Dialogue and Discourse 4(2) (2013) 87-117 doi: 10.5087/dad.2013.205

# A corpus of science journalism for analyzing writing quality

**Annie Louis** LANNIE@SEAS.UPENN.EDU
*University of Pennsylvania*
*Philadelphia, PA, 19104, USA*

**Ani Nenkova** NENKOVA@SEAS.UPENN.EDU
*University of Pennsylvania*
*Philadelphia, PA, 19104, USA*

**Editors:** Stefanie Dipper, Heike Zinsmeister, Bonnie Webber

## Abstract

We introduce a corpus of science journalism articles, categorized in three levels of writing quality.[1] The corpus fulfills a glaring need for realistic data on which applications concerned with predicting text quality can be developed and evaluated. In this article we describe how we identified, guided by the judgements of renowned journalists, samples of excellent, very good and typical writing. The first category comprises extraordinarily well-written pieces as identified by expert journalists. We expanded this set with other articles written by the authors of these excellent samples to form a set of very good writing samples. In addition, our corpus also comprises a larger set of typical journalistic writing. We provide details about the corpus and the text quality evaluations it can support.

Our intention is to further extend the corpus with annotations of phenomena that reveal quantifiable differences between levels of writing quality. Here we introduce two such annotations that have promise for distinguishing amazing from typical writing: text generality/specificity and communicative goals. We present manual annotation experiments for specificity of text and also explore the feasibility of acquiring these annotations automatically. For communicative goals, we present an automatic clustering method to explore the possible set of communicative goals and develop guidelines for future manual annotations. We find that the annotation of general/specific nature on sentence level can be performed reasonably accurately fully automatically, while automatic annotations of communicative goals reveals salient characteristics of journalistic writing but does not align with categories we wish to annotate in future work. Still with the current automatic annotations, we provide evidence that features based on specificity and communicative goals are indeed predictive of writing quality.

**Keywords:** Text quality, Readability, Science journalism, Corpus

## 1. Introduction

Text quality is an elusive concept. It is difficult to define text quality precisely but it has huge potential for transforming the way we use applications to access information. For example in information retrieval, rankings of documents can be improved by balancing the relevance of the document to the user query and the estimated quality of the document. Similarly, tools for writing support can be developed to help people improve their writing. Text quality prediction requires that texts can

1. The corpus can be obtained from `http://www.cis.upenn.edu/~nlp/corpora/scinewscorpus.html`

 Submitted 03/12; Accepted 12/12; Published online 05/13

be automatic analyzed in multiple dimensions and naturally links work from seemingly disparate fields of natural language processing such as discourse analysis, coherence, readability, idiomatic language use, metaphor identification and interpretation and automatic essay grading.

To develop models to predict text quality, we need datasets where articles have been marked with a quality label. But the problem of coming up with a suitable definition for article quality has hindered corpus development efforts. There are reasonable sources of text quality labels for low level aspects such as spelling and grammar in the form of student essays and search query logs. Such data, however, have remained proprietary and not available for research outside of the specific institutions. There are also datasets where automatic machine summaries and translations have been rated by human judges. Yet systems trained on such data are very unlikely to transfer with the same accuracy to texts written by people. On the other hand, studies that were not based on data from automatic systems have made a simple assumption that articles of bad quality can be obtained by manipulating good articles. One approach used to evaluate many systems that predict organization quality is to take an article and randomly permute its sentences to obtain an incoherent sample. However, such examples are rather unrealistic and are not representative of problems generally encounted in writing. Experiments on such datasets may give optimistic results that do not carry over to real application settings. Well-written texts are also marked by properties beyond general spelling, grammar and organization. They are beautifully written, use creative language and are on well-chosen topics. There are no datasets on which we can explore linguistic correlates of such writing.

The goal of the work that we describe here is to provide a realistic dataset on which researchers interested in text quality can develop and validate their theories about what properties define a well-written text. Our corpus consists of science journalism articles. This genre has unique features particularly suitable for investigating text quality. Science news articles convey complex ideas and research findings in a clear and engaging way, targeted towards a sophisticated audience. Good science writing explains, educates and entertains the reader at the same time. For example, consider the snippets in Table 1.

They provide a view of the different styles of writing in this genre. The impact and importance of a research finding is conveyed clearly and the research itself explained thoroughly (Snippet A). Snippet B shows the use of creative writing style involving unexpected phrasing (“fluent in firefly”) and visual language (“rise into the air and begin to blink on and off”). The last snippet also presents humour and a light-hearted discussion of the research finding. These examples show how this genre can be interesting for studying text quality. Note also how these examples are in great contrast with research descriptions commonly provided in journals and conference publications.

In this regard, these articles offer a challenging testbed for computational models for predicting to what extent a text is clear, interesting, beautiful or well-structured. Our corpus contains several thousand articles, divided into three coarse levels of writing quality in this genre. All articles have been published in the New York Times (NYT), so the quality of a typical article is high. A small sample of GREAT articles was identified with the help of subjective expert judgements of established science writers. A substantially larger set of VERY GOOD writing was created by identifying articles published in the NYT and written by writers whose texts appeared in the GREAT section of the corpus. Finally, science-related pieces on topics similar to those covered in the GREAT and VERY GOOD articles but written by different authors formed the set of TYPICAL writing. In Section 3 we provide specific details about the collection and categorization of articles in the corpus. Our corpus is also realistic compared to data sets used in the past and also relevant for many application

**A. Clarity in explaining research**
The mystery of time is connected with some of the thorniest questions in physics, as well as in philosophy, like why we remember the past but not the future, how causality works, why you can't stir cream out of your coffee or put perfume back in a bottle. ... Dr. Fotini Markopoulou Kalamara of the Perimeter Institute described time as, if not an illusion, an approximation, "a bit like the way you can see the river flow in a smooth way even though the individual water molecules follow much more complicated patterns."

**B. Creative language use**
Sara Lewis is fluent in firefly. On this night she walks through a farm field in eastern Massachusetts, watching the first fireflies of the evening rise into the air and begin to blink on and off. Dr. Lewis, an evolutionary ecologist at Tufts University, points out six species in this meadow, each with its own pattern of flashes.

**C. Entertaining nature**
News flash: we're boring.
New research that makes creative use of sensitive location-tracking data from 100,000 cellphones in Europe suggests that most people can be found in one of just a few locations at any time, and that they do not generally go far from home.

Table 1: Example snippets from science news articles

settings such as article search and recommedation. A further attractive aspect of our corpus is that the average quality of an article is high, they were all edited and published in a leading newspaper. Therefore our corpus allows for examining linguistic aspects related to quality at this higher end of the quality spectrum. Such distinctions have been unexplored in past studies.

Given the examples in Table 1, we can see that automatic identification of many aspects such as research content, metaphor, visual language can be helpful for predicting text quality. As a first step, manual annotations of these dimensions would be necessary to train systems to automatically annotate new articles. In Sections 5 and 6, we focus on the annotation of two aspects–general/specific nature and communicative goals of sentences. The general-specific distinction refers to the high level statements and specific details in the article. We believe that a proper balance between the use of general and specific information can contribute to text quality. Communicative goals refer to the author's intentions for every sentence. For example, the intention may be to define a concept, provide an example, narrate a story etc. Certain sequences of intentions may work better and create articles of finer quality compared to other intention patterns. Both these aspects, text specificity and communicative goals, appear to be relevant for quality given the explanatory nature of the science journalism genre.

In Section 5 we give further motivation for annotating specificity. We also present results on sentence-level annotation of sentence specificity both by annotators and automatically; we derive article-level characterizations of the distribution of specificity in the text and perform initial experiments which indicate that text specificity, computed fully automatically, varies between texts from different categories in the corpus. In Section 6 we discuss the need for annotation of communicative goals. We experiment with an unsupervised approach for capturing types of communicative goals. We find that the model reveals systematic differences between categories of writing quality

but does not directly correspond to distinctions we intended to capture. In Section 8 we conclude with discussion of further cycles of introducing factors related to text quality and annotating data either manually or automatically.

## 2. Background

In this section, we describe the different aspects of text quality which have been investigated in prior work, with special focus on the datasets used to validate the experiments.

There is the rich and well-developed field of readability research where the aim is to predict if a text is appropriate for a target audience. The audience is typically categorized by age, education level or cognitive abilities (Flesch, 1948; Gunning, 1952; Dale and Chall, 1948; Schwarm and Ostendorf, 2005; Si and Callan, 2001; Collins-Thompson and Callan, 2004). Most of the data data used for these experiments accordingly come from educational material designed for different grade levels. Some studies have focused on a special audience such as non-native language learners (Heilman et al., 2007) and people with cognitive disabilities (Feng et al., 2009). In Heilman et al. (2007), the authors use a corpus of practice reading material designed for English language learners. Feng et al. (2009) create a corpus of articles where for each article, they had their target users answer comprehension questions and the average score obtained by the users for each article was used as a measure of that article's difficulty. But readability work does not traditionally address the question of how we can rank texts suitable for the same audience into well or poorly written ones. In our corpus, we aim to differentiate texts that are exceptionally well-written from those which are typical in the science section of a newspaper.

Some recent approaches have focused on this problem of capturing differences in coherence within one level of readership. For example, they seek to understand what is a good structure for a news article when we consider a competent, college-educated adult reader. Rather than developing scores based on cognitive factors, they characterize text coherence by exploring systematic lexical patterns (Lapata, 2003; Barzilay and Lee, 2004; Soricut and Marcu, 2006), entity coreference (Barzilay and Lapata, 2008; Elsner et al., 2007; Karamanis et al., 2009) and discourse relations (Pitler and Nenkova, 2008; Lin et al., 2011) from large collections of texts. But since suitable data for evaluation is not available, a standard evaluation technique for several of these studies is to test to what extent a model is able to distinguish a text from a random permutation of the sentences in the same text. This approach removes the need for creating dedicated corpora with coherence ratings but severely limits the scope of findings because it remains an open question if results on permutation data will generalize for comparisons of text in more realistic tasks. One particular characteristic of the permutations data is that both the original and permuted texts are the same length. This setting might be easier for systems than the case where texts to be compared do not necessarily have the same length. Further, incoherent examples created by permuting sentences resemble text generated by automatic systems. For articles written by people there may be more subtle differences than the clearly incoherent permutation examples and performance may be unrealistically high on the permutations data. In this area specifically the need for realistic test data is acute.

Work on automatic essay grading (Burstein et al., 2003; Higgins et al., 2004; Attali and Burstein, 2006) and error correction in texts written by non-native speakers (De Felice and Pulman, 2008; Gamon et al., 2008; Tetreault et al., 2010) has utilized actual annotations on student writing. However, most datasets for these tasks are proprietary. In addition, they seek to identify potential deviations from a typical text that indicates poor mastery of language. The corpus that we collect targets analy-

sis at the opposite end of the competency spectrum, seeking to find superior texts among reasonably high typical standard.

The largest freely available annotations of linguistic quality come from manual evaluations of machine generated text such as machine translation and summarization. In several prior studies researchers have designed metrics that can replicate the human ratings of quality for these texts. However, the factors which indicate well-written nature of human texts are rather different from those useful for predicting quality of machine generated texts; recent findings suggest that prediction methods trained on both machine translated sentences or machine produced summaries perform poorly for articles written by people (Nenkova et al., 2010). So these annotations have little use outside the domain of evaluating and tuning automatic systems.

Another aspect that has not been considered in prior work is the effect of genre on metrics for writing quality. Readability studies have taken motivation in cognitive factors and their measures based on word familiarity and sentence complexity are assumed to be generally applicable for most texts. On the other hand, data-driven methods aiming to learn coherence indicators are proposed such that they can utilize any collection of texts of interest and learn their properties. However, a well-written story would be characterised by different factors compared to good writing in academic publications. There are no existing corpora that can be used to develop and evaluate metrics that are unique for a genre.

We believe that our corpus will help to evaluate not just readability and coherence, but the overall well-written nature or text quality. For this, we need to have texts that are not focused on competency levels of the audience. Science journalism is intended for an adult educated audience and our corpus picks out articles that are considered extremely well-written compared to typical articles appearing in the science section of a newspaper. So our corpus ratings directly reflect differences in writing quality for the same target audience. Secondly, our corpus will also enable the study of genre-specific metrics beyond ease of reading. These articles can be examined to understand linguistic correlates of interesting and clear writing and of the style that accomodates and encourages lay readers to understand research details. We believe that these important characteristics of science writing can be better studied using our corpus.

In terms of applications of measuring text quality, some recent work on information retrieval have incorporated readability ratings in the ranking of web page results (Collins-Thompson et al., 2011; Kim et al., 2012). But these studies so far have utilized the notion of grade levels only and the dominant approach is to predict the level using the vocabulary differences learnt from the training corpus. We believe that our corpus can provide a challenging setting for such a ranking task. In our corpus, for every well-written article, we also identify and list other topically similar articles which have more typical or average writing. While in the case of retrieving webpages on different topics, it may be difficult to apply genre-specific measures, we hope that for our corpus one can easily explore the impact of different measures: relevance to query, generic readability scores and metrics specific to science journalism. So far, only some studies on social media have been able to incorporate domain-specific features such as typos and punctuation for the ranking of questions and answers in a forum (Agichtein et al., 2008).

## 3. Collecting a corpus of science journalism

Ratings for text quality are highly subjective. Several factors could influence judgements such as choice of and familiarity with the topic and personal preference, particularly for specialized domains

such as science. In our work we have chosen to adopt the opinion of renowned science journalists to identify articles deemed to be GREAT. We then expand the corpus using author and topic information to form a second set of VERY GOOD articles. A third level of TYPICAL articles on similar topics but by different authors was then formed.

All articles in the corpus were published in the New York Times between 1999 and 2007.

### 3.1 Selecting GREAT articles

The GREAT articles in our corpus come from the "Best American Science Writing" annual anthologies. The stories that appear in these anthologies are chosen by prominent science journalists who serve as editors of the volume, with a different editor overseeing the selection each year. In some of the volumes, the editors explain the criteria they have applied while searching for articles to include in the volume:

- "First and most important, all are extremely well written. This sounds obvious, and it is, but for me it means the pieces impart genuine pleasure via the writers' choice of words and the rhythm of their phrases... "I wish I'd written that", was my own frequent reaction to these articles." (2004)
- "The best science writing is science writing that is cool... I like science writing to be clear and to be interesting to scientists and nonscientists alike. I like it to be smart. I like it, every once in a while, to be funny. I like science writing to have a beginning, middle and end—to tell a story whenever possible." (2006)
- "Three attributes make these stories not just great science but great journalism: a compelling story, not just a topic; extraordinary, often exclusive reporting; and a facility for concisely expressing complex ideas and masses of information." (2008)

Given the criteria applied by the editors, the "Best American Science Writing" articles present a wonderful opportunity to test computational models of structure, coherence, clarity, humor and creative language use.

From the volumes published between 1999 and 2007, we picked out articles that originally appeared in the New York Times newspaper. It was straightforward to obtain the full text of these articles from the New York Times corpus (Sandhaus, 2008) which has all articles published in the NYT for 20 years, together with extensive metadata containing editor assigned topic tags. All articles in our corpus consist of the full text of the article and the associated NYT corpus metadata.

There are 52 articles that appear both in the anthologies and the NYT corpus and they form the set of GREAT writing.

Obviously, the topic of an article will influence the extent to which it is perceived as well-written. One would expect that it is more likely to write an informative, enjoyable and attention-grabbing newspaper article related to medicine and heath compared to writing a piece with equivalent impact on the reader on the topic of coherence models in computational linguistics.

We use the NYT corpus topic tags to provide a first characterization of the articles we got from the "Best American Science Writing" anthology. There are about 5 million unique tags in the full NYT corpus and most articles have five or six tags each. The number of unique tags for the set of GREAT writing articles is 325 which is too big to present. Instead, in Table 2 we present the tags that appear in more than three articles in the GREAT set. Medicine, space and physics are the most popular subjects in the collection. Computers and finance topics are much lower in the list.

In the next section we describe the procedure we used to further expand the corpus with samples of VERY GOOD and TYPICAL writing.

| Tag | No. articles |
|---|---|
| Medicine and Health | 18 |
| Research | 17 |
| Science and Technology | 12 |
| Space | 11 |
| Physics | 9 |
| Archaeology and Anthropology | 7 |
| Biology and Biochemistry | 7 |
| Genetics and Heredity | 6 |
| DNA (Deoxyribonucleic Acid) | 5 |
| Finances | 5 |
| Animals | 4 |
| Computers and the Internet | 4 |
| Diseases and Conditions | 4 |
| Doctors | 4 |
| Drugs (Pharmaceuticals) | 4 |
| Ethics | 4 |
| Planets | 4 |
| Reproduction (Biological) | 4 |
| Women | 4 |

Table 2: Most frequent metadata tags in the GREAT writing samples

### 3.2 Extraction of VERY GOOD and TYPICAL writing

The number of GREAT articles is relatively small—just 52—so we expanded the collection of good writing using the NYT corpus. The set of VERY GOOD writing contains NYT articles about research that were written by authors whose articles appeared in the GREAT sub-corpus. For the TYPICAL category, we pick other articles published around the same time but were neither chosen as best writing nor written by the authors whose articles were chosen for the anthologies.

The NYT corpus contains every article published between 1987 to 2007 and has a few million articles. We first filter some of the articles based on topic and research content before sampling for good and average examples. The goal of the filtering is to find articles about science that were published around the same time as our GREAT samples and have similar length. We consider only:

- Articles published between 1999 and 2007. The best science writing anthologies have been published since 1997 and the NYT corpus contains articles upto 2007.
- Articles of at least 1,000 words. All articles from the anthologies had that minimum length.
- Only science journalism pieces.

In the NYT metadata, there is no specific tag that identifies all the science journalism articles. So, we create a set of metadata tags which can represent this genre. Since we know the GREAT article set to be science writing, we choose the minimal subset of tags such that at least one tag per GREAT article appears on the list. We dub this set as the "science tags". We derived this list using greedy selection, choosing the tag that describes the largest number of GREAT articles, then the tag that appears in most of the remaining articles, and so on until we obtain a list of tags that covers all GREAT articles. Table 3 lists the eleven topic tags that made it into the list.

| | | | |
|---|---|---|---|
| Medicine and Health | Research | Space | Physics |
| Computers and the Internet | Brain | Evolution | Disasters |
| Religion and Churches | Language and Languages | Environment | |

Table 3: Minimal set of "science tags" which cover all GREAT articles

| **People** | **Process** | **Topic** | **Publications** | **Endings** | **Other** |
|---|---|---|---|---|---|
| researcher | discover | biology | report | -ology | human |
| scientist | discuss | physics | published | -gist | science |
| physicist | experiment | chemistry | journal | -list | research |
| biologist | work | anthropology | paper | -mist | knowledge |
| economist | finding | primatology | author | -uist | university |
| anthropologist | study | | issue | -phy | laboratory |
| environmentalist | question | | | | lab |
| linguist | project | | | | |
| professor | | | | | |
| dr | | | | | |
| student | | | | | |

Table 4: Unique words from the research word dictionary

We consider an article to be science related if it has one of the topic tags in Table 3 and also mentions words related to science such as 'scientist', 'discover', 'found', 'physics', 'publication', 'study'. We found the need to check for words that appear in the article because in the NYT, research-related tags are assigned even to articles that only cursorily mention a research problem such as stem cells but otherwise report general news. We used a hand-built dictionary of research words and remove articles that do not meet a threshold for research word content. The dictionary comprises a total of 71 lexical items including morphological variants. Six of the entries in the dictionary are regular expression patterns that match endings such as "-ology" and "gist" that often indicate research related words. The unique words from our list are given in Table 4. We have grouped them into some simple categories here. An article was filtered out when (a) fewer than 10 of its tokens matched any entry in the dictionary or (b) there were fewer than 5 unique words from the article that had dictionary matches. This threshold keeps articles that have high frequency of research words and also diversity in these words. The threshold values were tuned such that all the articles in the GREAT set scored above the cutoff. After this step, the final *relevant* set has 13,466 science-related articles on the same topics as the GREAT samples.

The GREAT articles were written by 42 different authors. Some authors had more than one article appearing in that set, and a few have even three or more articles in that category. It is reasonable to consider that these writers are exceptionally good, so we extracted all articles from the relevant set written by these authors to form the VERY GOOD set. There are 2,243 in that category.

The remaining articles from the relevant set are grouped to form the TYPICAL class of articles, a total of 11,223.

A summary of the three categories of articles is given in Table 5.

| Category | No. articles | No. sentences | No. tokens |
|---|---|---|---|
| GREAT | 52 | 6,616 | 163,184 |
| VERY GOOD | 2243 | 160,563 | 4,054,057 |
| TYPICAL | 11223 | 868,059 | 21,522,690 |
| Total | 13518 | 1,035,238 | 25,739,931 |

Table 5: Overview of GREAT, VERY GOOD and TYPICAL categories in the corpus

### 3.3 Ranking corpus

As already noted in the previous section, the articles in science journalism span a wide variety of topics. The writing style for articles from different topics, for example health vs. religion research would be widely different and hard to analyze for quality differences. Further, the sort of applications we envision, article recommendation for example, would involve ranking articles within a topic. So we pair up articles by topic to also create a ranking corpus.

For each article in the GREAT and VERY GOOD sets, we associate a list of articles from the TYPICAL category which discuss the same or closely related topic. To identify topically similar articles, we compute similarity between articles. Only the descriptive topic words identified via a log likelihood ratio test are used in the computation of similarity. The descriptive words are computed by the TopicS tool[2] (Louis and Nenkova, 2012b). Each article is represented by binary features which indicate the presence of each topic word. The similarity between two articles is computed as the cosine between their vectors. Articles with similarity equal to or above 0.2 were matched. Table 6 gives an example of two matched articles.

The number of TYPICAL articles associated with each GREAT or VERY GOOD article varied considerably: 1,138 of the GOOD articles were matched with 1 to 10 TYPICAL articles; 685 were matched with 11 to 50 TYPICAL articles and 59 were matched with 51 to 140 TYPICAL articles. If we enumerate all pairs of (GREAT or VERY GOOD, TYPICAL) articles from these matchings, there are a total of 25032 pairs.

In this way, for many of the high quality articles we have collected examples with hypothesized inferior quality, but on the same topic. The dataset can be used for exploring ranking tasks where the goal would be to correctly identify the best article given a good article among a set of similar articles.

The corpus can support investigations into what topics, events or stories draw readers' attention. High quality writing is interesting and engaging and there are few studies that have attempted to predict this characteristic. McIntyre and Lapata (2009) present a classifier for predicting interest ratings for short fairy tales and Pitler and Nenkova (2008) report that ratings of text quality and reader interest are complementary and do not correlate well. To facilitate this line of investigation, as well as other semantically oriented analysis, in the corpus we provide a list of all descriptive topic words which were the basis for article mapping.

## 4. Annotating elements of text quality

So far, we described how we have collected a corpus of science journalism writing with varying quality. In later work we would like to further annotate these articles with different elements indica-

2. http://www.cis.upenn.edu/ lannie/topicS.html

GREAT **writing sample**

Kristen Ehresmann, a Minnesota Department of Health official, had just told a State Senate hearing that vaccines with microscopic amounts of mercury were safe. Libby Rupp, a mother of a 3-year-old girl with autism, was incredulous. “How did my daughter get so much mercury in her?” Ms. Rupp asked Ms. Ehresmann after her testimony. “Fish?” Ms. Ehresmann suggested. ”She never eats it,” Ms. Rupp answered. “Do you drink tap water?” “It’s all filtered.” “Well, do you breathe the air?” Ms. Ehresmann asked, with a resigned smile. Several parents looked angrily at Ms. Ehresmann, who left. Ms. Rupp remained, shaking with anger. That anyone could defend mercury in vaccines, she said, ”makes my blood boil.” Public health officials like Ms. Ehresmann, who herself has a son with autism, have been trying for years to convince parents like Ms. Rupp that there is no link between thimerosal – a mercury-containing preservative once used routinely in vaccines – and autism. They have failed. The Centers for Disease Control and Prevention, the Food and Drug Administration, the Institute of Medicine, the World Health Organization and the American Academy of Pediatrics have all largely dismissed the notion that thimerosal causes or contributes to autism. Five major studies have found no link. Yet despite all evidence to the contrary, the number of parents who blame thimerosal for their children’s autism has only increased. And in recent months, these parents have used their numbers, their passion and their organizing skills to become a potent national force. The issue has become one of the most fractious and divisive in pediatric medicine.

---

**Topically related** TYPICAL **text**

Neal Halsey’s life was dedicated to promoting vaccination. In June 1999, the Johns Hopkins pediatrician and scholar had completed a decade of service on the influential committees that decide which inoculations will be jabbed into the arms and thighs and buttocks of eight million American children each year. At the urging of Halsey and others, the number of vaccines mandated for children under 2 in the 90’s soared to 20, from 8. Kids were healthier for it, according to him. These simple, safe injections against hepatitis B and germs like haemophilus bacteria would help thousands grow up free of diseases like meningitis and liver cancer. Halsey’s view, however, was not shared by a footnotesize but vocal faction of parents who questioned whether all these shots did more harm than good. While many of the childhood infections that vaccines were designed to prevent – among them diphtheria, mumps, chickenpox and polio – seemed to be either antique or innocuous, serious chronic diseases like asthma, juvenile diabetes and autism were on the rise. And on the Internet, especially, a growing number of self-styled health activists blamed vaccines for these increases.

Table 6: Snippets from topically related texts belonging to the great and typical categories

tive of text quality and use these aspects to compute measurable differences between the TYPICAL science articles and the extraordinary writing in the GREAT and VERY GOOD classes. Some annotations may be carried out manually, others could be done on a larger scale automatically whenever system performance is good.

In the remainder of this article we discuss two types of annotation for science news: the general-specific nature of content and the communicative goals conveyed by sentences in the texts. We explore how these distinctions can be annotated and also present automatic methods which use surface clues in sentences to predict the distinctions. Using these automatic methods, we show how significant accuracies above baseline can be obtained for separating out articles according to text quality.

For the empirical studies reported later in the paper, we divide our data into three sets:

**Best** is the collection of GREAT articles from the corpus. No negative samples are used in this set for our analysis.

We analyze the distinction between VERY GOOD and TYPICAL articles matched by topic using the ranking corpus we described in Section 3.3. These articles are divided into two sets.

**Development corpus** comprised of 500 pairs of topically-matched VERY GOOD and TYPICAL articles. These pairs were randomly chosen.

**Test corpus** of the remaining (VERY GOOD, TYPICAL) pairs, totaling 24,532.

## 5. Specificity of content

This aspect is based on the idea that texts do not convey information at a constant level of detail. There are portions of the text where the author wishes to convey only the main topic and keeps such content devoid of any details. Certain other parts of the text are reserved for specific details on the same topics. For example, consider the two sentences below.

> *Dr. Berner recently refined his model to repair an old inconsistency.* [general]
> The revision, described in the May issue of The American Journal of Science, brings the model into closer agreement with the fact of wide glaciation 440 million years ago, yielding what he sees as stronger evidence of the dominant role of carbon dioxide then. [specific]

The first conveys only a summary of Dr. Berner's new research finding and alerts the reader to this topic. The next sentence describes more specifically what was achieved as part of Dr. Berner's research.

We put forward the hypothesis that the balance between general and specific content can be used to predict the text quality of articles. To facilitate this analysis, we created annotations of general and specific sentences for three of our articles and also developed a classifier that can automatically annotate new articles with high accuracy. Using features derived from the classifier's predictions, we obtained reasonable success in distinguishing the text quality levels in our corpus. This section describes the general-specific distinction, the annotations that we obtained for this aspect and experiments on quality prediction.

If texts gave only specific information, it would read like a bulleted list and less of an article or essay. In the same way that paragraphs divide an article into topics, authors introduce general

statements and content that abstract away from the details to give a big picture. Of course these general statements need to be supported with specific details elsewhere in the text. If the text only conveyed general information, readers would regard it as superficial and ambiguous. Details must be provided to augment the main statements made by the author. Books that provide advice about writing (Swales and Feak, 1994; Alred et al., 2003) frequently emphasize that mixing general and specific statements is important for clarity and engagement of the reader. This balance between general and specific sentences could be highly relevant for science news articles. These articles convey expert research findings in a simple manner to a lay audience. Science writers need to present the material without including too much technical details. On the other hands, an overuse of general statements such as quotes from the researchers and specialists can also lower the reader's attention.

Similarly, conference publications of research findings have been found to have a distinctive structure with regard to general-specific nature (Swales and Feak, 1994). These articles have a hour-glass like structure with general material presented in the beginning and end and specific contents in between. Given that even articles written for an expert audience have such a structure, we expected that the distinction would also be relevant for the genre of science news.

Yet while widely acknowledged to be important for writing quality, there is no work that explores how general and specific sentences can be identified and used for predicting the quality of articles. Our previous work (Louis and Nenkova, 2011a,b) presented the first study on annotating general-specific nature, building an automatic classifier for the aspect and using it for quality prediction. This work was conducted on articles from the news genre. In this section, we provide some details about this prior study and then explain how similar annotations and experiments were performed for articles from our corpus of science news.

In our prior work we focused on understanding the general-specific distinction and obtaining annotated data in order to train a classifier for the task. There was no annotated data for this aspect before our work but we observed that an approximate but suitable distinction has been made in existing annotations for certain types of discourse relations. Specifically, the Penn Discourse Treebank contains annotations for thousands of Instantiation discourse relations which are defined to hold between two adjacent sentences where the first one states a fact and the second provides an example of it. These annotations were created over Wall Street Journal articles which mostly discuss financial news. As a rough approximation, we can consider the first sentence as general and the second as specific. These general and specific sentences from the Instantiation annotations, disregarding pairing information, provided 1403 examples for each type. These sentences also served as development data to observe properties associated with general and specific types. We developed features that captured sentiment, syntax, word frequency, lexical items in these sentences and the combination of features had a 75% success rate for distinguishing the two types on this approximate data.

We then obtained a new held-out test set by eliciting direct annotations from people. We designed an annotation scheme and recruited judges on Amazon Mechanical Turk who marked sentences from news articles. We used news from two sources, Associated Press and Wall Street Journal to obtain sentences for annotation. The annotators had fairly good agreement on this data. We used our classifier trained on discourse relations to predict the sentence categories provided by our judges and our accuracy remained consistent at 75% as on the Instantiations data.

Also in prior work, we successfully used the automatic prediction of sentence specificity to predict writing quality (Louis and Nenkova, 2011b). We obtained general-specific categories on a

corpus of news summaries using our automatic classifier (trained on discourse relations). These summaries were created by multiple automatic summarization systems and the corpus also has manually assigned content and linguistic quality ratings for these summaries by a team of judges. We developed a specificity score that combined the general-specific predictions at sentence level to provide a characterisation for the full summary text. This specificity score was shown to be a significant component for predicting the quality scores on the summaries. The general trend was that summaries that were rated highly by judges showed greater proportion of general content compared to the lower rated summaries.

In the next sections, we explore annotations and classification experiments for the science writing genre.

### 5.1 Annotation guidelines

We chose three articles from the GREAT category of our corpus for annotation. Each article has approximately 100 sentences, providing a total of 308 sentences. The task for annotators was to provide a judgement for a given sentence in isolation, without the surrounding context in which the sentence appeared. We used Amazon Mechanical Turk to obtain the annotations and each sentence was annotated by five different judges. The judges were presented with three options—general, specific and cannot decide. Sentences were presented in random order, mixing different texts.

The judges were given minimal instructions about criteria they should apply while deciding what label to choose for a sentence. Our initial motivation was to identify the types of sentences where intuitive distinctions are insufficient, in order to develop more detailed instructions for manual annotation. The complete instructions were worded as:

> "Sentences could vary in how much detail they contain. One distinction we might make is whether a sentence is general or specific. General sentences are broad statements made about a topic. Specific sentences contain details and can be used to support or explain the general sentences further. In other words, general sentences create expectations in the minds of a reader who would definitely need evidence or examples from the author. Specific sentences can stand by themselves. For example, one can think of the first sentence of an article or a paragraph as a general sentence compared to one which appears in the middle. In this task, use your intuition to rate the given sentence as general or specific."
>
> **Examples:** (*G* indicates general and *S* specific)
>
> [G1] A handful of serious attempts have been made to eliminate individual diseases from the world.
> [G2] In the last decade, tremendous strides have been made in the science and technology of fibre optic cables.
> [G3] Over the years interest in the economic benefits of medical tourism has been growing.
> [S1] In 1909, the newly established Rockefeller Foundation launched the first global eradication campaign, an effort to end hookworm disease, in fifty-two countries.
> [S2] Solid silicon compounds are already familiar–as rocks, glass, gels, bricks, and of course, medical implants.
> [S3] Einstein undertook an experimental challenge that had stumped some of the most adept lab hands of all time–explaining the mechanism responsible for magnetism in iron.

| Agreement | General | Specific |
|---|---|---|
| 5 | 32 (25.6) | 50 (27.7) |
| 4 | 48 (38.4) | 73 (40.5) |
| 3 | 45 (36.0) | 57 (31.6) |
| total | 125 | 180 |
| no majority = 3 | | |

Table 7: The number (and percentage of sentences) for different agreement levels. The numbers are grouped by the majority decision as general or specific.

| | |
|---|---|
| **Agreement 5** | |
| General | At issue is whether the findings back or undermine the prevailing view on global warming. |
| Specific | The sudden popularity of pediatric bipolar diagnosis has coincided with a shift from antidepressants like Prozac to far more expensive atypicals. |
| **Agreement 3** | |
| General | "One thing is almost certain," said Dr. Jean B. Hunter, an associate professor of biological and environmental engineering at Cornell. |
| Specific | The other half took Seroquel and Depakote. |

Table 8: Example sentences for different agreement levels

### 5.2 Annotation results

Table 7 provides the statistics about annotator agreement. We do not compute the standard Kappa scores because different groups of annotators judged different sentences. Instead, we report the fraction of sentences which had 5, 4 or 3 annotators agreeing on the class. When fewer than three annotators agree on the class there is no majority decision, and this situation occurred only for three out of the 308 annotated sentences.[3]

We found that two-thirds of the sentences were annotated with high agreement: Either four or all the five judges gave these sentences the same class. The other one third of the sentences are on the borderline with a slim majority, three out of five judges assigning the same class. Table 8 shows some examples of sentences with full agreement and those that have only agreement of 3. These sentences with low agreement have a mix of both types of content. For example, the general sentence with agreement 3 has a quote the content of which is general, at the same time, the descriptions of the person presents specific information. Such sentences are genuinely hard to annotate. We hypothesize that we can address this problem better by selecting a different granularity such as clauses rather than sentences. Our annotations were also obtained by presenting each sentence individually but another aspect to explore is how these judgements would vary if the context of adjacent sentences is also provided to the annotators. For example, the presence of pronouns, discourse connectives and such references in a sentence taken out of context (for example the last sentence in Table 8) could make the sentence appear more vague but in context, they may be interpreted differently.

In terms of distribution of general and specific sentences, Table 7 shows that there were more specific (59%) sentences than general on this dataset. We expect that this distribution would vary by

3. For the three sentences that did not obtain majority agreement, two of the judges had picked general, two chose specific and the fifth label is 'cannot decide'.

| Transition type | Number (%) |
|---|---|
| GS | 66 (22) |
| GG | 58 (19) |
| SS | 114 (38) |
| SG | 63 (21) |

Table 9: Number and percentage of different transition types. *G* indicates general and *S* specific.

| Block type | G1 | G2 | GL | S1 | S2 | SL |
|---|---|---|---|---|---|---|
| Number | 35 | 19 | 13 | 21 | 17 | 28 |
| % | 26 | 14 | 10 | 16 | 13 | 21 |

Table 10: Number and percentage of different block types. G1, G2 and GL indicate general blocks of size 1, 2 and >=3. Similarity for the specific category.

genre. In the annotations that we have collected previously on Wall Street Journal and Associated Press articles, we found Wall Street Journal articles to contain more general than specific content while the Associated Press corpus collectively had more specific than general (Louis and Nenkova, 2011a). This finding could result from the fact that our chosen WSJ articles are written with a more 'essay-like' style, for example one article discusses the growing use of personal computers in Japan, while the AP articles were short news reports.

A simple plot of the sentence annotations for the three articles (Figure 1) shows how the annotated general and specific sentences are distributed in the articles. The X axis indicates sentence position (starting from 0) normalized by the maximum value. On the Y axis, 1 indicates that the sentence was annotated as specific (any agreement level) and -1 otherwise. These plots show that often there are large sections of continuous specific content and the general sentence switch happens between these blocks. Some of these specific content blocks on examination tended to be detailed research descriptions. An example is given below. In this snippet, a general sentence is followed by 8 consecutive specific sentences. This finding is interesting because it shows that research details are padded with general content providing the context of the research.

> *(General) Still, no human migration in history would compare in difficulty with reaching another star.* The nearest, Alpha Centauri, is about 4.4 light-years from the Sun, and a light-year is equal to almost six trillion miles. The next nearest star, Sirius, is 8.7 light-years from home. To give a graphic sense of what these distances mean, Dr. Geoffrey A. Landis of the NASA John Glenn Research Center in Cleveland, pointed out that the fastest objects humans have ever dispatched into space are the Voyager interplanetary probes, which travel at about 9.3 miles per second.
>
> "If a caveman had launched one of those during the last ice age, 11,000 years ago," Dr. Landis said, "it would now be only a fifth of the way toward the nearest star."
>
> ... (5 more specific sentences)

These patterns are also indicated quantitatively in Tables 9 and 10. Table 9 gives the distribution of transitions between general and specific sentences, calculated from pairs of adjacent sentences only. We excluded the three sentences that did not have a majority decision. In this table, $G$ stands for general, $S$ for specific and the combinations indicate a pattern between adjacent sentences. The most frequent transition is *SS*, specific sentence immediately followed by another specific sentence,

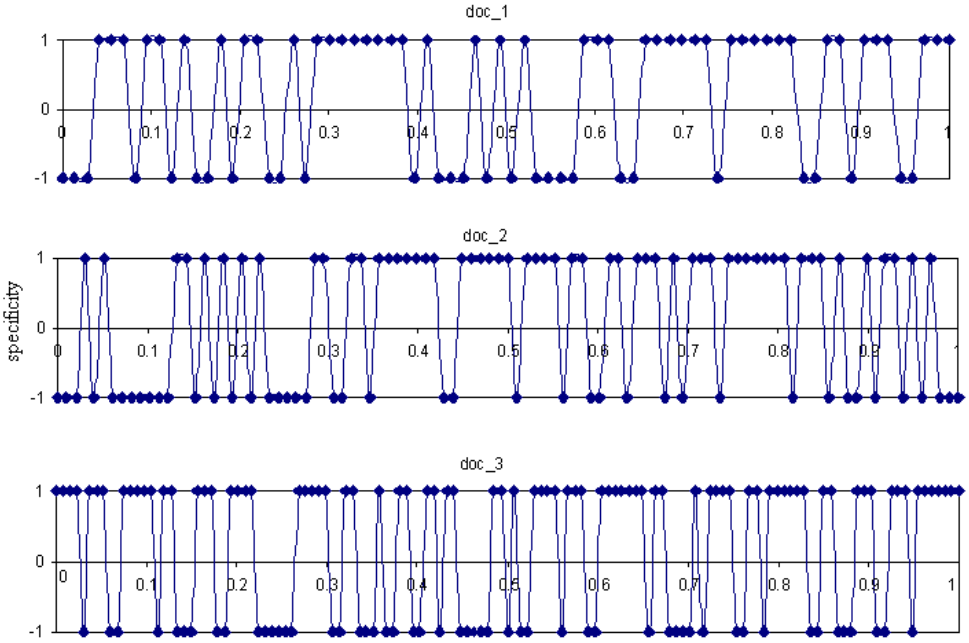


Figure 1: Specificity of annotated sentences

which accounts for 40% of the total transitions. Other types are distributed almost equally, 20% each. Table 10 shows the sizes of contiguous blocks of general and specific sentences. The minimum block size is one whereas the maximum turned out to be 7 for general and 8 for specific.

### 5.3 Automatic prediction of specificity

This section explains the classifier that we developed for predicting sentence specificity and results on the data annotated from the science news articles. Our classifier trained on the Instantation relations data obtains 74% accuracy in making the binary general or specific prediction on the science news data we described in the previous section. This high accuracy has enabled us to create specificity predictions on a larger scale for our full corpus and analyze how the aspect is related to text quality. The experiment on text quality is reported in the next section.

Our features use different surface properties that indicate specificity. For example, one significant feature indicative of general sentences is plural nouns which often refer to classes of things. Other features involve word-level specificity, counts of named entities and numbers, and likelihood under language model (general sentences tend to have lower probability than specific under a language model of the domain), and binary features indicating the presence of each word. These features are based on syntactic information as well as ontology such as WordNet. For example, the word-level specificity feature is computed using the path length from the word to root of WordNet via hypernym relations. Longer paths could be indicative of a more specific word, while words that are closer to the root could be related to more general concepts. We also add features that capture the length of adjective, adverb, preposition and verb phrases. To obtain the syntax features, we parsed the sentences using the Stanford Parser (Klein and Manning, 2003). We call the word features *lexical* and the set of all other features are *non-lexical*. Descriptions of these features and their

| Example type | Size | Lexical | Non-lexical | All features |
|---|---|---|---|---|
| Agree 5 | 82 | 74.3 | 92.2 | 82.9 |
| Agree 4+5 | 203 | 66.0 | 85.2 | 76.3 |
| Agree 3+4+5 | 305 | 58.3 | 74.4 | 67.2 |

Table 11: Accuracies for automatic classification of sentences on sentences with different agreement levels. Agree 4+5 indicates that sentences with both agreement 5 and 4 were combined. Agree 3+4+5 indicates all the annotated examples.

individual strengths are given in Louis and Nenkova (2011a). Here we report the results on categories of features (Table 11) on the science news annotations (302 sentences). (We excluded three sentences that did not have majority judgement.) We used a logistic regression classifier available in R (R Development Core Team, 2011) for our experiment.

We take the majority decision by annotators as the class of the sentences in our test set. Specific sentences comprise 59% of the data and so the majority class baseline would give 59% accuracy. The accuracy of our features is given in Table 11. For this test dataset (all sentences where a majority agreement was obtained) the accuracies are indicated in the last line of the table. The best accuracy is obtained by the non-lexical features, 74.4%. The word features are much worse. One reason for poor performance of the word features could be that our training sentences are taken from articles published in the Wall Street Journal. The lexical items from these finance news articles may not appear in science news from the New York Times leading to low accuracy. The lexical features could also suffer from data sparsity and need large training sets to perform well. On the other hand, the non-lexical indicators appear to carry over smoothly. The performance of these features on science news is close to the accuracy when the WSJ trained classifier was tested on held-out sentences also from WSJ. The combination of lexical and non-lexical features is not better than the non-lexical features alone.

When we consider examples with higher agreement, rows 1 and 2 of the table, the results for all feature classes are higher, reaching even 92.2% for the non-lexical features. In further analyses, we found that the confidence from the classifier is also indicative of the agreement from annotators. When the classifier made a correct prediction, the confidence from the logistic regression classifier is higher on examples which had more agreement. On examples on the borderline between general and specific classes, even when the prediction was correct, the confidence was quite low. The trend for wrong examples was opposite. Examples with high agreement were predicted wrongly with lower confidence whereas the borderline cases had high confidence even when predicted wrongly. This pattern suggested that the confidence values from the classifier can be used as generality/specificity scores in addition to the binary distinction.

### 5.4 Relationship to writing quality

Since automatic prediction could be done with very good accuracies, we used our classifier to analyze the different categories of articles in our corpus. We trained a classifier on data derived from the discourse annotations and using the set of all non-lexical features. We obtained the predictions from the classifier for each sentence in our corpus and then composed several features to indicate specificity scores at article level. These features are described below.

| Feature | Mean value in category | | P-value from t-test |
|---|---|---|---|
| | VERY GOOD | TYPICAL | |
| SPEC_WTD | 0.56 | 0.58 | 0.004 |
| SL | 0.09 | 0.10 | 0.035 |

Table 12: Features related to general and specific content which had significantly different mean values in the VERY GOOD and TYPICAL categories. The p-value from a two-sided t-test is also indicated.

**Overall specificity:** One feature indicates the fraction of specific sentences in the article (SPEC_FRAQ). Since we use a logistic regression classifier for predicting specificity, we also obtain from the classifier a confidence value for the sentence belonging to each class. We use the confidence of each sentence belonging to the 'specific' class and compute the mean and variance of this confidence measure across the article's sentences as features (SPEC_MEAN, SPEC_VAR). But these scores do not consider the lengths of the different sentences. A longer sentence constitutes a greater portion of the content of the article and its general or specific nature should have higher weight while composing the score for the full article. For this purpose, we also add another feature—the weighted average of the confidence values where the weights are the number of words in the sentence (SPEC_WTD).

**Sequence features:** Proportions of different transition and block types that we discussed in Section 5.2 are added as features (GG, GS, SS, SG, G1, G3, GL, S1, S2, SL). In contrast to the manual markings in that section, here the transitions are computed using the automatic annotations from the classifier.

We first tested how these features vary between good and average writing using a random sample of 1000 articles taken from the GREAT and VERY GOOD categories combined and another 1000 taken from the TYPICAL category. No pairing information from the ranking corpus was used during this sampling as we wanted to test overall if these features are indicative of good articles rather than their variation within a particular topic. A two-sided t-test between the feature values in the two categories showed the mean values for two of the features to vary significantly (p-value less than 0.05). The test statistic and mean values for the significant features are shown in Table 12.

The good articles have a lower degree of specific content (measured by the confidence weighted score SPEC_WTD). The other trend was a higher proportion of large specific blocks (SL) in the typical articles compared to the good ones. Both these features indicate that more specific content is correlated with the TYPICAL articles. None of the transition features were significantly different in these two sets.

All the features (significant and otherwise) were then input to a classifier that considered a pair of articles and decided which article is the better one. The features for the pair are the difference in feature values for the two individual articles. A random baseline would be accurate 50% of the time. We used the test set described in Section 4 and performed 10-fold cross validation over it using a logistic regression classifier. The accuracy using our features is 54.5% indicating that the general-specific distinction is indicative of text quality. The improvement is low but statistically significant. When the features were analyzed by category, the specificity levels gave 52% and the transition/block features gave 54.4% accuracy. So both aspects are helpful for making the general-specific distinction but the transition features are stronger than overall specificity levels.

The combination of the two feature types is not much different than transition features alone. But we expect both feature sets to be useful when combined with other aspects of writing not related to specificty.

These results provide evidence that content specificity and the sequence of varying degrees of specificity are predictive of writing quality. They can be annotated with good agreement and we can design automatic ways to predict the distinction with high accuracy.

## 6. Communicative goals of the article

Our second line of investigation concerns the intentions or communicative goals behind an article. This work is based on the idea that an author has a purpose for every article that he writes and the structure that he chooses for organizing the article's content is such that it will help him to convey his purpose. A detailed theory of intentional structure and its role in the coherence of articles was put forth in an influential study by Grosz and Sidner (1986). The theory has two proposals: that the overall intention for an article is conveyed by the combined intentions of smaller level discourse segments in the article and that the smaller segments are linked by relations which combine their purposes. Since the organization of low-level intentions in the article contributes to the coherence by which the overall purpose is conveyed, if a method of identifying intentions for small text segments is developed we can use it for text quality prediction. This section describes our attempt to develop an annotation method for sentence-level communicative goals in science news articles and the usefulness of these annotations for predicting well-written articles.

For the genre of science writing, the author's goal is to convey information about a particular research study, its relevance and impact. At a finer level in the text, many different types of sentences contribute to achieving this purpose. Some sentences define technical terms, some provide explanations and examples, others introduce the people and context of the research and some sentences are descriptive and present the results of the studies. Science articles are also sometimes written as a story and this style of writing creates sentences associated with the narrative and dialog in the story. There are numerous books on science writing (Blum et al., 2006; Stocking, 2010) that prescribe ideas for producing well-written articles in different styles—narratives, explanatory pieces and interviews. Such books often point to examples of specific sentences in articles that convey an idea effectively. For example, we might expect that good science articles are informative and contain more sentences that define and explain aspects of the research. Similarly, well-written articles might also provide examples after a definition or explanation. So we expect that a large scale annotation of sentences with communicative goals and building classifiers to automatically do such annotations can be quite useful for text quality research.

Studies on register variation (Biber, 1995; Biber and Conrad, 2009) provide evidence that there is noticeable variation in linguistic forms depending on the communicative purpose and situations. There have also been successful efforts to manually annotate and develop automatic classifiers for predicting communicative goals on the closely related genre of academic writing ie. conference and journal publications. These articles have a well-defined purpose, also for different sections of the paper. Studies in this area have identified that conference publications have a small set of goals such as motivation, aim, background, results and so on (Swales, 1990; Teufel et al., 1999; Liakata et al., 2010). Such well-definable regularities in academic writing have lead to studies that manually annotated these sentence types and developed supervised methods to do the annotation

| | |
|---|---|
| Definition | In this theory, a black hole is a tangled melange of strings and multidimensional membranes known as "D-branes." |
| Example | To see this shortcoming in relief, consider an imaginary hypothesis of intelligent design that could explain the emergence of human beings on this planet. |
| Explanation | If some version of this hypothesis were true, it could explain how and why human beings differ from their nearest relatives, and it would disconfirm the competing evolutionary hypotheses that are being pursued. |
| Results | The new findings overturn almost everything that has been said about the behavior and social life of the Mandrillus sphinx to date, and also call into question existing models of why primates form social groups. |
| About people | A professor in Harvard's department of psychology, Gilbert likes to tell people that he studies happiness. |
| Story line | "Happiness is a seven," she said with a triumphant laugh, checking the last box on the questionnaire. |

Table 13: Examples for some sentence types from the science news corpus

automatically (Teufel and Moens, 2000; Guo et al., 2011). Subsequently, these annotations have been used for automatic summarization and citation analysis tasks.

However, our annotation task for science journalism is made more difficult by the wider and unconstrained styles of writing in this genre. The number of communicative goals is much higher and varied compared to academic writing. In addition, academic articles are also structured into sections whereas there is no such subdivision in the essay-like articles in our genre. Perhaps the most difficult aspect is that even coming up with a initial set of communicative goals to annotate is quite difficult. Table 13 shows examples for some of the sentence types we wish to annotate for our corpus based on our intuitions. But these categories only capture some of the possible communicative goals.

As an alternative to predefining catgories of interest, we explored an unsupervised method to group sentences into categories. Our method relies on the idea that similarities in syntactic structure could indicate similarities in communicative goals (Louis and Nenkova, 2012a). For example, consider the sentences below.

A. An aqueduct is a water supply or navigable channel constructed to convey water.
B. A cytokine receptor is a receptor that binds cytokines.

Both sentences A and B are definitions. They follow the regular pattern for definitions which consists of the term to be defined followed by a copular predicate which has a relative or reduced relative clause. In terms of syntax, these two sentences have a high degree of similarity even though their content is totally different. Other sentences types such as questions also have a unique and well-defined syntax.

We implemented this hypothesis in the form of a model that performs syntactic clustering to create categories and assigns labels to sentences indicating which category is closest to the syntax of the sentence (Louis and Nenkova, 2012a). Then the sequence of sentences in an article can be analyzed in terms of these approximate category label sequences. We ran our approach on a corpus of academic articles. We showed that the sequence of labels in well-written and coherent articles is different from that in an incoherent article (created artificially by randomly permuting the sentences in the original article). We also found that the automatic categories created by our model have some

correlation to categories used in the manual annotations of intentional structure for the academic genre.

Given this success on academic articles, we expected that syntactic clustering could help us uncover some of the approximate categories prevalent in science journalism. Such a method is in fact more suitable for our genre where it is much harder to define the intention categories.

Our plan for using the approximate categories is two-fold. Firstly, we want to test how much these approximate categories would align with our intuitions about sentence types in this genre and also how accurately we can predict text quality using sequences of these categories. On the other hand, we expect this analysis to help create guidelines for a fuller manual annotation of communicative goals on these articles. The clusters would give us a closer view of different types of sentences present in the articles and enable us to define categories we should annotate. In addition, when a sentence is annotated, annotators can look at similar sentences from other articles using these clusters and these examples would help them to make a better decision on the category.

### 6.1 Using syntactic similarity to identify communicative goals

Our approach uses syntactic similarity to cluster sentences into groups with (hopefully) the same communicative goal. The main goal is to capture these categories and also obtain information about what are the likely sequences of such categories in good and typical articles. To address this problem, we employ a coherence model that we have recently developed (Louis and Nenkova, 2012a).

Our implementation uses a Hidden Markov Model (HMM). In the HMM, each state represents the syntax of a particular type of sentence and transitions between two states $A$ and $B$ indicates how likely it is for a sentence of type $B$ to follow a sentence of type $A$ in articles from the domain. We train the model on examples of good writing and learn the sentence categories and transition patterns on this data. We then apply this model to obtain the likelihood of new articles. Poorly-written articles would obtain a lower probability compared to well-written articles.

Accordingly, for our corpus, we use the articles from the best category (GREAT) for building the model. We also need to define a measure of similarity for clustering the sentences. Our similarity measure is based on the constituency parse of the sentence. We represent the syntax of each sentence as a set of context free rules (productions) from its parse tree. These productions are of the form LHR $\rightarrow$ RHS where LHS is a non-terminal in the tree whose children are the set of non-terminals and terminal nodes indicated by RHS. We compute features based on this representation. Each production serves as a feature and the feature value for a given sentence is the number of times that production occurred in the parse of the sentence. We obtained the parse trees for the sentences using the Stanford Parser (Klein and Manning, 2003). All sentences from the set of articles are clustered according to these syntactic representations by maximizing the average similarity between items in the same cluster. These clusters form the states of the HMM. The emissions from a state are given by a unigram language model computed from the productions of the sentences clustered into that state. Transitions between states are computed as a function of number of sentences from cluster $A$ that follow those of cluster $B$ in the training articles.

We also need to decide how many clusters are necessary to adequately cover the sentences types in a collection of articles. We pick the number of hidden states (sentence types) by searching for the value that provides the best accuracy for differentiating good and poorly-written articles in our development set (see Section 4). Specifically, for each setting of the number of clusters (and other

model parameters) we obtained the perplexity of the good and typical articles under the model. The article with lower perplexity is considered as the one predicted as better written. We use perplexity rather than probability to avoid the influence of article length on the score. This development set contained 500 pairs of GOOD and TYPICAL articles. The smoothing parameters for emission and transition probabilities were also tuned on this development data. The parameters that gave the best accuracy were selected and the final model trained using these settings.

## 7. Analysis of sentence clusters

There are 49 clusters in our final model trained on the best articles. We manually analyzed each cluster and marked whether we could name the intended communicative goal on the basis of example sentences and descriptive features. The descriptive features were computed as follows. Since the sentences were clustered on the basis of their productions and a cosine similarity function was used, we selected those productions which contribute to more than 10% of the average intra-cluster cosine similarity.

The intrepretability of the clusters varied greatly. There were 27 clusters for which we could easily provide names. Some of the remaining clusters also had high syntactic similarity, however we could not easily interpret them as related to any communicative goal. So we have not included them in our discussion. In this section, we provide descriptions and examples for sentences in each of these clusters. We ignore the remaining clusters for now but their presence indicates that further in-depth analyses of sentences types must be conducted in order to be able to create annotations that cover all sentences. An overview of the 27 clusters is given in Tables 14 and 15. We also indicate the descriptive features and some example sentences for each cluster. For the example sentences, we have also marked the span covered by the LHS of each descriptive production. These are indicated with '[' and ']' braces attached with the label of the LHS non-terminal.

We also divide these clusters into categories for discussion. Some clusters are created such that the sentences in them have matching productions that cover long spans of texts in the sentences. The descriptive features for these clusters are productions that involve the full sentence or complete predicates of sentences. Such clusters are in contrast to others where the sentences are similar but only based on productions that involve rather small spans of text. For example the descriptive productions for these clusters could involve smaller syntactic categories such as noun or prepositional phrases. We provide a brief study of our clusters below.

**Sentence-level matches.** Clusters *a*, *b* and *c* correspond to quotes, questions and fragments (Table 14) respectively. The match between the sentences in all these clusters is on the basis of sentence level productions as indicated by the most descriptive features. For example, ROOT→SQ the descriptive production for cluster 2 is a topmost level production which groups the yes/no questions. For quotes-related clusters, the descriptive features have quotation marks enclosing an $S$ node or a NP VP sequence. An example sentence from this category is shown below.

ROOT-[Is there one way to get Parkinson's, or 50 ways? "]-ROOT

Category *c* has many sentences which are quite short and have a style that would be unconventional for writing but would be more likely in dialog. They are fragment sentences that create a distinctive rhythm: "In short, no science"; "And the notion of mind?"; "After one final step, this."

These clusters were easiest to analyze and directly matched our intuitions about possible communicative goals.

**Predicate-level match.** For clusters in categories *d*, *e*, *f* and *g*, the descriptive features often cover the predicate of the sentences.

For clusters in category *d* several of these predicates are adjectival phrases as in the following example.

> The number of applications VP-[is ADJP-[mind-boggling]-VP]-ADJP .

The primary goal of these sentences is to describe the subject and in our examples these phrases often provided opinion and evaluative comments. Other clusters in this category have clauses headed by 'wh' words which also tend to provide descriptive information.

Category *e* has a number of clusters where the main verb is in past tense but with different attributes. For example, in cluster 12, the predicate has only the object or the predicate has a prepositional phrase in addition to the object as in cluster 11. Most of these sentences (two are shown below) tended to be part of the narrative or story-related parts of the article.

> (i) He VP-[entered the planet-searching business through a chance opportunity]-VP .
> (ii) S-[He VP-[did a thorough physical exam]-VP and VP-[took a complete history]-VP .]-S

Categories *f* and *g* comprised sentences with adverbs, discourse connectives or modal verbs. These sentences show how comparisons, speculations and hypotheses statements are part of motivating and explaining a research finding. An example is shown below.

> If a small fraction of the subuniverses VP-[can support life]-VP , then there is a good chance that life VP-[will arise somewhere]-VP, Dr. Susskind explained.

**Entity or verb category matches.** In categories *h* and *i*, we have listed some of the clusters where the descriptive features tended to be capturing only one or two words at most. So the descriptive feature does not relate the sentences by their overall content but are based on a phrase such as NP.

The two clusters under category *h* have proper names as the descriptive features as in the sentence below.

> Dr. Strominger remembered being excited when he found a paper by the mathematician NP-[Dr. Shing-Tung Yau]-NP , now of Harvard and the Chinese University of Hong Kong .

These sentences have either two or three proper name tokens in sequence and were often found to introduce a named entity which is why they are referred to using their full names. These sentences often have an appositive following the name which provides more description about the entity.

Other clusters group different types of noun phrases and verbs. For example, cluster 27 groups sentences containing numbers and 42 has sentences with plural nouns. The top descriptive feature for cluster 15 is present tense verbs. Some examples from these clusters with their descriptive features are shown below.

> (i) The big isolated home is what Loewenstein , NP-[48]-NP , himself bought.
> (ii) The news from NP-[environmental organizations]-NP is almost always bleak.
> (iii) The first departments we VP-[checked]-VP were the somatosensory and motor regions.
> (iv) Periodically, he went to Methodist Hospital for imaging tests to measure NP-[the aneurysm's size]-NP .

While some intrepetation can be provided for the communicative goals corresponding to these sentences, these clusters are somewhat divergent from our idea of detecting communicative goals. The similarity between sentences in these clusters is based on a rather short span such as a single noun phrase and less likely to give reliable indications of the intention of the full sentence.

Overall the clusters discovered by the automatic approach were meaningful and coherent, however, they did not not match our expected communicative goals from Table 13. While we expected several categories related to explaining a research finding such as definition, examples and motivation, we did not find any clearly matching clusters that grouped such sentences. We hypothesize that this problem could arise because several science articles embed research related sentences as part of a larger narrative or story and information about scientists and other people. Hence the input to clustering is a diverse set of sentences and clustering places research and non-research sentences in the same cluster. This issue can be addressed by separating out sentences which describe research findings from those that are related to the story line as a first step. The clusters can then be built on the two sets of sentences separately and we expect their output to be closer to the types we had envisioned. We would then obtain clusters related to the core research findings that are reported and a separate set of clusters related to the story line of the articles. We plan to explore this direction in our future work. However, we believe that this experiment has provided good intuitions and directions for further annotation of communicative goal categories for our corpus.

### 7.1 Accuracy in predicting text quality

Despite the fact that the clusters are different from what we expected to discover, we evaluate if these automatic and approximate sentence types contribute to the prediction of text quality for our corpus. Good results on this task would provide more motivation for performing large scale annotation of communicative goals. We describe our experiments in this section and show that positive results were obtained using these approximate sentence categories.

For each article in our test set (details in Section 4), we obtained the most likely state sequence under the HMM model using the Viterbi decoding method. This method assigns a category label to each sentence in the article such that the sequence of labels is the most likely one for the article under the model. Then features were computed using these state predictions. Each feature is the proportion of sentences in the article that belong to a state.

We first analyzed how these features vary between a random set of 1000 good articles and a set of 1000 typical articles, without topic pairing. A two-sided t-test was used to compare the feature values in the two categories. Table 16 shows the 10 clusters whose probability was significantly different (p-value less than 0.05) between the good and typical categories.

| Clusters | Descriptive features | Example sentences |
|---|---|---|
| **a. Quotes:** Sentences from dialog in the articles | | |
| 1 | S→NP VP . ” | S-[We were sometimes lax about personal safety . ”]-S |
| 6 | S→” NP VP . ” | S-[“That ’s one reason people are interested in time travel.”]-S |
| 10 | S→” S , ” NP VP . ; VP→VBD | S-[” I thought Ben made an error , ” he VP-[said]-VP .]-S |
| **b. Questions:** Direct questions, some are rhetorical, some are yes/no type | | |
| 2 | ROOT→SQ | ROOT-[Would they know about stars?]-ROOT |
| 3 | ROOT→SBARQ ; SBARQ→WHADVP SQ . | SBARQ-[ROOT-[Why are atoms so tiny and stars so big ?]-]ROOT]-SBARQ |
| **c. Conversation style:** Sentences that are unconventional in written texts but more likely in dialog | | |
| 4 | ROOT → FRAG | ROOT-[For what ?]-ROOT |
| **d. Descriptions:** The predicate is mainly intended to provide description often containing evaluative comments and opinion | | |
| 5 | VP→VBD ADJP ; ADJP→JJ | She knew only that it VP-[was ADJP-[contagious]-VP]-ADJP and that it VP-[was ADJP-[nasty]-VP]-ADJP . |
| 7 | VP→VBZ ADJP ADJP→JJ | Emotion VP-[is central to wisdom]-VP, yet detachment VP-[is ADJP-[essential]-VP]-ADJP . |
| 13 | ADJP→JJ S | And yet as psychologists have noted, there is a yin-yang to the idea that makes it ADJP-[difficult to pin down]-ADJP . |
| 20 | VP→VBD NP ; SBAR→WHNP S | Later, Ardelt settled on 39 questions SBAR-[that, in her judgment, VP-[captured the elusive concept of wisdom]-VP]-SBAR |
| 29 | ADJP→JJ ; VP→VBP ADJP | Cosmologists have found to their astonishment that life and the universe VP-[are ADJP-[strangely]-ADJP and deeply connected]-VP . |
| 30 | SBAR→WHADVP S ; WHADVP→WRB | The problem, as Gilbert and company have come to discover, is that we falter WHADVP-[SBAR-[when]-WHADVP it comes to imagining WHADVP-[SBAR-[how]-WHADVP we will feel about something in the future]-SBAR]-SBAR . |
| **e. Story line:** These clusters mostly capture events associated with the story presented in the article | | |
| 11 | VP→VBD NP PP | We VP-[saw activity in the hippocampus]-VP . |
| 12 | VP→VBD NP ; S→NP VP . | The more complex the task, the more VP-[dispersed the brain ’s activity]-VP . |

Table 14: Example clusters discovered by syntactic similarity. Multiple descriptive features for a cluster are separated by a ‘;’.

| Clusters | Descriptive features | Example sentences |
|---|---|---|
| **f. Discourse relations / negation:** sentences that are part of comparison and expansion relations and those having adverbs | | |
| 31 | VP→VBZ RB VP ; S→CC NP VP . | S-[But the picture, as most good evolutionary psychologists point out, is more complex than this .]-S |
| **g. Modals / adverbs:** speculations, hypothesis. Sometimes adverbs are used with the modals increasing the intensity | | |
| 40 | VP→MD VP ; SBAR→IN S | The next big milestone, astronomers say, VP-[will be the detection of Earth-size planets, SBAR-[although that VP-[will require going to space]-VP]-VP]-SBAR . |
| 41 | VP→VB NP ; VP→TO VP | A few alien bacteria in a mud puddle someplace would VP-[change science]-VP . |
| 46 | VP→MD ADVP VP ; ADVP→RB | It is about a kind of energy we ADVP-[often]-ADVP rue but VP-[would ADVP-[surely]-ADVP miss.]-VP |
| **h. About named entities:** sentences where the main focus is a person or other named entity | | |
| 14 | NP→NNP NNP NNP | "It touches on philosophical issues that scientists oftentimes skirt," said NP-[Dr. John Schwarz]-NP , a physicist and string theorist at the California Institute of Technology . |
| 45 | NP→NNP NNP | NP-[Dr. Lomborg]-NP also takes issue with some global warming predictions. |
| **i. Others:** noun phrases with different attributes, different types of verb phrases | | |
| 27 | NP→CD | It also made me curious about Clayton, who disappeared from academia in NP-[1981]-NP . |
| 39 | NP→NNP | NP-[Ben]-NP was once a mail carrier and a farmer and cattle rancher. |
| 26 | NP→NP NN ; NP→NNP POS | Dr. Hawking 's celebrated breakthrough resulted partly from a fight . |
| 34 | NP→JJ NN | To date , the proponents of NP-[intelligent design]-NP have not produced anything like that. |
| 42 | NP→JJ NNS | NP-[Tropical forests]-NP are disappearing. |
| 15 | VP→VBZ | Ardelt acknowledges that no one really knows what wisdom VP-[is]-VP . |
| 32 | VP→VBD | But Dr. DeBakey 's rescue almost never VP-[happened]-VP . |

Table 15: Example clusters discovered by syntactic similarity (continued from Table 14)). Multiple descriptive features for a cluster are separated by a ';'.

| Cluster no. | Type | Mean proportion in category | | p-value from t-test |
|---|---|---|---|---|
| | | VERY GOOD | TYPICAL | |
| | **Higher probability in good articles** | | | |
| 5 | descriptions | 0.004 | 0.003 | 0.038 |
| 30 | descriptions | 0.085 | 0.079 | 0.001 |
| 40 | modals | 0.028 | 0.026 | 0.014 |
| 34 | others (nouns with adjective modifiers) | 0.019 | 0.018 | 0.036 |
| | **Lower probability in good articles** | | | |
| 6 | quotes | 0.003 | 0.004 | 0.021 |
| 46 | modals and adverbs | 0.004 | 0.005 | 0.003 |
| 15 | other (present tense verbs) | 0.002 | 0.003 | 0.018 |
| 27 | other (numbers) | 0.022 | 0.025 | 0.003 |
| 39 | other (proper names) | 0.014 | 0.016 | 0.024 |
| 42 | other (plural nouns with adjective modifiers) | 0.039 | 0.043 | 0.003 |

Table 16: Clusters that have significantly different probabilities in good and typical articles. The mean value for the cluster label in the good and typical articles and the p-value from a two-sided t-test comparing these means are also indicated.

Four states occurred with higher probability in good writing. Two of them are from the description type category that we identified. Particularly, in cluster 5, as we already noted, several of the adjective phrases provided evaluative comments or opinion. Such sentences are more prevalent in good articles compared to the typical ones. Sentences that belong to cluster 34 also have a descriptive nature, however, only at noun phrase level. Cluster 30 groups sentences which provide explanation in the form of 'wh' clauses. The prevalence of modal containing sentences (cluster 30) indicates that hypothesis and speculation statements are also frequently employed. Most of these clusters appear to be associated with descriptions and explanations.

On the other hand, there are six states that appeared with significantly greater probability in the typical articles. One of these clusters is state 6 depicting quotes. Several of the other clusters come from the category that we grouped as 'other' class in Table 15. These are clusters whose descriptive features are numbers, proper names, plural nouns and present tense verbs. Even though the descriptive features for these clusters mostly indicated properties at noun phrase or verb level, we find that they provide signals that can distinguish good from typical writing. Further annotation of the properties of such sentences could provide more insight into why these sentences are not preferred.

Since these still approximate categories are significantly differently distributed in very good and typical categories, we also examined the use of features derived from them for pairwise classification. For each article, we calculate the proportion of sentences labelled with a particular state label. Similar to the experiments for general-specific content, the features for a pair of articles are computed as the difference in feature values of the two constituent articles. We evaluated these features using 10-fold cross validation on the test corpus we described in Section 4. We obtained an accuracy of 58%. While low, the accuracy is significantly higher than baseline giving motivation that sentence types could be valuable indicators of quality. We plan to experiment with other ways of computing similarity and also create manually annotate articles for communicative goals which will enable us to develop supervised methods to do the intention classification.

## 8. Closing discussion

We have presented a first corpus of text quality ratings in the science journalism genre. We have shown that existing judgments of good articles from New York Times can be combined with the New York Times corpus to create different categories of writing quality. The corpus can be obtained from our website.[4]

There are several extensions to the corpus which we plan to carry out. All the three categories GREAT, VERY GOOD and TYPICAL involve writing samples from professional journalists working for the New York Times. So these articles are overall nicely written and the distinctions that we make are aimed at discovering and characterizing truly great writing. This aspect of the corpus is advantageous because it allows researchers to focus on the upper end of the spectrum of writing quality, unlike work dealing with student essays and foreign language learners.

However, it would also be useful to augment the corpus with further levels of inferior writing using the same topic matching method that we used for our ranking corpus. One level can be articles from other online magazines—*average* writing. In addition we can elicit essays from college students on similar topics to create a *novice* category. Such expansion of the corpus will further enable us to identify how different aspects of writing change along the scale.

We believe that the corpus will help researchers to focus on new and genre-specific measures of text quality. In this work, we have presented preliminary ideas about how annotations for new aspects of text quality can be carried out and showed that automatic metrics for these factors can also be built and are predictive of text quality. In future work, we plan to build computational models for other aspects that are unique to science writing and which can capture why certain articles are considered more clearly written, interesting and concise. These include identifying figurative language (Birke and Sarkar, 2006) and metaphors (Fass, 1991; Shutova et al., 2010), and work that aims to produce sentences that describe an image (Kulkarni et al., 2011; Li et al., 2011). We wish to test how much these genre-specific metrics would improve prediction of text quality in addition to regular readability type features. Moreover, we can expect that the strength of these metrics would vary for novice versus great writing.

## Acknowledgements

We would like to thank the anonymous reviewers of our paper for their thoughtful comments and suggestions. This work is partially supported by a Google research grant and NSF CAREER 0953445 award.

## References

E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of WSDM*, pages 183–194, 2008.

G.J. Alred, C.T. Brusaw, and W.E. Oliu. *Handbook of technical writing*. St. Martin's Press, New York, 2003.

Y. Attali and J. Burstein. Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning and Assessment*, 4(3), 2006.

4. http://www.cis.upenn.edu/~nlp/corpora/scinewscorpus.html

R. Barzilay and M. Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008.

R. Barzilay and L. Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of NAACL-HLT*, pages 113–120, 2004.

D. Biber. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press, 1995.

D. Biber and S. Conrad. *Register, genre, and style*. Cambridge University Press, 2009.

J. Birke and A. Sarkar. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL*, 2006.

D. Blum, M. Knudson, and R. M. Henig, editors. *A field guide for science writers: the official guide of the national association of science writers*. Oxford University Press, New York, 2006.

J. Burstein, M. Chodorow, and C. Leacock. Criterion online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, 2003.

K. Collins-Thompson and J. Callan. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT-NAACL*, pages 193–200, 2004.

K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag. Personalizing web search results by reading level. In *Proceedings of CIKM*, pages 403–412, 2011.

E. Dale and J. S. Chall. A formula for predicting readability. *Edu. Research Bulletin*, 27(1):11–28, 1948.

R. De Felice and S. G. Pulman. A classifier-based approach to preposition and determiner error correction in l2 english. In *Proceedings of COLING*, pages 169–176, 2008.

M. Elsner, J. Austerweil, and E. Charniak. A unified local and global model for discourse coherence. In *Proceedings of NAACL-HLT*, pages 436–443, 2007.

D. Fass. met*: a method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17:49–90, March 1991.

L. Feng, N. Elhadad, and M. Huenerfauth. Cognitively motivated features for readability assessment. In *Proceedings of EACL*, pages 229–237, 2009.

R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32:221 – 233, 1948.

M. Gamon, J. Gao, C. Brockett, A. Klementiev, W. B. Dolan, D. Belenko, and L. Vanderwende. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of IJCNLP*, 2008.

B. Grosz and C. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 3(12):175–204, 1986.

R. Gunning. *The technique of clear writing*. McGraw-Hill; Fouth Printing edition, 1952.

Y. Guo, A. Korhonen, and T. Poibeau. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of EMNLP*, pages 273–283, 2011.

M. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of HLT-NAACL*, pages 460–467, 2007.

D. Higgins, J. Burstein, D. Marcu, and C. Gentile. Evaluating multiple aspects of coherence in student essays. In *Proceedings of HLT-NAACL*, pages 185–192, 2004.

N. Karamanis, C. Mellish, M. Poesio, and J. Oberlander. Evaluating centering for information ordering using corpora. *Computational Linguistics*, 35(1):29–46, 2009.

J. Y. Kim, K. Collins-Thompson, P. N. Bennett, and S. T. Dumais. Characterizing web content, user interests, and search behavior by reading level and topic. In *Proceedings of WSDM*, pages 213–222, 2012.

D. Klein and C.D. Manning. Accurate unlexicalized parsing. In *Proceedings of ACL*, pages 423–430, 2003.

G. Kulkarni, V. Premraj, S. Dhar, Siming Li, Yejin Choi, A.C. Berg, and T.L. Berg. Baby talk: Understanding and generating simple image descriptions. In *Proceedings of CVPR*, pages 1601 –1608, 2011.

M. Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL*, pages 545–552, 2003.

S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of CoNLL*, pages 220–228, 2011.

M. Liakata, S. Teufel, A. Siddharthan, and C. Batchelor. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of LREC*, 2010.

Z. Lin, H. Ng, and M. Kan. Automatically evaluating text coherence using discourse relations. In *Proceedings of ACL-HLT*, pages 997–1006, 2011.

A. Louis and A. Nenkova. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of IJCNLP*, pages 605–613, 2011a.

A. Louis and A. Nenkova. Text specificity and impact on quality of news summaries. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation, ACL-HLT*, pages 34–42, 2011b.

A. Louis and A. Nenkova. A coherence model based on syntactic patterns. In *Proceedings of EMNLP*, pages 1157–1168, 2012a.

A. Louis and A. Nenkova. Automatically assessing machine summary content without a gold-standard. *Computational Linguistics*, 2012b.

N. McIntyre and M. Lapata. Learning to tell tales: A data-driven approach to story generation. In *Proceedings of ACL-IJCNLP*, pages 217–225, 2009.

A. Nenkova, J. Chae, A. Louis, and E. Pitler. Structural features for predicting the linguistic quality of text: applications to machine translation, automatic summarization and human-authored text. In Emiel Krahmer and Mariët Theune, editors, *Empirical methods in natural language generation*, pages 222–241. Springer-Verlag, Berlin, Heidelberg, 2010.

E. Pitler and A. Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of EMNLP*, pages 186–195, 2008.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2011.

E. Sandhaus. The New York Times Annotated Corpus. *Corpus number LDC2008T19, Linguistic Data Consortium, Philadelphia*, 2008.

S. Schwarm and M. Ostendorf. Reading level assessment using support vector machines and statistical language models. In *Proceedings of ACL*, pages 523–530, 2005.

E. Shutova, L. Sun, and A. Korhonen. Metaphor identification using verb and noun clustering. In *Proceedings of COLING*, pages 1002–1010, 2010.

L. Si and J. Callan. A statistical model for scientific readability. In *Proceedings of CIKM*, pages 574–576, 2001.

R. Soricut and D. Marcu. Discourse generation using utility-trained coherence models. In *Proceedings of COLING-ACL*, pages 803–810, 2006.

S. H. Stocking. *The New York Times Reader: Science and Technology*. CQ Press, Washington DC, 2010.

J. Swales. *Genre analysis: English in academic and research settings*, volume 11. Cambridge University Press, 1990.

J. M. Swales and C. Feak. *Academic writing for graduate students: A course for non-native speakers of English*. Ann Arbor: University of Michigan Press, 1994.

J. Tetreault, J. Foster, and M. Chodorow. Using parse features for preposition selection and error detection. In *Proceedings of ACL*, pages 353–358, 2010.

S. Teufel and M. Moens. What's yours and what's mine: determining intellectual attribution in scientific text. In *Proceedings of EMNLP*, pages 9–17, 2000.

S. Teufel, J. Carletta, and M. Moens. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of EACL*, pages 110–117, 1999.