily achieved through communication during the authoring process, the disassociation between author and standard-reader can tend to lead to comments about standards being impenetrable. The majority of stages through which ISO documents transition entail assessment – review and commentary – to be made of the document. Specific documents that will be reviewed and commented upon in the development life cycle of an ISO include the Working draft (WD), Committee draft (CD), Draft International Standard (DIS) and Final DIS (FDIS). Comments are generally provided through National Standards Bodies (NSB), where they may or may not be (re-)validated against principles and methods used for the production of the standards. The amount of hidden effort in the production of standards can be significant: assume: these four documents alone, consider the possible commentary from twenty NSBs, where each NSB has a number of people involved who read these documents and who provide comments, and the potential scale of duplicated, or even contradictory, comments across NSBs becomes apparent. Comments from NSBs then have to be merged, filtered, and subsequently dealt with by the editor of the standard.

A variety of measures of readability have been constructed on the basis that sentence length and word length, and in some cases as a function of the number of syllables, are determining factors (Kitson 1921). The results of applying these formulae attempt to indicate the level of education or provide a difficulty score on a scale of 1-100: these formulae and the elements on which they rely are presented in Table 1 – further discussion of these can be found elsewhere (Gillam and Newbold 2007).

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

Table 1: Features of readability metrics

A common feature missing from these metrics is accounting for additional background knowledge that can be provided alongside the document. For example, complexity based on the number of syllables fails to take into account existence of definitions. One outcome from our work, in due course, will be alternative measures for readability that takes account of such considerations.

3. Document Content Management System

New components that have been (re-)engineered and integrated with GATE, building around ANNIE. Existing GATE plug-ins from ANNIE were used for the preliminary NLP tasks, leading into newly devised processing resources. These additional resources and the results of analysis emerging from them will be described as follows:

- Terminology Lookup (3.1)
- Linguistic Term Finder (3.2)
- Keyword Extractor (3.3)
- Statistical Term Finder (3.4)
- SimpleText Analyser (3.5)
- Annotation Controller (3.6)
- Readability Analyser (3.7)
- Replacer (3.8).

The pipeline for these resources is shown in Figure 2, below, with brief descriptions of each component following to provide an indication of the approach. It should be noted that the readability analyser can be run at two separate points in the pipeline, the latter prior to committing changes.

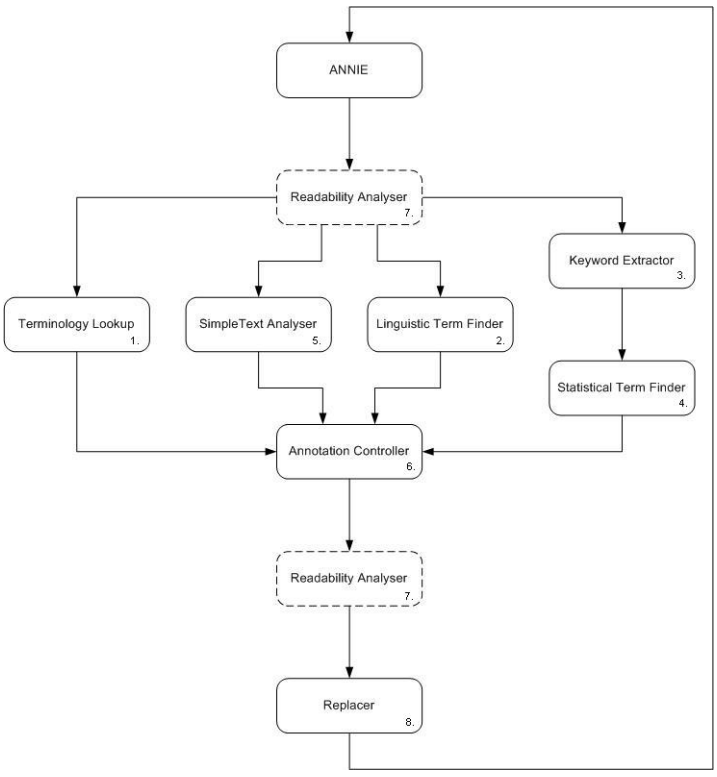

Figure 1: Pipeline for the prototype document content management system

3.1 Terminology Lookup

The *Terminology Lookup* plug-in analyses documents and annotates term entries. It uses an ISO 16642 compatible XML-based terminological markup file containing a snapshot of the terminology collection. The terminology is available in both English and French, potentially providing some assistance for translators also. Providing such a file containing terminology for different domains/applications is a possibility. In this instance, extant terms are annotated using our own collection, though in future this could interoperate directly with the iTerm TMS that has recently been populated with ISO TC37 terminology.

3.2 Linguistic Term Finder

The *Linguistic Term Finder* identifies candidate terms according to specified patterns of part of speech annotations (e.g. Jacquemin 2001, p27) using the ANNIE POS tagger within GATE.

3.3 Keyword Extractor

The *Keyword Extractor* calculates distributions of frequency and weirdness as outlined by Gillam (2004). We use frequency information from the 100 million word tokens of the British National Corpus (BNC) to act as a reference corpus. The extent to which annotations are applied can be adjusted by modifying parameters for the distributions and their combinations.

3.4 Statistical Term Finder

The *Statistical Term Finder* takes input from the Keyword Extractor (3.3). This plug-in examines collocations of the keywords and identifies patterns occurring with statistical significance, following on from the work of Smadja (1993) and Gillam (2004). We use defined thresholds for identification: if a word consistently appears in the user-defined neighbourhood size above the threshold value, it is considered a potential new term. This can be undertaken iteratively (re-collocation). We are exploring automatic determination of this threshold in related work.

3.5 SimpleText Analyser

The *SimpleText Analyser* uses a thesaurus containing words and phrases identified as verbose, and hence deprecate, by either the Plain English Campaign[1] or ASD Simplified Technical English[2]. The thesaurus contains 1302 such entries, offering one or more preferred alternatives for each. The SimpleText Analyser identifies these phrases within the text and offers potential replacements for the expression. We are examining improvements to automation of this aspect.

3.6 Annotation Controller

The *Annotation Controller* is used to reduce the quantity of overlapping annotations being produced by prioritising some annotations over others. This component was added after discovering that conflicting suggestions for improvements were being produced by the components operating in parallel.

3.7 Readability Analyser

The *Readability Analyser* computes the number of words, syllables, sentences, characters and polysyllabic words contained within a document as required by current readability formulae. These values are used for the calculation of readability formulas such as the Kincaid formula, Flesch Index, SMOG, ARI and Fog Index.

3.8 Replacer

The *Replacer* substitutes the text in a document with user-selected SimpleText replacements (3.5). If no "best replacement" is selected, the text is left unchanged. Once the Replacer has finished, the whole procedure can be repeated. When the Readability Analyser is subsequently run, the effect the replacements had on the readability scores of the document are displayed and the user can decide whether any further replacements or additions to the subject terminology are appropriate.

4. Results

To demonstrate results of this analysis, two standards being developed within the LIRICS project, at various stages of the ISO process, have been analysed. The documents 'Lexical markup framework (LMF)' (at Draft International Standard stage[3]) and 'Syntactic Annotation Framework (SynAF)' (at Working Draft stage[4]) were chosen to show the output obtained from the various stages of the analysis.

4.1 Terminology Lookup

All known terms were annotated including, for purposes of inspection, those occurring containing another term within the annotation. For example, 'object language' contains another known term 'object'. The annotation allows access to the definition for the term, and allows relationships between terms to be investigated.

4.2 Term Finder (Keywords, Statistical and Linguistic)

Similar to Terminology Lookup, a discovered term can have annotated elements. For example, in Figure 4 the candidate term 'syntactic annotation' also contains 'annotation'. This new candidate could then become an extension of the existing terminology, potentially referring to 'annotation' as the broader concept. Decisions over the use of such relationships need to be considered. The numbers of known and discovered terms (total count) found in the two documents are detailed in Table 2.

[1] http://www.asd-ste100.org/
[2] http://www.plainenglish.co.uk/

[3] Revision 13, available at: http://lirics.loria.fr/doc_pub/N330_LMF_rev13_For_CD_Ballot.pdf
[4] Available at: http://lirics.loria.fr/doc_pub/SynAF_WD_2006-01-22.pdf

Document	Known Terms	Discovered Terms
Lexical markup framework (LMF)	466	3712
Syntactic Annotation Framework (SynAF)	96	1125

Table 2: Known and discovered terms in the standards

The 'LMF' document was roughly three times the size of 'SynAF', but appears to have substantially more terminological content. The top 10 known terms and their frequencies in 'LMF' are shown with in Table 3, while the top 10 known terms found in the document 'SynAF' along with their frequencies are shown in Table 4.

Term	Count
Class	238
Form	99
Lexical Entry	59
Word	57
Lexicon	46
Data	44
Paradigm	39
Paradigm Class	31
Lemma	30
Extension	27

Table 3: The top 10 known terms with their frequencies in 'LMF'

Term	Count
type	28
label	15
data	14
definition	9
object	9
information	6
merge	2
parsing	2
read	2
context	1

Table 4: The top 10 known terms with their frequencies in 'SynAF'

Discovered term candidates were evaluated to determine which could be considered as terms. The terms highlighted by both linguistic and statistical methods were prioritised for consideration. Terms such as 'syntactic annotation', 'annotation', 'SynAF' and 'morph' were identified. Further filtering of this list is required, but frequency information can be helpful here also; variations by part of speech can lead to duplications, for example for 'SynAF'. Examples of discovered terms from SynAF are shown in Table 5. Additionally, the linguistic and statistical methods for discovering terms found numerous valid two word expressions that were regularly used. Examples of these are shown in Table 6.

Term	Linguistic Discovery	Statistical Discovery	Count
* annotation	N	Y	42
head	Y	N	33
value name	Y	N	22
partec	Y	Y	21
* synaf	Y	Y	19
value	Y	N	18
edge label	Y	N	14
syntactic annotation	Y	Y	13
mod	Y	N	11
morph	Y	Y	11
* synaf	N	Y	11
word	Y	N	11
* annotation	Y	Y	10
constituency	Y	N	10

Table 5: Examples of highly frequent discovered terms in 'SynAF', including duplications due to different parts of speech (*)

Term	Linguistic Discovery	Statistical Discovery	Count
sense class	Y	N	14
lexicon instance	Y	N	8
core package	Y	N	6
sense instance	Y	N	6
external system	Y	N	5
lemma class	Y	N	4
narrative description	Y	N	4
word forms	Y	N	4
affix class	Y	N	3
affix slot	Y	N	3

Table 6: Examples of frequent bigrams in 'LMF'

Further notable keywords (single words) are shown in Table 7.

Term	Linguistic Discovery	Statistical Discovery	Count
LMF	N	Y	34
ISO	Y	N	27
subcategorization	N	Y	24
multilingual	N	Y	18
verb	Y	N	11
inflectional	N	Y	9
agglutination	Y	N	8
UML	N	Y	8

Table 7: Examples of discovered single-word terms in 'LMF'

Discovered terms of increased length at lower frequencies indicate the existence of potentially highly complex expressions. Here the statistical approach requires further treatment that is beyond the scope of this paper. The combination of the two methods of identifying potential new terms allows for readability issues that might be caused by ambiguous bracketing to be highlighted. Such a readability issue can be demonstrated by the first item in Table 8, the "complex knowledge organization system":

1. **[**complex knowledge**] [**organization system**]**: an organization system for complex knowledge, simple knowledge is excluded?
2. **[**complex**] [**knowledge organization system**]**: a knowledge organization system that is somehow complicated?
3. **[**complex knowledge organization**] [**system**]**: the system is for an intricately arranged "knowledge organization"?

Table 8 further demonstrates term inclusion: "data category" is a term from "ISO 1087-2:2000 Terminology work - Vocabulary - Part 2: Computer applications" and "data category selection" is defined in "ISO 12620:1999 Computer applications in terminology - Data categories". We find 2 instances of "lmf data category selection procedures", which appears to extend this notion somehow. Interpretation, however, remains an exercise to be undertaken by the document author, for the two instances, or an exercise that will be undertaken by the readers. The discovery of the "multi-layered annotation" and its "strategy", or perhaps the "annotation strategy" and its multiple layers, may also suggest the correct interpretation should be made clear.

Term	Linguistic Discovery	Statistical Discovery	Count
complex knowledge organization system	Y	N	4
lmf data category selection procedures	Y	N	2
semantic predicate class section	Y	N	2
dual use mrd metamodel	Y	N	2
dual use mrd package	Y	N	2
multi-layered annotation	N	Y	3
multi-layered annotation strategy	Y	N	2

Table 8: Examples of potential multiword terms that were discovered in 'LMF'

4.3 SimpleText Analyser

In analysis using a subset of the Plain English substitutions against a further document, ISO/DIS 12620, a report of substitutions for words and phrases deemed unnecessarily complex was produced. The first 200 suggested replacements were analysed manually, with 33 individual replacements found to be suitable. Every further instance of these replacements was tested throughout the remainder of the document, 183 instances in total, to see if the replacements were appropriate in every instance. 65 of these potential replacements were valid, providing us with initial indications of the likelihood of success in making replacements automatically, though substantial further testing is required. Some replacements were appropriate in every further instance such as "comprises", "in order to", "permissible" and "thus". However, some rarely had correct replacements and in particular 'application' and 'component' were never suitable. A large proportion of the proposed SimpleText replacements were found not to be suitable within the contexts we encountered, however the potential remains for more in-depth analysis of these constructions. The suggestions, replacements evaluations are detailed in Table 9.

Phrase	Replacement	Occurs In Text	Replaced	% Correct
application	use	17	1	5.88%
by means of	by	2	2	100.00%
component	part	68	1	1.47%
comprises	is made up of	4	4	100.00%
consequence	result	1	1	100.00%
essential	important	2	2	100.00%
frequently	often	1	1	100.00%
in conjunction with	with	2	2	100.00%
in order to	to	4	4	100.00%
instances	cases	3	3	100.00%
latest	last	2	2	100.00%
nature	type	1	1	100.00%
needed	necessary	1	1	100.00%
permissible	allowed	4	4	100.00%
provide	give	19	3	15.79%
represent	show	6	2	33.33%
requirements	rules	4	2	50.00%
restrict	limit	1	1	100.00%
revised	changed	1	1	100.00%
specified	given	5	4	80.00%
thus	therefore	4	4	100.00%
utilize	use	1	1	100.00%
various	different	10	4	40.00%
within	in	20	14	70.00%

Table 9: Replacements filtered from initial suggestions, with the number of times the replacements were correct throughout the rest of the document

4.4 Readability analysis

Having produced a limited number of effective substitutions, the Replacer (3.8) was run against the modified text to determine the influence on readability scores. Scores for Kincaid, Flesch and ARI reduced slightly while FOG and SMOG increased slightly. For

FOG and SMOG, this is likely due to the fact that some SimpleText replacements increase the number of words in the document while not reducing the number of *complex words*. The most common example is the substitution of "comprises" for "is made up of". Other replacements such as "important" for "essential" have no effect on readability scores as the number of syllables and characters are identical. Readability scores before and after the replacements are shown in Table 13.

Score	Before	After
Kincaid	14.753	14.747
Flesch	28.534	28.611
FOG	*17.234*	*17.254*
SMOG	*15.432*	*15.447*
ARI	14.408	14.398

Table 10: Readability scores before and after the SimpleText process

5. Conclusion

To provide for a variety of aspects of additional quality control, in addition to the extant processes of ISO, we have integrated and used a variety of supporting resources and components for the standards development process, including a Plain English thesaurus, lookup of ISO TC 37 terminology provided from a terminology management system (TMS) via ISO 16642, automatic terminology discovery using statistical and linguistic techniques, and readability metrics. These components have been re-engineered from the University of Surrey Department of Computing's content analysis applications (System Quirk), developed in prior research, including EU co-funded projects, and integrated with the University of Sheffield's GATE system. These efforts were undertaken to demonstrate the potential for controlled authoring in the International Standards environment. The result of these efforts leads us to the development of an assistive tool for authors of standards based around, and evaluated against, LIRICS work.

These experiments helped us to provide some additional commentary into ISO on several standards documents at various stages of the ISO process; fuller sets of commentary for the LIRICS standards are at various stages of production and this deliverable presents some examples of how these can be formulated. Human interpretation of, and action upon, the results being produced by these components is still required to varying extents, however the analysis of language simplicity and consistency, identification of known and unknown terms, and the generation of "understandability" metrics have all been implemented and demonstrate interesting and potentially highly-valuable results. There are a number of further investigations required in relation to these components, and further evaluation efforts are needed to assess the results being produced, to improve the treatment provided and to improve the formulation of feedback on the document or documents being analysed. The ideal outputs would be provided directly to standards authors prior to the submission of a document into the ISO processes, potentially leading to a reduction in both the workload associated within the process, and the cognitive load of the reader. Further work to foster adoption of such an approach into the authoring process is still required.

6. Acknowledgements

This work has been supported, in part, by the EU eContent project LIRICS (22236) and the UK's EPSRC project REVEAL (GR/S98443/01).

7. References

Boldyreff, C., Burd, E., Donkin, J. and Marshall, S. (2001). The Case for the Use of Plain English to Increase Web Accessibility. In *Proceedings of the 3rd Intl. Workshop on Web Site Evolution (WSE'01).*

Cunningham, H., Maynard, D., Bontcheva, K. and Tablan, V. (2002). 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.

Gillam, L. (2004). Systems of concepts and their extraction from text. Unpublished PhD thesis, University of Surrey

Gillam, L. and Newbold N. (2007) Quality Assessment. Deliverable 1.3 of EU eContent project LIRICS. URL: http://lirics.loria.fr/doc_pub/T1.3Deliverable.final.2.pdf. Last accessed 19 Feb. 2008.

Jacquemin, C. (2001) *Spotting and Discovering Terms through Natural Language Processing*. The MIT Press.

Kitson, H. D. (1921). *The mind of the buyer*. New York: Macmillan.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1) pp.143--178.