Assessing the Readability of Clinical Documents in a Document Engineering Environment

Mark Truran
School of Computing
Teesside University
United Kingdom
m.a.truran@tees.ac.uk

Gersende Georg
Haute Autorité de Santé
Saint-Denis La Plaine Cedex
France
g.georg@has-sante.fr

Marc Cavazza
School of Computing
Teesside University
United Kingdom
m.cavazza@tees.ac.uk

Dong Zhou
Dept. of Computer Science
Trinity College Dublin
Ireland
dong.zhou@cd.tcd.ie

ABSTRACT

Previous work has established that specific linguistic markers present in specialised medical documents (clinical guidelines) can be used to support their automatic structuring within a document engineering environment. This technique is commonly used by the French Health Authority (la Haute Autorité de Santé) during elaboration of clinical guidelines to improve the quality of the final document. In this paper, we explore the readability of clinical guidelines. We discuss a structural measure of document readability that exploits the ratio between these linguistic markers (deontic structures) and the remainder of the text. We describe an experiment in which a corpus of 10 French clinical guidelines is scored for structural readability. We correlate these scores with measures of textual cohesion (computed using latent semantic analysis) and the results of a readability survey performed by a panel of domain experts. Our results suggest an association between the density of deontic structures in a clinical guideline and its overall readability. This implies that certain generic readability measures can henceforth be utilised in our document engineering environment.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Medical information systems; I.7.2 [**Document and Text Processing**]: Document Preparation—*Markup languages*; I.2.7 [**Natural Language Processing**]: Text Analysis

General Terms

Experimentation, Languages, Theory, Measurement, Human Factors

DocEng2010, September 21–24, 2010, Manchester, United Kingdom.

Keywords

Latent semantic analysis, LSA, readability, cohesion, medical document processing

1. INTRODUCTION

Clinical guidelines are specialised medical documents describing the appropriate treatment and care for patients with specific conditions. They are evidence-based, best practice resources designed to assist healthcare professionals in their work and medical students in their training. The elaboration of these guidelines is an inherently complex task, typically requiring thematic panels of expert physicians working within a complex production life cycle. For this reason, the generation (and dissemination) of clinical guidelines has often been delegated to national bodies who can command sufficient resources and expertise (e.g. the National Institute for Health and Clinical Excellence (NICE) in the United Kingdom, the National Guidelines Clearinghouse (NGC) in the United States and la Haute Autorité de Santé (HAS) in France).

In recent years the elaboration of clinical guidelines at HAS has been simplified by the introduction of a text analysis environment known as G-DEE (for *Guidelines Document Engineering Environment*) [14, 12]. The G-DEE platform supports a number of functions useful during the development of clinical guidelines. One of these functions involves the automatic identification of *deontic structures* in the text indicating the presence of *clinical recommendations*. These structures are recognised using shallow natural language processing (NLP) techniques and automatically marked-up to provide the first level of document structuring. Since 2007 the G-DEE platform has become firmly embedded in the daily workflow of the HAS - it is commonly used on all incarnations of a guideline, from the first draft to the definitive version.

In this paper we describe a new readability measure for clinical guidelines that exploits deontic structure. The origin of this measure can be found in [13] where Georg. et al. speculated that the distribution of recommendations within a clinical guideline could be used as a quality indicator during the guideline elaboration process (i.e the more recommendations a guideline contains, the better structured the

document and therefore the easier it is to read). In the following pages we attempt to provide an empirical foundation for this speculation by examining the relationship between the deontic density of a guideline and its overall readability.

The remainder of the paper is organised as follows - in §2 we discuss previous work on readability measures. In §3 we try to correlate deontic density with the overall readability of a clinical guideline. In §4 we present our experimental results. In §5 we discuss the limitations of our experiments and propose future research directions. In §6 we conclude our discussion.

2. RELATED WORK

2.1 Traditional Readability Indices

Readability indices have been commonly used by writers and educators for over 80 years [35, 29]. Traditionally, these indices exploit the lexical and syntactic features of text (e.g. sentence lengths, syllable counts, word frequencies etc.) to produce a numerical score indicating its reading difficult [10, 7, 5]. These measures have proved consistently popular due to their computational simplicity and predictive reliability [6]. However, readability formulae of this type have attracted sustained criticism for a variety of reasons [27]:

- They attempt to reduce an inherently complex process to a limited set of measurable variables.
- Their emphasis on surface readability features fails to take into account the way in which a reader interacts with a text [8].
- They are not based on any theory of reading or reading comprehension, just empirical correlations
- They are generally based on traditional student populations reading academic texts. Their suitability for other materials (e.g. technical texts, domain specific material) and other reading populations is questionable [32].
- Although most of the readability metrics use the same ground truth (i.e. typical American grade schools reading levels), inter-correlating sets of results generated by different tests produces serious anomalies [27].

2.2 The 'New Readability'

For the reasons discussed above, the so-called 'traditional' readability formulae have been partially displaced during the last three decades by measurements that recognise the psycholinguistic factors of reading *comprehension*. This 'new readability' is essentially a mixed model approach, a confluence of several disciplines (e.g. computational linguistics, corpus linguistics, information extraction and discourse analysis) that together '*supercede surface components of text and language comprehension and instead explore deeper, more global attributes of language*'[6]. Comprehension measures eschew the 'micro-structure' of text in favour of its 'macro-structure' [2], appropriating techniques such as POS tagging, statistical language modelling, machine learning and term frequency analysis to go beyond shallow syntactic analysis. Readability measures derived using these newer techniques do not, as a rule, correlate well with traditional readability formulae [26].

One useful concept that has been developed by researchers seeking to estimate reader comprehension is 'textual cohesion'. In the field of discourse analysis, a cohesive discourse is a text that 'hangs together' (in both a logical and a rhetorical sense). Numerous influential studies have shown that there is a direct correlation between the cohesion of a text and its reading comprehension [34, 26], so that changes to the structural or explanatory cohesion of a text result in significant increases in recall amongst readers [3].

2.3 Measuring Cohesion

The estimation of textual cohesion has been attempted using a variety of techniques. Most of these techniques address one (or more) of the five cohesive devices identified by Halliday and Hasan [20]. The first of these devices is the *reference*. Referential cohesion between textual units can be measured by identifying the incidence of *argument overlap* [25]. Argument overlap is something that '*...occurs when a noun, pronoun, or NP is one sentence is a co-referent of a noun, pronoun or NP in another sentence*'[16]. Several measurements exploiting this basic propositional model have been developed, including variants exploiting noun, stem and content word overlap at both local and global levels.

The second cohesive device is the *conjunction*, which establishes a relationship between two clauses. Connectives of this sort are extremely important when assessing text cohesion generally. Researchers have examined both the density of connectives and the frequency of differing types of connectives (additive, temporal, clarifying etc.) as potential indicators of overall textual cohesion [17]. Causal connectives (e.g. because, accordingly) have been used by researchers to measure the *casual cohesion* of text. Causal cohesion is an appropriate measure when the underlying text describes a series of events or actions that are related causally, like the sequences in a play or steps in a recipe [18].

The third and fourth cohesive devices identified by Halliday and Hasan are *ellipsis* - which occurs when specific words are omitted after a more specific mention - and *substitution*, which occurs when a word is substituted for a more general word. These devices are less popular as indicators reflecting the cohesion of text, presumably due to the computational complexity of establishing their frequency and effect. The final cohesive device relates to *textual cohesion*. Textual cohesion is created by repetition of the same word or of a set of lexemes sharing the same semantic features. A number of different approaches have been used to measure textual cohesion automatically [31]. The most successful to date uses an approach called *latent semantic analysis* (LSA)[9].

2.4 Latent Semantic Analysis

Latent semantic analysis is a well-established method for computing the contextual-usage meaning of words. In its approach LSA shares many similarities with the vectorial methods employed in information retrieval. The technique takes as its input a large training corpus segmented into meaningful *passages* (e.g. documents, paragraphs, sentences etc.). This corpus is represented as a rectangular matrix in which columns represent passages and rows represent unique words occurring in two or more passages. Each cell in this word-by-context matrix contains a frequency value weighted to reflect the overall distribution of the term across the corpus. Initially a log-entropy transform was employed to ef-

fect this weighting, but several variants (including term frequency x inverse document frequency $TF \times IDF$) have become popular [30].

This weighted rectangular matrix is subjected to *singular value decomposition* (SVD), which is a form of factor analysis, thereby creating a high dimensional *semantic space*. Thereafter, textual units extracted from a target text can be compared for semantic relatedness using this semantic space in combination with the well-known *cosine similarity measure*. Two textual units using the same words with the same frequency would score 1 when measured. Two textual units that do not share any semantically related terms would score near -1. The majority of pairwise comparisons would produce a value somewhere in between. Crucially, the order of the words in the textual units is immaterial and literal word overlap is unnecessary [28].

Latent semantic analysis has proven remarkable popular during the last two decades. Its core applications cluster around the fields of natural language processing [4] and information retrieval [9], but it has also been applied to a wide variety of other problem spaces including automatic assessment [19], information filtering, machine learning and speech recognition. One further application field for LSA corresponds to the prediction of reader comprehension via textual cohesion [11, 16]. In this context, textual cohesion is typically measured by comparing the semantic relatedness of each pair of adjacent sentences in the target text (using the cosine measure discussed above). Textual cohesion is generally assumed to increase as the mean cosine value for the target text increases.

3. METHODOLOGY

3.1 Overview

In the following section we describe two experiments which attempt to correlate the deontic density of a clinical guideline with its overall readability. The first experiment considers the relationship between deontic density and the readability of each guideline *as a whole* (i.e. document-level readability). The second experiment examines the relationship between deontic density and the readability of finer grained text passages (i.e. paragraph-level readability). Further details describing the technical aspects of each experiment, as well as the necessary pre-processing, are provided below. Our results can be found in §4.

3.2 Document-Level Readability

Our experimental corpus was a set of 10 plain text clinical guidelines published by the HAS. These guidelines cover a wide range of medical issues including Alzheimers disease, heart disease, obesity and cocaine addiction. The first stage of our experiment involved processing these guidelines using the G-DEE document engineering platform, essentially repeating the experiment described in [13]. During processing, the text of each guideline was analysed using a finite state transition network (FSTN), a relatively simple NLP device for recognising/producing text strings. This network is constructed using around 170 syntactic patterns extracted during corpus analysis, corresponding to 65 deontic operators (e.g. forbid (*interdire*), authorise (*autoriser*), ought to (*devrait*) etc.). There were approximately 12,000 FSTN nodes in the network (taking morphological variants into account). Individual deontic expressions were grouped together when they shared common syntactic patterns. Processing time using this FSTN was extremely reasonable. A 26 page document, approximately 920 lines of text, was processed in 300ms on a standard PC.

The output of this processing was a set of 10 clinical guidelines marked up to indicate the presence of deontic structures. We calculated the *raw* frequency of deontic structures in each document d as well as a *normalised* deontic frequency NDF as follows:

$$NDF = \frac{DSW}{L} \tag{1}$$

where

DSW is the total number of words in deontic structures

L is the total number of words in the document

This calculation produced a number in the interval [0,1]. We used this number to perform a rank ordering of the clinical guidelines in which the assigned rank of a guideline was positively related to its normalised deontic frequency (i.e. the highest ranking was assigned to the guideline with the highest NDF score).

In the next stage of the experiment we used latent semantic analysis to measure the cohesion of each clinical guideline. We started with the 10 plain text medical guidelines, encoded in ISO 8859-1 (Western Europe). Before any preprocessing was applied to this set of guidelines it contained 2989 sentences with 8578 unique terms. We converted the guidelines to UTF-8 encoding, then applied a stemming algorithm designed for French text (part of the Apache Lucene project - http://lucene.apache.org/) and a French stopword list provided by the University of Neuchatel [33]. Next we segmented each guideline into sentences treating structural headings (e.g. *s.1 Introduction*) and individual bullet points as sentences. Finally, we removed all sentences containing less than three words (a common technique in this context). After pre-processing, the set of clinical documents contained 2215 sentences and 3339 unique terms.

The next task was construction of the semantic space. Given the nature of the documents under consideration, we required a specialised textual representation. We built this using web resources provided by the Grenoble Faculty of Medicine (the ALPESMED corpus) [22]. These resources were developed as a teaching reference for 2nd and 3rd year medical students. We selected this corpus because it contained a large proportion of the terms used in the HAS guidelines and it had been approved by a European quality assurance body (MEDCIRCLE). The Grenoble corpus is divided into 31 high level topics (anaesthesia, cardiology etc.) which are split into individual modules (for example, urology is broken down into '*congenital malformations of the urinary tract*' and '*Kidney trauma*'). There were 275 modules in total. Each module contained information in both HTML and PDF formats.

We performed a recursive crawl of the Grenoble corpus using the WGET tool (http://www.gnu.org/software/wget/). This produced 893 files altogether, a collection of HTML, PDF, JPG and other image files approximately 70 MB in size. We separated the PDF files from the rest of the corpus and converted each one to a UTF-8 encoded plain text file using the *xpdf* suite of tools (http://www.foolabs.com/xpdf/). This produced 275 documents containing 547,788 words in

total. We selected 60 of these documents at random (4922 sentences, 63398 terms, 4922 unique terms) and we built a rectangular matrix (*terms* × *sentences*), with raw TF values in the cells of the matrix. We used the Terrier information retrieval platform (http://terrier.org/) to index the ALPESMED documents and we calculated IDF for every term in the corpus. Then we weighted the values in the matrix cells using the TF x IDF weighting scheme and performed the SVD using JAMA (Java Matrix package), a free linear algebra package for manipulating real, dense matrices [1]. The SVD operation took approx. 6 hours on a standard desktop PC. The output of the operation was a very large text file approaching 900MB in size.

Having pre-processed the clinical guidelines and computed a suitable semantic space, we measured the textual cohesion of each clinical guideline in turn. We did this by loading the SVD text file into memory (an operation requiring 2-3 seconds), then computing the cosine similarity between all pairs of adjacent sentences $cos_{si,si+1}$ in every guideline (with dimensionality determined using the *Frobenius* norm [21]). Each sentence-sentence comparison was very quick (less than 400ms). Processing each guideline took less than a minute. We then calculated the overall cohesion COH of a document d having n sentences as follows:

$$COH = \frac{\sum_{n-1}^{i=1} cos_{i,i+1}}{n-1} \quad (2)$$

This calculation yielded a number in the range [-1,1]. We then performed a rank ordering of the clinical guidelines in which the assigned rank of a guideline was positively related to its overall cohesion (i.e. the more cohesive a guideline was, the better the assigned rank).

Finally, to provide a useful comparison, we calculated the readability of each guideline using a 'traditional' readability metric. The metric we applied was the Flesch Reading Ease (RE) index calibrated for use with French text [23]. To determine the readability of a text using this index we calculated:

$$FRE = \frac{209 - (0.68 \times SW) - (1.15 \times SL)}{100} \quad (3)$$

where

SW is the average number of syllables per word

SL is the average sentence length (in words)

This operation produces a value in the range [0-1] with 1 indicating a text which is very easy to read. We calculated FRE for each clinical guideline and used this value to produce an ordered ranking of all the guidelines wherein ranking was positively related with reading ease (i.e the higher the score, the higher the rank).

3.3 Paragraph-Level Readability

In the second experiment we wanted to examine the performance of the structural readability measure on finer grained text passages. We began by extracting a sample of paragraphs from the raw clinical guidelines. Our sampling procedure was designed to ensure a loose uniformity across the samples consistent with the various readability analyses (automated or otherwise) we needed to apply. For each guideline in the corpus-

1. We removed the *Introduction* section entirely. It is a generic section present in all 10 guidelines. It could be removed with no impact on the experiment.
2. We removed all numbered headings (e.g s.2 Douleur chronique : définition et épidémiologie (*trans. Chronic pain: definition and epidemiology*))
3. We removed all obvious unnumbered headings.
4. We assumed that all remaining blocks of text separated by 2 carriage returns (or more) were paragraphs.
5. We removed all paragraphs having less than three sentences.
6. We removed all paragraphs having less than 100 words in total.
7. We removed all paragraphs having more than 200 words in total.

The filtering process described above left us with 87 chunks of text. These paragraphs ranged from 101 to 181 words in length, with a mean word count of 129.

We scored the textual cohesion of each paragraph using LSA (as described above). This produced a fairly even spread of values in the range of 0.84 to -0.20 (arithmetic mean 0.16, stdev 0.214). Using an empirical distribution function, we established quartiles for the raw data ($Q_1 = -0.007, Q_2 = 0.115, Q_3 = 0.275$) and randomly selected 20 paragraphs with cohesion scores above the third quartile (i.e. the most coherent paragraphs) and 20 paragraphs with cohesion scores below the first quartile (i.e. the least coherent paragraphs).

In the next stage of the experiment, these paragraphs were passed to 8 individuals closely involved in the elaboration of clinical guidelines. These domain experts were asked to read all 40 paragraphs and score their readability using a 5-point Likert scale (see Figure 1)). For each paragraph, the experts selected one of the following statements:

1. Le sujet du paragraphe est immédiatement reconnaissable et son contenu est clairement accessible (*The topic under discussion can be identified immediately and the section contents are clear*).
2. Le paragraphe est compréhensible mais le style est un peu lourd (*The section is readable but rather tedious*).
3. Le paragraphe demande des efforts de lecture pour être bien compris (*The section requires reading effort to be properly understood*).
4. J'ai du relire plusieurs fois certaines phrases pour être sur(e) (*I had to read some sentences again to understand their meaning*).
5. Même après plusieurs lectures du paragraphe, il reste des ambiguïtés (*Even after reading the section several times, its meaning is not entirely clear*).

We merged the combined responses of the panel to determine a score for each paragraph (lower scores indicating *higher* readability). Finally, we calculated the normalised deontic frequency and Flesch index score for each chunk of text. The results of this experiment are shown below.

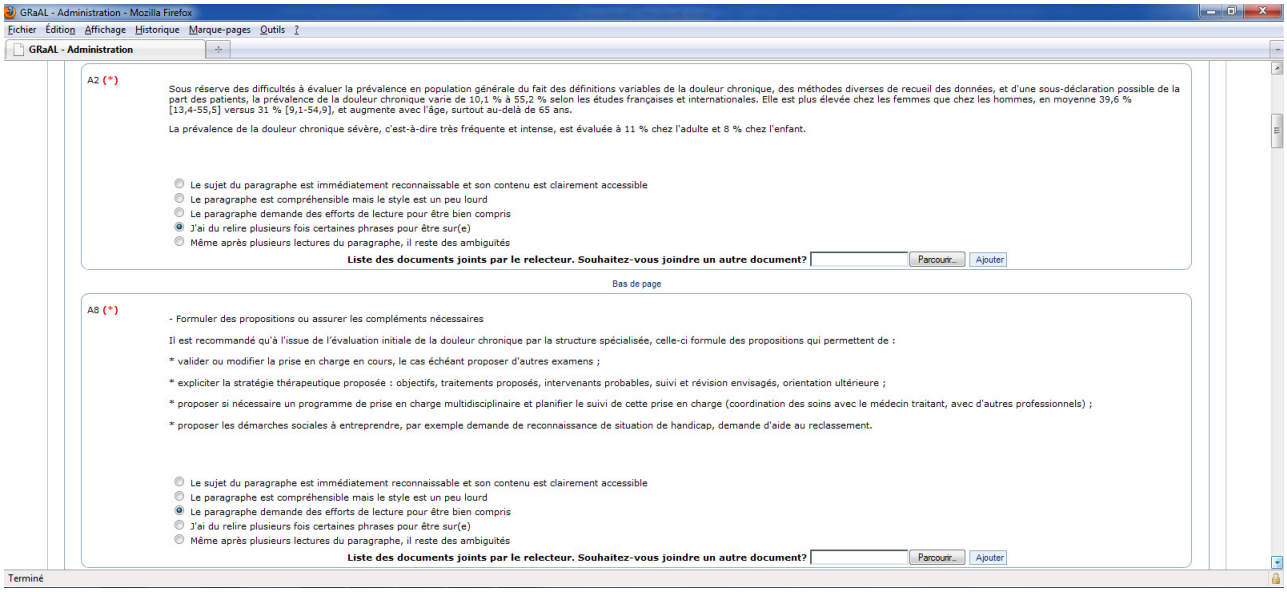

Figure 1: Screen shot of the survey software used to assess the readability of 40 randomly selected paragraphs

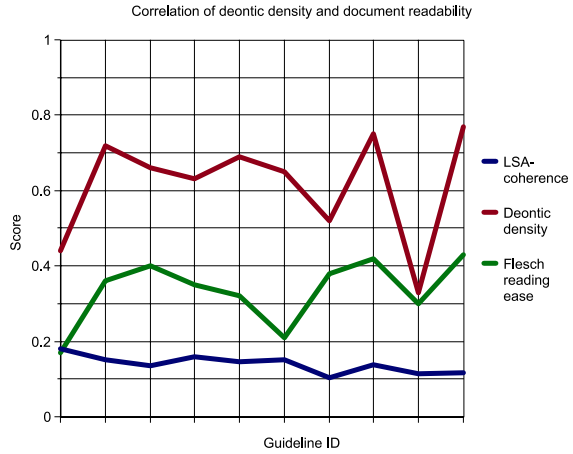

Figure 2: Graph illustrating the relationship between document-level readability and deontic density

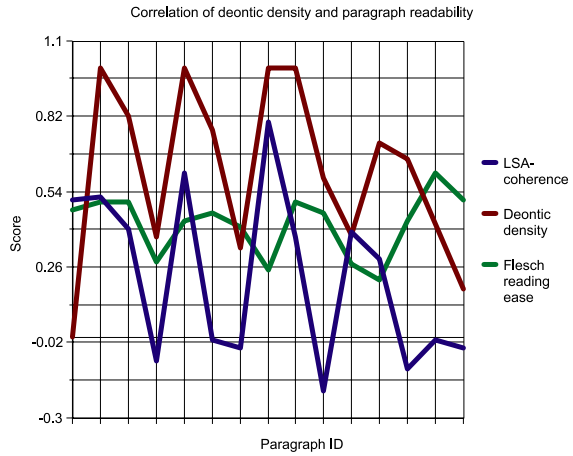

Figure 3: Graph illustrating the relationship between paragraph-level readability and deontic density

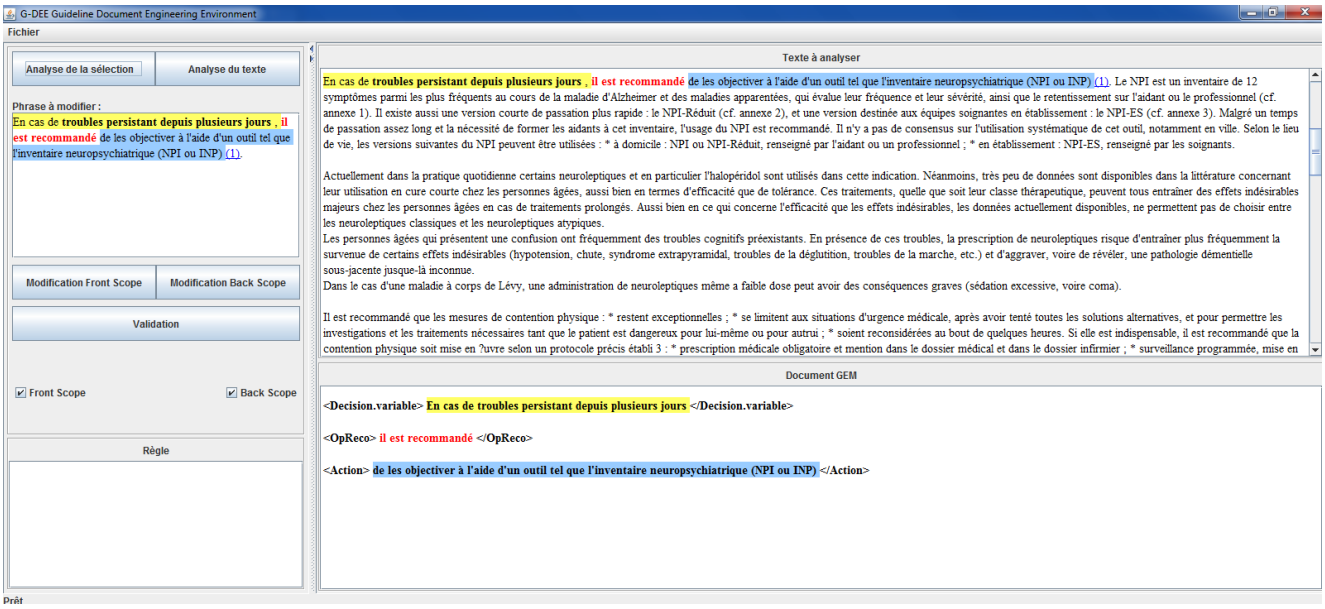

Figure 4: Identifying the conditional statements in the clinical recommendations

RANK	NDF	LSA	FRE
1	J (0.77)	A (0.181)	J (0.43)
2	H (0.75)	D (0.159)	H (0.42)
3	B (0.72)	B (0.152)	C (0.40)
4	E (0.69)	F (0.150)	G (0.38)
5	C (0.66)	E (0.145)	B (0.36)
6	F (0.65)	H (0.139)	D (0.35)
7	D (0.63)	C (0.135)	E (0.32)
8	G (0.52)	J (0.118)	I (0.30)
9	A (0.44)	I (0.114)	F (0.21)
10	I (0.33)	G (0.104)	A (0.17)

Table 1: Rank orderings of 10 clinical guidelines (A-J) by normalised deontic frequency, latent semantic analysis, and the Flesch Reading Ease index

4. RESULTS AND ANALYSIS

4.1 Document-Level Readability

Table 1 presents the rank orderings imposed on the set of 10 clinical documents (here labelled A-J) using normalised deontic frequency, latent semantic analysis and the Flesch reading ease index. We evaluated this raw data using a non-parametric (distribution-free) statistic known as the Kendall tau (τ) rank correlation coefficient [24]. This coefficient is defined in the following way:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{4}$$

where

n_c is the number of pairs that agree

n_d is the number of pairs that disagree

This coefficient produces a number in the interval [-1,1] with 1 indicating complete agreement in the rankings and -1 indicating complete disagreement. A result of 0 is obtained when the rankings are independent.

Reviewing our research question (*Is there is significant positive correlation between the deontic density of a document and its overall readability?*) it is useful to clarify what we were actually testing in this experiment. We were attempting to correlate the scores generated by two automatic means for assessing cohesion/readability (LSA and FLESCH) with the concentration of recommendations in the text. Therefore, for each of the readability measures, we needed to test the following null hypothesis H_0 and alternative hypothesis H_1:

H_0 There is no relationship between deontic density and the readability measure at the document level (i.e. $\tau = 0$).

H_1 There is a significant positive correlation between deontic density and the readability measure at the document level (i.e. $\tau \neq 0$).

Interpreting a τ value involves converting the coefficient into a two-sided p-value (a measure of the statistical significance of the result). Using the (standard) alpha (α) level of 0.05 (where α is the probability of making a type I 'false positive' error), the null hypothesis (H_0) can be rejected (and therefore H_1 accepted) where $p - value < \alpha$.

Table 2 shows the Kendall tau correlation coefficient for each pair of rank orderings. Unfortunately, the rankings imposed by normalised deontic frequency could not be meaningfully correlated with the rankings imposed by LSA-cohesion ($\tau = 0.11, p - value = 0.72$). We suspect that this lack of correlation is due to the *extreme smoothing effect* implicit in the LSA-cohesion measure applied to large passages of text. In practical terms this smoothing effect robs the measure of discriminatory power at this level of granularity.

However, when we compared the deontic scores with the Flesch Reading Ease index (calibrated for French texts) we calculated a tau coefficient of 0.55. This exceeds the critical value for Kendall tau when using small samples (0.511

when N=10, two tailed test at 5%) and has a p-value of 0.031. Therefore, with respect to a traditional readability metric, we were able to reject the null hypothesis H_0 when considering the document as a whole (see Figure 2 for an illustration).

The experiment was encouraging. We decided to see if this positive correlation persisted when finer grained sections of documents were examined (see below).

4.2 Paragraph-Level Readability

In this experiment we wanted to see if the deontic density of paragraph-sized chunks of text could be correlated with any (or all) of three different measures of document readability (i.e. Flesch reading ease, LSA-cohesion and a panel of human experts). Therefore, for each of these three readability measures, we needed to test the following null hypothesis H_0 and alternative hypothesis H_1:

H_0 There is no relationship between deontic density and the readability measure at the paragraph level (i.e. $\tau = 0$).

H_1 There is a significant positive correlation between deontic density and the readability measure at the paragraph level (i.e. $\tau \neq 0$).

Table 3 presents the raw data for each of the 40 sampled paragraphs. We analysed this data using the Kendall tau (τ) rank correlation coefficient, as described above. As shown by Table 4, the initial results of our analysis were initially quite disappointing - we could not find a statistically significant correlation between deontic density and any of the three different approaches to measuring the readability/cohesion of a paragraph.

In an effort to discover the root cause of this negative result, we re-examined the characteristics of the 40 sampled paragraphs. One striking feature of these paragraphs was the extremely high incidence of protracted bullet point structures. To determine the confounding effect of these structures on the two automatic readability measures, we removed all of the paragraphs whose text was exclusively contained in these extended lists. This left us with a subsample of 15 slightly more 'conventional' paragraphs (selected paragraphs shown in bold on Table 3).

We then re-calculated the tau coefficient for all three sets of readability scores (see Table 5). All of the coefficients improved (i.e. moved closer towards the anticipated correlation with NDF), but the change for one particular result was pronounced. The correlation between normalised deontic density and LSA-cohesion jumped from its mediocre 40-paragraph value of $\tau = 0.18$ (p-value=0.13) to a highly significant $\tau = 0.50$ (p-value=0.013) (see Figure 3 for an illustration). Thus, we were able to reject the null hypothesis H_0 and accept the alternative hypothesis H_1 by accepting a restricted definition of the term *paragraph* which excluded chunks of text made up exclusively of lengthy enumerative structures.

5. DISCUSSION

Overall, we think that the results presented above are persuasive. In two separate experiments we have demonstrated that there *is* a relationship between the distribution of recommendations within a clinical guideline and its overall readability. However, there are still a number of inconsistencies which need further investigation, as described below.

5.1 Segmentation Issues

The first inconsistency relates to our use of 'old' and 'new' readability measures. In the experiment examining document-level readability, deontic density correlated strongly with the Flesch reading index, but not the measure utilising LSA-cohesion. In the second experiment, which scored paragraph-level readability, the exact opposite was true. This was an entirely unexpected result. Before we ran the experiment, we anticipated a close relationship between LSA-cohesion and deontic density at all levels of granularity. We decided to include Flesch in the analysis purely to illustrate the weakness of 'old' readability metrics when applied to expert level documents. We were therefore extremely surprised when Flesch outscored LSA-coherence at the document level as a predictor for deontic density.

Given the above, the obvious question is this - why does deontic density mimic a 'traditional' readability metric when guidelines are considered as a whole, and a measure based on LSA-cohesion when paragraph-level segmentation is applied. A follow up experiment using a more comprehensive range of text segments could be used to scrutinise this rather puzzling result in more detail.

5.2 Paragraph Extraction

As documented above, our second experiment centred on paragraph-level readability. This experiment proved problematic. In the absence of a pilcrow, deciding what does and what does not constitute a 'paragraph' can be a fairly subjective exercise. Lacking an objective definition, we were forced to improvise rather arbitrary criteria for their identification and extraction (see §3.3). Subsequently, we were forced to adjust our criteria to exclude 'abnormal' paragraphs consisting almost exclusively of bullet points, thereby reducing the set of paragraphs under consideration to a somewhat disappointing 15. This was an obvious weakness in our experimental design. If the second experiment were repeated today, it is likely that we would use randomly selected text segments of *fixed size* (e.g. 300 words) as an alternative.

5.3 Semantic Spaces and Readability Panels

As discussed above, the semantic space used during these experiments was constructed using resources developed as a teaching reference for 2nd and 3rd year medical students. This space was then used to score the similarity of adjacent sentences extracted from the medical guidelines. An obvious criticism, which we concede, would target the mis-alignment between the semantic space (intended for students) and the guidelines (which are intended for practitioners). Use of the ALPESMED medical corpus, it could be argued, results in a semantic space more heterogeneous than appropriate for the task of analysing clinical documents. A follow up experiment could be employed to examine the confounding effect of the semantic space. This experiment could contrast the LSA-coherence scores achieved using three different semantic spaces as follows:

1. A space built using non-expert, non-medical documents on diverse topics (e.g. sampled from fr.wikipedia.org).
2. A space built using ALPESMED, as described above (e.g. intermediate level, medical reference material).
3. A space built using a large collection (200+) of expert level medical guidelines, contributed by HAS.

	LSA	NDF	FRE
LSA	-	-0.11 (p-value=0.72)	-0.38 (p-value=0.15)
NDF	-0.11 (p-value=0.72)	-	**0.55 (p-value=0.03)**
FRE	-0.38 (p-value=0.15)	**0.55 (p-value=0.03)**	-

Table 2: Kendall tau (τ) correlation coefficients (document-level readability)

PARA	NDF	LSA	FRE	HUM	PARA	NDF	LSA	FRE	HUM
1	**0.00**	**0.51**	**0.47**	**0.21**	**21**	**0.77**	**-0.01**	**0.46**	**0.22**
2	0.92	-0.02	0.00	0.24	22	1.00	-0.03	0.27	0.1
3	0.00	-0.02	0.49	0.09	23	0.92	0.52	0.00	0.11
4	1.00	0.36	0.00	0.13	**24**	**0.33**	**-0.04**	**0.41**	**0.23**
5	0.25	-0.03	0.41	0.10	**25**	**1.00**	**0.84**	**0.25**	**0.11**
6	**1.00**	**0.52**	**0.5**	**0.13**	26	0.23	0.31	0.03	0.12
7	1.00	-0.06	0.36	0.09	27	0.43	-0.05	0.39	0.14
8	**0.82**	**0.40**	**0.50**	**0.11**	28	0.62	-0.01	0.04	0.17
9	**0.73**	**-0.12**	**0.43**	**0.14**	29	0.00	0.75	0.28	0.16
10	1.00	-0.04	0.00	0.11	**30**	**0.59**	**0.37**	**0.50**	**0.18**
11	**0.37**	**-0.09**	**0.28**	**0.14**	31	0.15	-0.04	0.30	0.18
12	1.00	0.52	0.51	0.22	**32**	**0.00**	**-0.20**	**0.46**	**0.1**
13	0.87	0.55	0.00	0.09	**33**	**0.38**	**0.39**	**0.27**	**0.12**
14	**0.49**	**-0.01**	**0.61**	**0.17**	**34**	**0.72**	**0.30**	**0.21**	**0.12**
15	0.72	-0.06	0.31	0.19	35	0.92	-0.02	0.00	0.12
16	0.57	0.38	0.00	0.16	36	0.68	0.40	0.27	0.17
17	1.00	-0.01	0.00	0.11	37	0.54	0.30	0.37	0.11
18	**1.00**	**0.61**	**0.43**	**0.12**	38	0.85	0.49	0.70	0.14
19	1.00	-0.06	0.24	0.12	39	0.58	0.36	0.31	0.09
20	**0.15**	**-0.05**	**0.51**	**0.14**	40	1.00	0.49	0.00	0.2

Table 3: Readability of 40 paragraphs according to normalised deontic frequency, LSA Flesch Reading Ease index and human experts (sub-sample shown in bold)

	NDF	LSA	FRE	HUM
NDF	-	0.18 (p-value=0.13)	-0.23 (p-value=0.05)	-0.10 (p-value=0.41)
LSA	0.18 (p-value=0.13)	-	-0.04 (p-value=0.72)	0.03 (p-value=0.78)
FRE	-0.23 (p-value=0.05)	-0.04 (p-value=0.72)	-	0.11 (p-value=0.34)
HUM	-0.10 (p-value=0.41)	0.03 (p-value=0.78)	0.11 (p-value=0.34)	-

Table 4: Kendall tau (τ) correlation coefficients (paragraph-level readability)

	NDF	LSA	FRE	HUM
NDF	-	**0.50 (p-value=0.013)**	-0.12 (p-value=0.58)	-0.24 (p-value=0.25)
LSA	**0.50 (p-value=0.013)**	-	-0.07 (p-value=0.72)	0.16 (p-value=0.45)
FRE	-0.12 (p-value=0.58)	-0.07 (p-value=0.72)	-	0.17 (p-value=0.42)
HUM	-0.24 (p-value=0.25)	0.16 (p-value=0.45)	0.17 (p-value=0.42)	-

Table 5: Kendall tau (τ) correlation coefficients (paragraph-level readability with bullet-point filter)

Further variations on a theme could be introduced by employing an additional readability panel. Our experiment employed a readability panel composed of HAS employees (i.e. experts in the production of clinical guidelines). A better experimental design would feature a second readability panel composed of medical practitioners and students (i.e. the actual target audience for the guidelines). Such a panel may have scored the clinical guidelines quite differently to the HAS experts.

5.4 LSA, Readability and Quality

One interesting feature of the second experiment was the lack of correlation between the human judgements and the LSA cohesion scores. Previous studies have shown that LSA provides an adequate reflection of human knowledge given a suitable semantic space (see §2.4). However, in our study, the tau coefficient expressing the correlation between the findings of the readability panel and the LSA-cohesion scores never approached a significant value.

A post-mortem of this rather surprising result suggests that the problem may lie in a mismatch between the related concepts of 'readability' and 'quality'. The 5 Likert statements in our questionnaire focused on textual readability. Respondents were asked to comment on the syntactic and lexical qualities of the various text passages rather than their quality, usability or fitness for purpose. This meant that low quality text passages (i.e. passages containing vague decision rules) were receiving good marks from the panel simply because they were easy to read.

In contrast, latent semantic analysis seems more attuned to deeper issues of document quality. We made this discovery when carrying out a small follow-up experiment. Previous work has established that clinical recommendations should include a 'condition' part and an 'action' part if they are to be useful to medical practitioners [15]. For instance, 'Les antipsychotiques peuvent être envisagés en cas de symptômes délirants ou hallucinatoires. (*Antipsychotic drugs can be considered in case of symptoms of delirium or hallucinations*). One feature of the G-DEE document engineering platform enables the automatic recognition of these conditional statements (either in the *front-scope* or the *back-scope* of a recommendation) (see Figure 4). We used this feature to calculate the percentage of conditional statements within recommendations for all 40 of the sampled paragraphs. We found that there was a statistically significant difference between the arithmetic mean we recorded for the high LSA-coherence group (41.8%) and the arithmetic mean scored by the low LSA-coherence group (14.3%). We regard this finding as important. The fact that semantic measures of document cohesion tend to agree with pragmatic measures of quality (based on the rhetorical aspects of recommendations) is extremely interesting, and clearly warrants further attention.

5.5 Practical Applications

The experiments described in this paper have established that there is a relationship between the density of deontic structures in French clinical documents and their overall readability. This finding has an immediate practical application. A readability measure based on deontic density, once perfected, can be integrated into the G-DEE document engineering platform and used to improve the quality of clinical guidelines produced by the HAS. However, beyond this worthy (but rather narrow) scope, are there any other applications for this research? We hope that there are. All prescriptive documents (whatever the language, whatever the topic) will contain deontic structures. Once a (language-specific) finite state transition network (FSTN) identifying these structures is built (a fairly trivial exercise), these documents can be analysed in exactly the same way as the clinical documents described above. We could test this theory quite easily e.g. by scoring a corpus of English clinical documents.

6. CONCLUSION

The purpose of this study was to confirm an intuition common amongst users of the G-DEE document engineering platform that high quality clinical guidelines are authored *around* the structure imposed by recommendations. The results discussed above, which show a statistically significant correlation between the density of recommendations (at both document and paragraph level) and overall readability, provides empirical support for this intuition. The next step involves using this information in an operational context. We plan to develop a lightweight plug-in for G-DEE that tracks normalised deontic frequency throughout the enumeration of a clinical guideline. This plug-in will provide a real-time, impartial assessment of the quality/readability of a guideline that can be used to steer the iterative authoring process.

Further work in this area could continue to explore the tenuous relationship between document quality (as assessed by end users) and document quality (according to various pragmatic, semantic and syntactic measurements). As mentioned above, a document which scores well on such tests is not necessarily a high quality document. A document can be easy to read and comprehend, but *still* utterly unfit for the purpose it was intended. Bridging the gap (between actual quality and its computational proxies) remains a grand challenge.

7. ACKNOWLEDGEMENTS

The authors would like to thank all of the people at la Haute Autorité de Santé who participated in the readability study - Joelle Andre-Vert, Muriel Dhenain, Michel Laurence, Estelle Lavie, Valérie Lindecker, Emmanuel Nouyrigat, Christine Revel Delhom and Pierre Durieux (Université Paris Descartes / Hopital Européen Georges Pompidou. Also, many thanks to Michel Laurence and Martine Franco (HAS) for their help with the survey software (GRaAL) and Stefan Darmoni (Medical Informatics and CISMeF team, Rouen University Hospital). We would also like to thank the anonymous referees for their suggestions on how to improve the paper. This research was partially supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at University of Dublin, Trinity College.

8. REFERENCES

[1] http://math.nist.gov/javanumerics/jama/.

[2] S. M. Aluísio, L. Specia, T. A. Pardo, E. G. Maziero, and R. P. Fortes. Towards brazilian portuguese automatic text simplification systems. In *DocEng '08: Proceeding of the eighth ACM symposium on Document engineering*, pages 240–248, New York, NY, USA, 2008. ACM.

[3] B. K. Britton and S. Gulgoz. Using kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology*, 83:329–45, 1991.

[4] A. M. Buckeridge and R. F. E. Sutcliffe. Disambiguating noun compounds with latent semantic indexing. In *International Conference On Computational Linguistics*, pages 1–7, 2002.

[5] M. Coleman and T. L. Liau. A computer readability formula designed for machine scoring. *Applied Psychology*, 60:283–284, 1975.

[6] S. A. Crossley, D. F. Dufty, P. M. McCarthy, and D. S. McNamara. Toward a new readability: A mixed model approach. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, 2007.

[7] E. Dale and J. Chall. A formula for predicting readability. *Educational Research Bulletin*, 27:11–20, 1948.

[8] A. Davison and R. N. Kantor. On the failure of readability formulas to define readable text: A case study from adaptations. *Reading Research Quarterly*, 17:187–202, 1982.

[9] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[10] R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233, 1947.

[11] P. W. Foltz, W. Kintsch, T. K. Landauer, and K. L. Thomas. The measurement of textual coherence with latent semantic analysis, 1998.

[12] G. Georg, H. Hernault, M. Cavazza, H. Prendinger, and M. Ishizuka. Analysing clinical guidelines' contents with deontic and rhetorical structures. *Lecture Notes in Artificial Intelligence*, pages 86–90, 2009.

[13] G. Georg, H. Hernault, M. Cavazza, H. Prendinger, and M. Ishizuka. From rhetorical structures to document structure: shallow pragmatic analysis for document engineering. In *DocEng '09: Proceedings of the 9th ACM symposium on Document engineering*, pages 185–192, New York, NY, USA, 2009. ACM.

[14] G. Georg and J. Marie-Christine. A document engineering environment for clinical guidelines. In *DocEng 07: Proceedings of the 2007 ACM symposium on Document engineering*, pages 69–78. ACM, 2007.

[15] G. Georg, B. Séroussi, and J. Bouaud. Does gem-encoding clinical practice guidelines improve the quality of knowledge bases? a study with the rule-based formalism. *Proceedings of the American Medical Informatics Association*, pages 254–258, 2003.

[16] A. C. Graesser, D. McNamara, M. Louwerse, and Z. Cai. Coh-metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers*, 36(2):193–202, 2004.

[17] A. C. Graesser, D. S. McNamara, and M. M. Louwerse. What do readers need to learn in order to process coherence relations in narrative and expository text. In *Rethinking reading comprehension*, pages 82–98. New York: Guilford Publications, 2003.

[18] A. C. Graesser, M. Singer, and T. Trabasso. Constructing inferences during narrative text comprehension. *Psychological Review*, 101:371–395, 1994.

[19] A. C. Graesser, P. Wiemer-Hastings, K. Wiemer-Hastings, D. Harter, and T. T. R. Group. Using latent semantic analysis to evaluate the contributions of students in autotutor. *Interactive Learning Environments*, 8(2):129–147, 2000.

[20] M. A. K. Halliday and R. Hasan. *Cohesion in English*. Longman, 1976.

[21] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1-2), 2001.

[22] http://www-sante.ujf grenoble.fr/SANTE/alpesmed/corpus.htm).

[23] L. Kandel and A. Moles. Application de l'indice de flesch a la langue francais. *Cahiers d'Etudes de Radio-Television*, 19:253–274, 1958.

[24] M. G. Kendall. *Rank Correlation Methods*. Griffin, 1976.

[25] W. Kintsch. The use of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95:163–182, 1988.

[26] W. Kintsch and D. Vipon. Reading comprehension and readability in educational practice and psychological theory. In *Perspectives on Memory Research*. Lawrence Erlbaum, 1979.

[27] G. Klare. Readability. In P. D. Pearson, R. Barr, M. L. Kamil, and P. Mosenthal, editors, *The Handbook of Reading Research*, pages 681–731. New York Longman, 1984.

[28] T. K. Landauer, D. Laham, B. B. Rehder, and M. Schreiner. How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pages 412–417, 1997.

[29] B. A. Lively and S. L. Pressey. A method for measuring the vocabulary burden of textbooks. *Educational Administration and Supervision*, 9:389–398, 1923.

[30] P. Nakov, A. Popova, and P. Mateev. Weight functions impact on lsa performance. In *EuroConference RANLP'2001 (Recent Advances in NLP*, pages 187–193, 2001.

[31] F. Rastier, M. Cavazza, and A. Abeille. Semantics for descriptions: From linguistics to computer science. *CLSI*, 2002.

[32] J. Redish. Readability formulas have even more limitations than klare discusses. *ACM J. Comput. Doc.*, 24(3):132–137, 2000.

[33] J. Savoy. A stemming procedure and stopword list for general french corpora. *J. Am. Soc. Inf. Sci.*, 50(10):944–952, 1999.

[34] T. A. van Dijk and W. Kintsch. *Strategies of Discourse Comprehension*. New York: Academic Press, 1983.

[35] M. Vogel and C. Vogel. An objective method of determining grade placement of children's reading material. *Elementary School Journal*, 28:373–381, 1928.