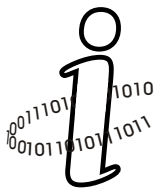
Faculty of Mathematics, Physics, and Informatics
Comenius University, Bratislava

Towards Advanced Text Quality Evaluation Modelling

Krištof Anetta, Martin Takáč

TR-2013-036



Technical Reports in Informatics

Towards Advanced Text Quality Evaluation Modelling

Kristof Anetta
University of Vienna
Universitaetsring 1, 1010 Vienna, Austria

Martin Takac
Centre for Cognitive Science, Faculty of Mathematics, Physics and Informatics
Comenius University in Bratislava
Mlynska dolina, 84248 Bratislava, Slovakia

Abstract

This paper tackles the phenomenon of judging text quality from the point of view of AI and cognitive science. Building on relevant literature, it starts with a general view of current computer text quality evaluation and then, discarding the fields where sufficient amount of research is taking place, narrows its focus down to questions of advanced text quality evaluation (literary style in particular). It suggests solutions on several levels, starting with subsemantic neural network evaluation and going through possibly implementable rules from writers‘ and editors’ practice both on the subsemantic and semantic level, extracted by means of content analysis out of a considerable corpus of mostly informal Internet articles.

1 RATIONALE

On the battlefield where humans and computers measure their skills against each other, the domain of judging text quality is, despite heavy losses of humans in many other fields of research, still deep in the human territory. And given the immense complexity of the task, mastery of which usually follows at least ten years of formal education of the human agent, it will most probably remain so for quite some time, unless an unexpected breakthrough in AI should occur. On the continuum of different levels of text quality, ranging from the more or less mechanical aspect of correct grammar and syntax to highly abstract criteria such as ideological relevance or intellectual coherence, computers did not record much success in judgements beyond the level of correct grammar and syntax, except for a few attempts at

extrapolating word-level statistical data to form "judgements" on sentence and text level, for example the latent semantic analysis (Landauer, Foltz and Laham, 1998) or the coherence model of Louis and Nenkova (Louis and Nenkova, 2012).

1.1 Text quality = readability?

While there are many possible ways of defining the concept of text quality, there is an important distinction to be made in order to specify the aim of this paper. Correct grammar and syntax, together with semantic coherence, are the necessary requirements for any text to be readable and understandable by a human agent. But this level of **readability**, necessary though it is, is often used as the only level of computer-modelled text quality evaluation.

Once we assume that the text is readable and want to apply methods of evaluation which are more subtle and discerning in the **more advanced aspects** of text quality (e.g. style, originality), current research ceases to present a coherent picture and leaves us with rather disconnected fragments of research. In practice, state-of-the-art computers are capable of assessing the grammatical and syntactical aspects of, say, student essays, but once these reach a certain degree of readability, they can no longer rank them based on their quality: if attempted, the results would be most probably different for each model and certainly not satisfactory. It appears that there is an implicit consensus that the frontier of computer-modelled text quality evaluation did not yet leave the level of readability. This paper will examine the question to what extent the advanced aspects of text quality might be reliably captured with state-of-the-art knowledge of computer-modelled text quality evaluation.

1.2 Where is the goal? The ultimate text evaluation practice

Let us have a short look at several titles of important works of the most advanced human text quality evaluation practice, literary criticism, published in the 20th century. For example, there were *Blurred Genres: The Refiguration of Social Thought* by Clifford Geertz, *Anti-Oedipus: Capitalism and Schizophrenia* by Gilles Deleuze and Félix Guattari, *The Heresy of Paraphrase* and *Irony as a Principle of Structure* by Cleanth Brooks and *The Sacrificial Crisis* by René Girard. Indeed, it does seem that the elusive aesthetic, cultural and intellectual aspects of literature considered by literary critics are not to be evaluated by computers in the

near future – if anything because they require human-level semantic networks and text understanding coupled with scholar-level knowledge. What is the point of this diversion? The point is to demonstrate that from a certain level upwards, text quality evaluation is an important constituent of the process of human self-understanding – we need it to make sense of our embodied existence in the world, be it revealing human aesthetic preferences or judging political and social theories. The fact that text evaluation has such meta-cognitive ends both ennobles the purpose of trying to model it and reminds us that the journey to perfection will probably take longer than expected.

2 MEASURING ADVANCED ASPECTS OF TEXT QUALITY

It is important to be aware of the fact that there is no established norm of text quality beyond grammar and syntax (and hopefully never will be). What this paper considers is an arbitrary selection of criteria based partly on common sense and partly on the notions of editorial practice collected from various sources.

Karen Schriver, even though she probably did not expect her work to occupy such a prominent position in a purely computational endeavour, outlined a highly relevant set of questions:

- What are the characteristics of an effective text?
- Can we agree on a working definition of text quality?
- What are the key skills and abilities involved in text evaluation?
- What do experienced evaluators do that inexperienced evaluators do not?
- What do writers learn from repeated experience in judging text quality?
- How can we improve evaluators' abilities to judge the quality of text?
- What are the tradeoffs associated with particular methods for judging text quality? What methods produce reliable and valid judgments?
- What aspects of text evaluation can we automate using the computer?
- How can the computer help reduce the burden of text evaluation?

(Schriver, 1989, p. 238)

Hayes et al. (1987) provided an interesting framework which attempts to serve as the basis to approaching text quality on any level:

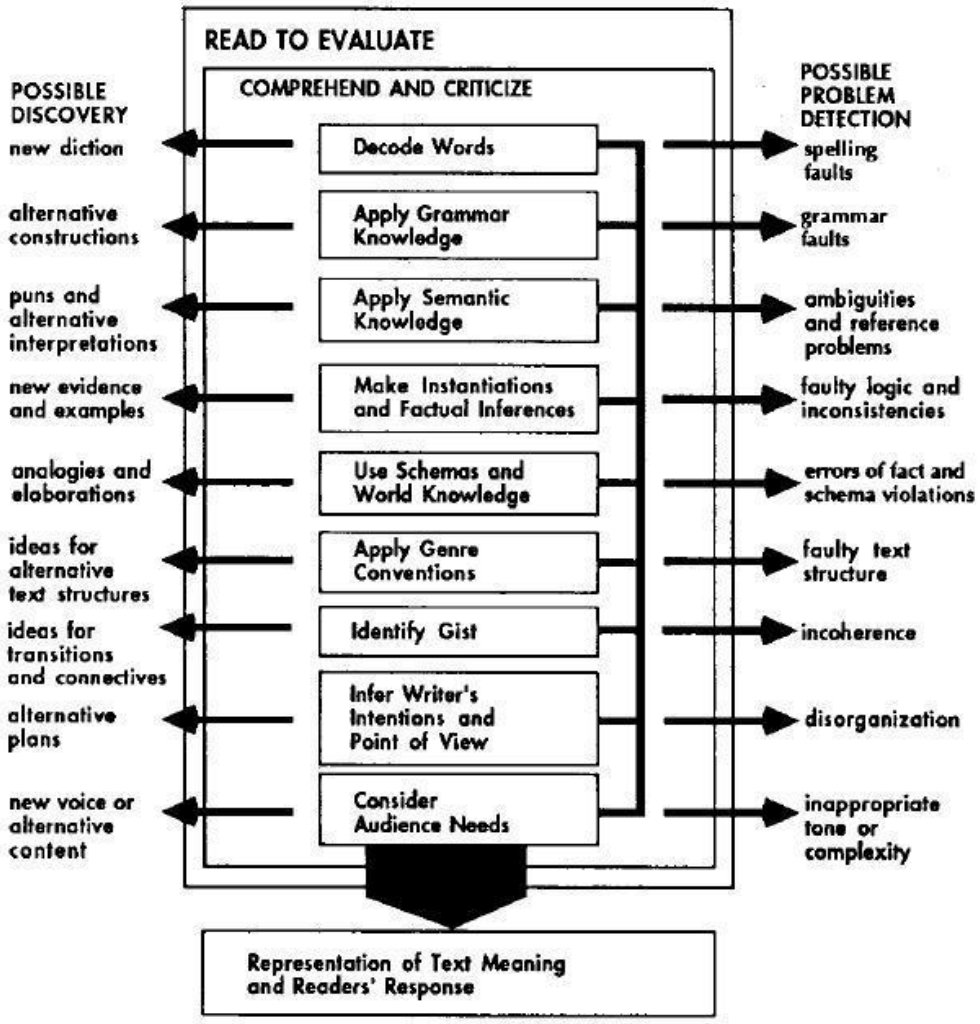


Figure 1: Schema of general text evaluation according to Hayes (1987)

In order to achieve a narrower focus, let us discard the levels of this diagram which are sufficiently covered by current research. Obviously, the levels *Decode Words* and *Apply Grammar Knowledge* have been given proper consideration, as seen for example in Microsoft Word's spelling and grammar checker. Similarly, the level named *Identify Gist* was addressed with the Latent Semantic Analysis (LSA) by Landauer, Foltz and Laham (1998) and the level *Consider Audience Needs* was extensively studied and reviewed in papers such as Paukkeri et al. (2013).

The remaining levels form two groups with regard to their computer requirements. Levels *Apply Semantic Knowledge*, *Make Instantiations and Factual Inferences*, *Use Schemas and World Knowledge*, and *Infer Writer's Intentions and Point of View* seem to be fully dependent

on the computer's ability to map the structures in the text onto a knowledge base. Because the question of semantics in natural language processing is an enormously large and complex field, state-of-the-art text quality evaluation based on semantic information is still very far from being able to evaluate advanced aspects of text quality. Therefore, this paper will not deal with these levels for the time being.

The rather inconspicuous level, *Apply Genre Conventions*, forms the second group. Although it does depend on the knowledge base to some extent, the word "genre" leaves us with a large, perhaps even unintentionally large space of interpretation. To see "faulty text structure" as the major possible problem on this level (as seen in the diagram) is a rather narrow view of genre and is probably caused by a utilitarian approach to the effectiveness of a text. If the term "genre" was taken with all its meanings, it would create a complex set of criteria which would judge almost anything from vocabulary to large plot components, either checking correct genre usage or establishing quality by the absence of genre limitations.

3 JUDGING LITERARY STYLE

After we have abandoned the ambition of revolutionising knowledge-based computer text understanding, we can focus on finding the computationally tractable components of advanced text quality evaluation. As individual words are rarely judged by human evaluators and whole texts have insatiable computational power requirements, this paper presupposes that **the central level for advanced text quality evaluation is the sentence level**. Similarly, Nenkova et al. (2010) writes that "Sentence structure is considered to be an important component of the overall linguistic quality of text" (p. 222). This is convergent with widespread notions of literary style as the ability to form eloquent and original sentences – one of the consequences being that quotations of famous people are so popular, as opposed to mere individual "words of famous people" (it is true that Shakespeare is known for many neologisms, but the general consensus is that vocabulary alone has little explanatory power - some of the best texts were written using only a standard set of words). It is the syntactical coupling which is able to create an infinite number of meaningful sentences out of a rather finite set of (several tens of thousands of) words.

Literary style can serve as a good example of an advanced aspect of text quality, as it combines being located mostly on sentence level and being advanced enough to require considerable human reading skills.

The question is, for example, how could these opening words of Vladimir Nabokov's *Lolita*:

"The tip of the tongue taking a trip of three steps down the palate to tap, at three, on the teeth. Lo. Lee. Ta."

be computationally distinguishable from a clichéd line such as:

"You are my heart and soul and I will love you forever and always."

as having a more original and more literary prose style.

4 DESIGNING THE EVALUATION MODEL

James Joyce's 1922 novel *Ulysses* and Stephanie Meyer's 2005 novel *Twilight* belong to completely different categories of literature – even if we do not consider *Twilight* outright faulty (many do), it certainly is a young-adult fantasy romance with all the ensuing limitations, while *Ulysses* is a revolutionary work of art, highly respected both for richness of style and innovative narrative. Both novels are grammatically and syntactically correct, but in the reading of an experienced evaluator, there should be a significant difference in quality as low as on the sentence level (note that this is an authorial assumption lacking reference, based on general knowledge of the literary canon).

Q1: Is it possible to capture the difference computationally, without semantic information?

Q2: If the answer is yes, is the result generalizable? Is it possible to use it to categorize other texts with a success rate beyond chance?

4.1 Straightforward neural network approach

The basic step to take is to train a neural network to differentiate between sentences from both books and then, without any further input, test the network with sentences from other works by Joyce and Meyer to see if the network acquired the ability to differentiate between

the styles of the authors. In the case of successful categorisation, works of other authors of both quality and allegedly non-quality fiction can be tested in order to determine to what extent is the difference separable from the respective writers and generalizable to the attribute of "literariness". If the generalization does not prove successful beyond chance, a possible elaboration of this step is to train the network on ten quality and ten non-quality novels written by authors of varied prose styles.

4.1.1 Rationale behind using a neural network

It is important to be aware of the fact that since we do not believe that an objective measure of the quality of texts is possible, our suggestion is not designed to identify any objective attributes. The basis of the neural network approach is to teach the network using already classified samples and let it generalize – in the same way as a student of literature, a neural network needs to be presented examples of quality literature and neutral literature to be able to generalize and distinguish between previously unseen texts. It can be argued that any arbitrary selection of learning material is inherently subjective, but this argument can be successfully countered by pointing out the way literary canon is taught in schools – indicating that the scientifically desired objective background of text selection (for example statistical evaluation of free-market development) is not to be found in human literary practice in the first place.

Simply put, people usually do not discover great literary style by themselves – they rather learn to appreciate it over time. In this paper, we suggest a number of ways to find the minimum requirements for being able to do so.

4.1.2 The neural network

Taking into account the temporal dimension of language, we propose that the suitable neural network architecture is the **recurrent network** as proposed by Elman (1990). An important attribute of recurrent networks is that it represents time implicitly in the form of dynamic memory.

4.2 Suggestions for additional preprocessing

To improve the richness of the neural network's input while still excluding semantic information, this paper proposes a number of major possibilities:

- to include vocabulary information (such as corpus based word frequencies)

- to include syntactic information

- to include corpus-based frequencies of phrases up to a specified length (cliché identification)

Theoretically, these additions abandon the notion that style is distinguishable in sentences as mere strings of characters – namely, they express the need of:

- relating the text to the existing corpora and thereby hypothesizing about the **role of statistical memory and volume of previously read text of the evaluator**

- identifying syntactical structures and thereby hypothesizing about the **link between syntax and style**

4.2.1 Vocabulary richness

The simplest way to compute the "original vocabulary quotient" of a text is to convert its lexemes into rank numbers on a language corpus frequency table and then calculate the average rank per word (eventually excluding function/grammatical/synsemantic words). An available solution commonly used by linguists in similar tasks is the 60,000 lemma frequency list of the *Corpus of Contemporary American English* (Davies) at http://corpus.byu.edu (paid). This particular frequency list also contains frequency information for different genres, thus accounting for differences in text purpose and target audience (for example between scientific articles and newspaper columns).

Laufer and Nation (1995), in their article on lexical richness in second language learners, mention several measures, namely:

- Lexical Originality (LO), the "percentage of words in a given piece of writing that are used by one particular writer and no one else in the group" (p. 309) (in our case, where context is

wider than just a classroom, a group can be viewed as a group of writers of one genre in one historical period). It is calculated as follows:

$$LO = \frac{Number\ of\ tokens\ unique\ to\ one\ writer * 100}{Total\ number\ of\ tokens}$$

- Lexical Density (LD), the "percentage of lexical words in the text" (p. 309). Obviously, this index is useful to see if the density is optimal for the genre, judging against both extremely dense and extremely sparse texts.

$$LD = \frac{Number\ of\ lexical\ tokens * 100}{Total\ number\ of\ tokens}$$

Banerjee and Papageorgiou (2009, p. 6) introduce a different equation for lexical density:

$$LD = \frac{\frac{High\ frequency\ lexical\ words}{2} * Low\ frequency\ lexical\ words}{Grammatical\ words}$$

- Lexical Sophistication (LS), the "percentage of 'advanced' words in the text" (p. 309), while the definition of "advanced" depends on the researcher. This index is also expected to have its genre-specific averages (even inside fiction – science fiction and romance would be diametrically different in this respect)

$$LS = \frac{Number\ of\ advanced\ tokens * 100}{Total\ number\ of\ lexical\ tokens}$$

- Lexical Variation (LV), the "type/token ratio, i.e. the ratio in per cent between the different words in the text and the total number of running words" (p. 310). Authors note that this index is affected by text length and is "unstable for short texts" (p. 310).

$$LV = \frac{Number\ of\ types * 100}{Total\ number\ of\ tokens}$$

Following the outline of these rather simple measures, Laufer and Nation propose to replace them with their own measure: the Lexical Frequency Profile (LFP). It is expressed in the form of four percentages (for example 75%-10%-10%-5%), which stand for, respectively: the 1,000 most frequent words of English; the second 1,000 most frequent words; UWL (University Word List, containing 836 word families not in the 2,000 most frequent words but still frequent in academic texts); and words which are in neither of these lists. Slight

genre-specific modifications (firstly, replacing the UWL with a general frequency-based list) might generate interesting information about lexical profiles of different prose styles. The authors concluded that it provided stable results for texts by the same author and was able to discriminate "between learners of different proficiency levels" (p. 319).

4.2.2 Cliché identification

To begin in the same vein as in the previous chapters, the simplest way to control for clichés would be to seek commonly used n-grams. But since English is an isolating language with many grammatical words (low morpheme-per-word ratio), the list of its most frequent n-grams is likely to contain many syntactical constructions which are not considered clichés. Unless we develop a framework categorizing n-grams with big *n*s based on lexical density, annotation seems to be necessary. There are some existing corpora of clichés that might be utilized, such as the *Cliché Finder* (Friedman), listing 3,300 clichés and *ClichéSite*, listing 2,100 clichés including explanations.

Smith, Zee and Uitdenbogerd, in their project rating song lyrics according to their clichédness, used trigrams and rhyme pairs aided by tf-idf (text frequency – inverse document frequency) weights and correlated these with human judgements.

A potentially strong argument against the use of n-grams and in favour of an annotated cliché database is put forward by Wray (2002), who claims that humans learn languages mostly on the level of ready-made formulaic phrases, thus reserving energy for generating and interpreting of ideas. Banerjee and Papageorgiou use this theory to suggest that formulaic phrases might be a dimension of vocabulary richness (p. 10) and therefore related to text quality. Lists of formulas such as that by Simpson-Vlach and Ellis (academic) might be useful if not for fiction evaluation, then at least for academic texts (more emphasis on systematic and standardized language).

4.2.3 Syntactical cues

Nenkova (2011), in the attempt to find correlates of human text quality ratings, found that neither average number of characters per word (not surprisingly) nor average number of words per sentence correlated with the ratings (p. 11). She also indicates that the positive

correlation between the number of subordinate clauses and text quality was close to significant (p. 11). On the other hand, the average number of verb phrases correlated significantly with text quality.

5 INFORMED BY PRACTICE: WRITERS AND EDITORS ON TEXT QUALITY

Scientists do not feel a need to clarify what exactly makes a good text – any opinion can be expected to seem too subjective to be defendable. However, writers and editors have to think about good and bad texts because their work depends on the very difference of quality this paper tries to address. Even though the rules may appear vague and subjective, in practice they are still better than nothing (as opposed to science). In this section, we will try to extract and operationalize information found in writers' and editors' advice about text quality with the help of both manual and automated methods.

For the automated quantitative content analysis, we chose KHCoder (Higuchi), an open-source program.

5.1 Content analysis of William Strunk's *Elements of Style*

Although the critical stance towards Strunk's famous 1918 prescriptive guide to style in English is not unequivocally positive, it is arguably the most influential collection of writing advice for the English language (and therefore for the Western canon regardless of language as well), resonating to this day in writing classes. In this part, we will try to interpret the results of quantitative analyses of the original 1918 edition.

The first step was to draw a co-occurrence network, omitting grammatical words, setting the minimum term frequency to 20 and using the settings "Thicker lines for stronger edges" and "Larger nodes for higher frequency words".

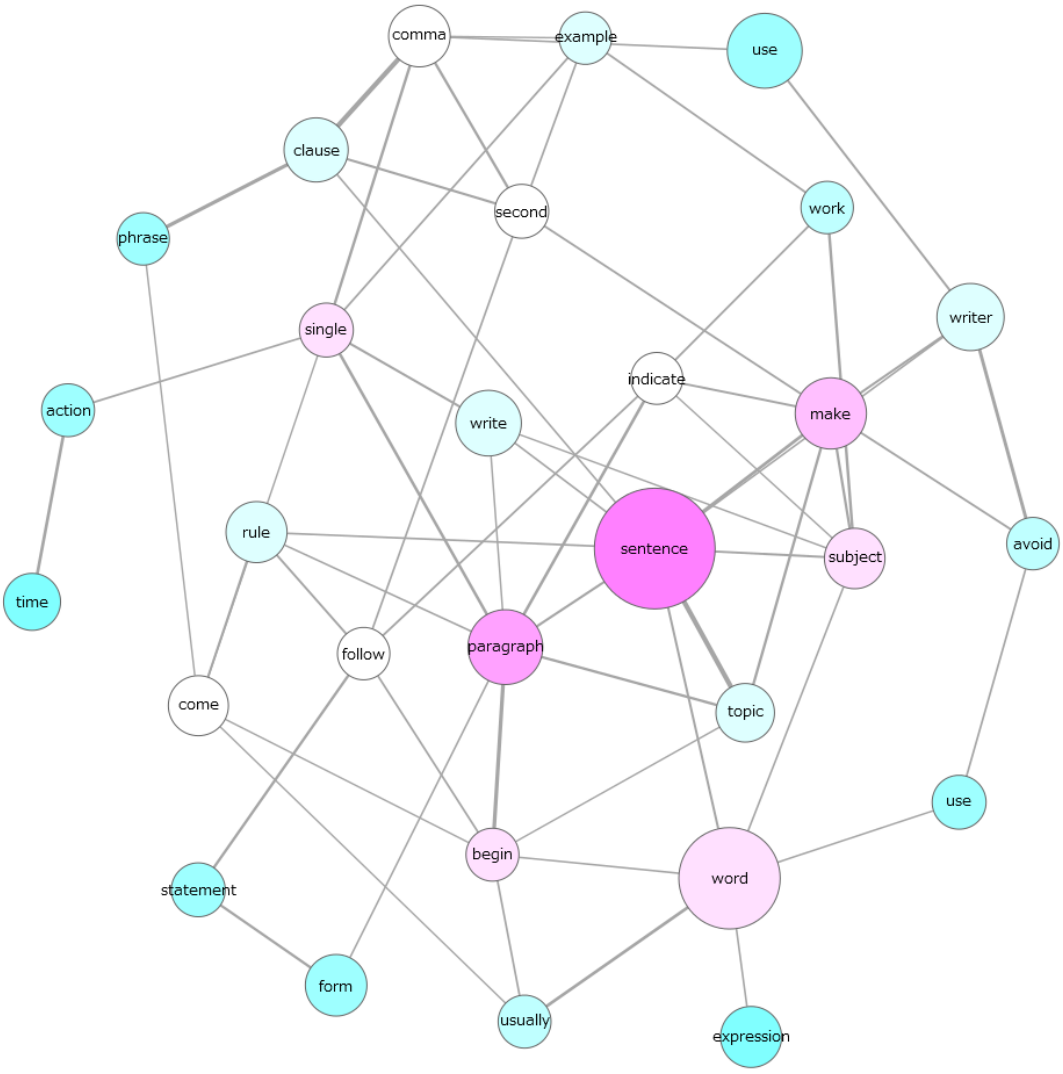


Figure 2: Word co-occurrence network of the text of Strunk (1918) coloured according to degree of centrality

This basic map of a small number of central words does not reveal much about the finer aspects of the text. However, it clearly highlights its focal points. Let us take the four different levels on which text quality can be observed: word, sentence, paragraph and text level. This network indicates that Strunk found sentence level to be the most important (we presuppose this in Chapter 3 independently of this finding). Following are the paragraph and word levels, "word" being more numerous and "paragraph" being more central. It is important to note the absence of the text level or any noun which could represent it – this

indicates that Strunk's advice targets only the constituents which stand as a part in a part-whole relationship (to the sentence, paragraph, text).

5.2 A comparative content analysis of Internet writing advice

A quick Google survey of writing advice on the Internet shows us a variety of different perspectives on writing. However, the majority of these fall into two categories:

- advice regarding the process of writing or writer's state of mind, unrelated to the actual text (such as "do not correct yourself while writing the first draft" or "be sincere")

- common mechanical writing problems ranging from grammar to syntax which are often difficult to search for algorithmically (such as apostrophe mistakes or misplaced modifiers)

Apart from these, a small third category can be identified – the category of comparatively implementable, concrete advice on the word, sentence and paragraph levels of writing. We collected 20 texts falling into this category to form a small corpus referenced in the part "Websites used for content analysis" just before "References".

5.2.1 Manual content analysis

In this part, we will go through implementable pieces of advice from the small corpus of 20 Internet articles and we will highlight the degree of overall agreement/disagreement in the particular questions and slight differences in formulation between multiple wordings of the same principle.

Use of lexical categories

A number of articles target a specific lexical category and its effect on text quality. Nouns and verbs are promoted as the major constituents of good style in [1] ("write with nouns and verbs"). [14] elaborates on this by claiming that nouns and verbs should be "strong", not needing adverbs and adjectives (run quickly => sprint). [7] is the only one to specify that verbs should be used at the expense of nouns ("use less nouns and more verbs").

Use of adverbs and adjectives is mostly discouraged. [14] warns against both. [1] quotes Mark Twain saying "when you catch adjectives kill most of them". [11] and [17] similarly discourage the use of adjectives. [6], being wholly dedicated to its title "Abolish the Adverbs", quotes Stephen King's "the road to hell is paved with adverbs". Although [7], [8], and [18] do not mention any lexical category, they specifically target the most widely used adverbs.

With respect to particular contexts and genres, [18] recommends avoiding the first person pronoun in formal writing and [15] suggests creatively replacing third person pronouns in fiction with nouns.

Synthesis: Since it is easy to computationally capture lexical categories by content analysis software, we suggest calculating the ratios of the major categories (noun, verb, adjective, adverb) in established, quality literature and see whether the results converge towards an optimum (an overall optimum, a genre optimum or one for smaller groups of texts).

Use of semantically labelled words

Commonly known as "show, don't tell" (mentioned in [4], [6], [14], [15], and [20]), this advice translates into the distinction between abstract and concrete words. To specify what is "concrete", [4] encourages description of outward appearance instead of inner states, [7] uses words "tangible", "imagery", and "motion", [10] mentions "stories" and "examples", [14] and [15] agree on the importance of conveying "physical sensation". [20] calls the quality "grounded in the real world". [8] uses a quote from Anton Chekhov saying "don't tell me the moon is shining, show me the glint of light on broken glass".

Synthesis: We are not aware of any dictionary which is consistently annotated based on the degree of abstraction of the word, or its sensory content. However, should such a database exist, calculating the "concreteness quotient" of a text might be a strong predictor of literary quality (given that necessary preconditions of readability and coherence are satisfied).

Lexical density

A frequently used adjective is "concise", found in [4], [6], [8], [13], and [19]. Oxford English Dictionary defines this word in this way: "Of speech or writing: Expressed in few words; brief and comprehensive in statement; not diffuse." [2] recommends using "the fewest words

that convey the sense of action and character". [11] and [12] advocate "economy" of words and language, [17] and [18] make the same point using different words ([17]: "if one word will work where three are, replace it.", "avoid wordiness" [18]: "brevity", "if it is possible to cut out a word, always cut it out (Orwell)").

[16] makes an indirect reference to lexical density by referring to "weak words" such as "idea, interesting, intriguing" and "vapid" words such as "issue", which supposedly decrease density of meaning.

Synthesis: It is very hard to capture lexical density computationally and the only way to minimize semantic redundancy that seems plausible to us is to proceed using semantic networks and internal representation of text meaning.

A partial solution might be to search for synonyms (with different degrees of synonymy) within a close range around the word.

Word choice

The major point regarding word choice is voiced in two distinct ways. One is to use short words ([1], [11], [14], [18]) and the other to use words with Anglo-Saxon origin ([11], [18]) (as opposed to Latin in most cases), which has basically the same implications.

As to the choice between registers, [1], [2], [3], [8], [10], [17], and [20] recommend some kind of simplicity. For [1] it is "familiar words"/"plain simple language", [2], [8], and [17] speak of a "simple way" or "manner" of writing, [10] and [20] specify that the goal is "conversational writing". [3] does not describe positive qualities, but it assumes a strong position against "tides of phony, posturing, pretentious, tired, imprecise slovenly language, which both suffocate and corrupt the mind".

[3] advises writers to avoid "overwriting". *Oxford English Dictionary* offers two relevant definitions: "To write too much" and "To write (a work, or aspect of a work, etc.) in too elaborate or ornate a style".

[11] warns against using slang, contractions and colloquialisms in formal writing.

Synthesis: Regarding the etymological origin of the words, many dictionaries (e. g. the Oxford English Dictionary) contain this kind of information. If one of these dictionaries was

properly integrated with a text analysis software, it would be possible to calculate the Anglo-Saxon quotient of a text as the ratio between the content words with Anglo-Saxon origin and the total number of content words (nouns, verbs and adjectives would probably suffice).

Choice between registers returns us to section 4.2.1, where we suggested calculating the average rank of words based on corpus frequency. Presuming that "plain simple language" are effectively the most frequent words, this advice suggests that the lower the "original vocabulary quotient" , the better. Eventually, a low optimum would probably be reached. (However, as practice shows, notable exceptions would have to be accepted (Joyce and Nabokov come to mind) and therefore we have to express caution and stress the need of empirical evaluation of any of these conclusions)

Slang words and colloquialisms are marked as such in dictionaries, and contractions are detectable by seeking the character ', which makes the informality check of formal writing an easily implementable task.

Voice

[1], [4], [7], [11], [14], [18], and [19] either recommend the active voice or warn against the passive. [19] argues specifically against the verb "to be".

Synthesis: We suggest calculating the "passivity quotient" by searching for past participles co-occurring with different forms of the verb "to be" and obtaining the ratio of these to all verbs used.

Sentence length

There are different opinions regarding sentence length. [1] directly recommends "short sentences", [11] speaks of "relative" shortness, [17] argues only for avoiding "long" sentences. [4] and [19] assert that "sentences should vary in length and structure". [10] provides the most precise delimitation of sentence length: "aim for 14-18 words per sentence on average".

[1], [2], [3], [8], [10], [17], and [20], which I quoted in the part “Word choice” as promoting simplicity of writing, can be also understood to advocate sentences short enough to be considered simple.

Synthesis: Any software which is complex enough to distinguish between the different uses of interpunction is capable of calculating the average length of sentences. The 14 to 18 word optimum should be tested on established texts of different genres. Standard deviation could indicate the variability in length.

Sentence syntax

All references to simplicity refer to syntax as well. In addition, [11] recommends to “begin most sentences with the subject, other provide useful variety only sparingly”. This is in slight conflict with [4] and [19], which say that sentences should vary in structure. [11] also offers specific advice by warning against “excessive use of dependent clauses, stringing together [of] prepositional phrases”. [19] agrees with avoiding “strings of prepositions [prepositional phrases]”

Synthesis: There is no consensus in the question of simplicity vs. variability. On the other hand, dependent clauses and prepositional phrases can be easily identified using syntactic parsing (for example Enju (“Enju”)). Excessive use of both of these could be highlighted by the program and recommended for review.

Repetition

[3], [4], [11], [16], and [18] warn against repetition. But except for [11], which mentions “close proximity”, these sources do not state within what radius multiple occurrences can be considered repetition. Except for [11] and [16], which target “words and phrases” and “redundancy, semantic repetition” respectively, the sources do not specify what is being repeated. There is also a completely different position indicated in [4] (“good repetition”, “repeating themes, symbols, and images can be powerful”) and [15] says “repeating choruses might let the story breathe”.

Synthesis: Because of the disagreement concerning repetition, we suggest that computationally identified repetition should not serve as a measure included in automated evaluation of text quality. However, a program pointing out repetition and letting a human evaluator decide whether it is justifiable could be an important contribution towards highly efficient computer-aided text quality evaluation.

Cliché

Cliché is a form of phrase/semantic repetition, differing in the span between individual uses. It could be defined as a phrase or image frequently repeated in the entire body of cultural discourse. [1] and [15] simply recommend avoiding clichés, [11] specifies them as "empty words and phrases" (related to "Lexical density"), and [16], without using the term, describes unwanted "sweeping statements" such as "from the dawn of humanity".

Synthesis: This issue was expounded on in the part 4.2.2 titled "Cliché identification".

Auditory characteristics

[1], [2], [14], [15], and [18] make a reference to the sound of the written text as one of the measures of quality. In [1], some cases of Hemingway's simple repetition (see part on "Repetition") are praised for their "rich and powerful rhythm". [2] encourages the writer to "cultivate a good 'ear', a sense of pitch". [14], [15], and [18] recommend reading a text out loud to discover imperfections.

Synthesis: Although it is obviously hard to capture the notion of "good rhythm" computationally, state-of-the-art text-to-speech software is advanced enough to possess all necessary information as to the phonemes and syllables pronounced. Taking good examples of rhythm from poetry and prose, a neural network could learn to distinguish the pleasing sound patterns from the neutral ones.

Imitation

Several texts encourage analysis and imitation of models. [1] quotes Wynford Hicks recommending "analysing the techniques you admire". Referring mostly to training, [7], [10], and [17] encourage writers to "copy" ([7]), "imitate" ([10]), and "use other writers' sentences

and paragraphs as models and emulate the syntactic structures with [their] own content" ([17]).

Synthesis: Since style/technique is an elusive category and this advice in its most observable demonstration targets training rather than practice, we do not consider imitation a meaningful addition to text quality evaluation. However, note that the issue of similar styles/techniques is addressable by part 4.1 on "Straightforward neural network approach" (in theory, a properly taught neural network could extract something similar to style from the texts).

6 CONCLUSIONS

This paper is an attempt to introduce a promising way of research – operationalizing the different criteria for advanced text quality evaluation. It seems to have been neglected due to the inherent subjectivity of judgement. Also, the diametrically different criteria of quality for different genres do not invite a unifying framework. However, as we have shown, a possible solution to the problems of subjectivity and genre specificity is to capture **multiple measurable dimensions which influence the quality of any text** and let the evaluating person choose which dimensions are used and which ranges of values are acceptable. A major caveat surfaced during our research: as certain novel and creative uses of language may elude the hitherto captured notions of quality, **it is strongly recommended to combine the automatic evaluation with expert human evaluation** in order to override the false assumptions of the automatic evaluation system, reflect on and edit its parameters accordingly and thus continually improve the framework. Used in this way, the automatic evaluation has the potential to be an immediately effective pre-processor and highlighter, helping human evaluators notice the important quantitative indicators.

Where does the way lead from here? Programmers and corpus linguists might implement the dimensions we suggested, test their consistency in general or within genres, and create a graphical user interface designed to assist human evaluators; for example, highlighting clichés, instances of repetition, or showing charts with word type ratios and average sentence length. Theorists and practicing human evaluators are free to add dimensions to the ones we suggested, given that there is a proof of a certain degree of objectivity of their observations.

To end the paper with a philosophical paragraph, we would like to note that our research is not intended to and should not appear as an assertion that one day, we will be able to capture every conceivable aspect of text quality in an automatic evaluation model. We maintain all due respect for the complexity of the foremost works of art. Rather, this project is meant to be a slightly sarcastic intrusion into elitist beliefs that literary taste is a fine, uncapturable skill of a higher order, as opposed to simple accumulation of training hours.

WEBSITES USED FOR CONTENT ANALYSIS

[1] Albert, Tim. "What is a good writing style?" *Tim Albert.* November 10, 2010. <http://www.timalbert.co.uk/page.php?action=article&ID=70>

[2] Clarke, Caro. "Style, the life and death of a writer." *Caro Clarke*. <http://www.caroclarke.com/style.html>

[3] Crossen, Cynthia. "What Makes Bad Writing". *The Wall Street Journal*. June 25, 2012. <http://online.wsj.com/article/SB10001424052702304898704577483133208871806.html>

[4] Donovan, Melissa. "36 Tips for Writing Just About Anything." *Writing Forward.* August 21, 2012. <http://www.writingforward.com/writing-tips/tips-for-writing-just-about-anything>

[5] ---. "Eight Characteristics of Good Writing". *Writing Forward.* January 12, 2012. <http://www.writingforward.com/better-writing/characteristics-of-good-writing>

[6] ---. "Writing Tips: Abolish the Adverbs". *Writing Forward.* January 30, 2013. <http://www.writingforward.com/writing-tips/writing-tips-abolish-adverbs>

[7] Dorian, Mars. "How to Create a Writing Style That Impacts Your Audience Like A Blaaazing Meteor". *Mars Dorian*. <http://www.marsdorian.com/2011/01/writing-style-that-impacts-people/>

[8] Falconer, Erin. "10 Writing Tips from the Masters". *Pick the Brain.* <http://www.pickthebrain.com/blog/art-of-writing/>

[9] Goins, Jeff. "Writng Tip: Be Specific". *Goins, Writer*. January 10, 2011. <http://goinswriter.com/2011/01/10/writing-tip-be-specific/>

[10] Gray-Grant, Daphne. "10 Most Common Writing Mistakes, Plus 10 Remedies". *Quips and Tips for Successful Writers*. December 7, 2010. <http://theadventurouswriter.com/blogwriting/most-common-writing-mistakes-writing-tips/>

[11] Griffin, Roger A. "Elements of a Good Writing Style". *Using the Internet as a Resource for Historical Research and Writing.* 1996. <http://www.austincc.edu/history/inres10a4style.html>

[12] Marcus, Karen. "What Makes Good Writing?". *Final Draft Communications.* August 14, 2009. <http://www.finaldraftcommunications.com/what-makes-good-writing/>

[13] Nordquist, Richard. "What Are the Characteristics of Good Writing?". *About.com Grammar and Composition*. <http://grammar.about.com/od/yourwriting/a/characteristics.htm>

[14] Owen, Audrey. "Writing Tips". *Writer's Helper.* <http://www.writershelper.com/writingtips.html>

[15] Palahniuk, Chuck. "36 Writing Essays by Chuck Palahniuk". *LitReactor*. September 17, 2011. <http://litreactor.com/essays/36-writing-essays-by-chuck-palahniuk>

[16] Scarce, Rick. "Writing Tips: Advice, Hints, and Teensy Pearls of Wisdom for Weary and Wary Writers". March 2012. <http://www.skidmore.edu/~rscarce/WritingTips.htm>

[17] Scocco, Daniel. "34 Writing Tips That Will Make You a Better Writer". *Daily Writing Tips*. <http://www.dailywritingtips.com/34-writing-tips-that-will-make-you-a-better-writer/>

[18] Smith, Peter. "Developing a writing style." *Logic Matters*. December 2012. <http://www.logicmatters.net/students/writing-style/>

[19] University of Illinois at Urbana-Champaign. "Writing Tips: Five Editing Principles". *The Center for Writing Studies.* <http://www.cws.illinois.edu/workshop/writers/tips/editing/>

[20] Wax, Dustin. “8 Qualities of Powerful Writing”. *Lifehack*. January 15, 2013. <http://www.lifehack.org/articles/communication/8-qualities-of-powerful-writing.html>

REFERENCES

Banerjee, Jayanti, and Spiros Papageorgiou. "Analysis of written language." *Qualitative Methods in Language Testing and Assessment*. 2009. Web. 26 Feb. 2013. <http://www.ealta.eu.org/conference/2009/docs/workshop2/D2_2_AWL.pdf>.

ClichéSite (2006). Web. 26 Feb 2013. <http://www.clichesite.com/>.

"concise, adj.". OED Online. December 2012. Oxford University Press. 3 March 2013 <http://www.oed.com/view/Entry/38276?rskey=Knt6hw&result=1&isAdvanced=false>.

Davies, Mark. (2008-) *The Corpus of Contemporary American English: 450 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/.

Elman, J. L. 1990. “Finding Structure in Time.” *Cognitive Science* 14: 179-211.

“Enju.” Version 2.4.2. The University of Tokyo, Department of Computer Science, Tsujii laboratory. June 16, 2011. <http://www.nactem.ac.uk/enju/>

Friedman, S. Morgan. *Cliché Finder* (1996). Web. 26 Feb 2013. <http://www.westegg.com/cliche/>.

Hayes, J. R., Flower, L., Schriver, K. A., Stratman, J., and Carey, L., “Cognitive Processes in Revision,” in *Advances in Applied Psycholinguistics, Volume 11: Reading, Writing, and Language Processing*, S. Rosenberg (ed.), Cambridge, England: Cambridge University Press, 1987, pp. 176-240.

Higuchi, Koichi. “KHCoder.” *Sourceforge*. 2001-2013. <http://khc.sourceforge.net/en/>

Hoover, D. L. 2003. Another Perspective on Vocabulary Richness. *Computers and the Humanities* 37: 151-178.

Landauer, T., Foltz, P., & Laham, D. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284. doi:10.1080/01638539809545028.

Laufer, B., Nation, P. 1995. "Vocabulary Size and Use: Lexical Richness in L2 Written Production." *Applied Linguistics*, Vol. 16, No. 3, 307-322.

Louis, A. and Nenkova, A. 2012. "A coherence model based on syntactic patterns." In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1157-1168.

Nenkova, Ani. "Automatic Text Understanding of Content and Text Quality." *2011 US Frontiers of Engineering Symposium.* Web. 26 Feb. 2013. <http://www.naefrontiers.org/File.aspx?id=31596>.

Nenkova, A., Chae, J., Louis, A., Pitler, E. 2010. "Structural Features for Predicting the Linguistic Quality of Text." In Krahmer, E., Theune, M. (Eds.). *Empirical Methods in NLG*, LNAI 5790, pp. 222–241, 2010. Heidelberg: Springer, 2010.

"overwrite, v.". OED Online. December 2012. Oxford University Press. 9 March 2013 <http://www.oed.com/view/Entry/135394?redirectedFrom=overwrite>.

Paukkeri, M., Ollikainen, M., Honkela, T. 2013. "Assessing user-specific difficulty of documents." In *Information Processing and Management* 49 (2013). 198-212.

Schriver, K. A. 1989. Evaluating Text Quality: The Continuum From Text-Focused to Reader-Focused Methods. *IEEE Transactions on Professional Communication*, Vol. 32, No. 4, December 1989. 238-255.

Simpson-Vlach, Rita, and Nick C. Ellis. "An Academic Formulas List: New Methods In Phraseology Research." *Applied Linguistics*. 31.4 (2010): 487-512. Print.

Smith, A. G., Zee, C. X. S., Uitdenbogerd, A. L. 2012. "In Your Eyes: Identifying Clichés in Song Lyrics." *Proceedings of the Australasian Language Technology Workshop*. Vol. 10, 2012. pp. 88-96. ISSN 1834-7037.

Strunk, William. *Elements of Style.* Ithaca: W. P. Humphrey, 1918. Print.

Qumsiyeh, R., Ng, Y. 2011. ReadAid: A Robust and Fully-Automated Readability Assessment Tool. *2011 23rd IEEE International Conference of Tools with Artificial Intelligence*. 539-546. DOI: 10.1109/ICTAI.2011.87

Wray, Alison. *Formulaic Language and the Lexicon*. London: Cambridge UP, 2002. Print.