# Automatic Evaluation of Information Ordering: Kendall's Tau

Mirella Lapata*
University of Edinburgh

*This article considers the automatic evaluation of information ordering, a task underlying many text-based applications such as concept-to-text generation and multidocument summarization. We propose an evaluation method based on Kendall's $\tau$, a metric of rank correlation. The method is inexpensive, robust, and representation independent. We show that Kendall's $\tau$ correlates reliably with human ratings and reading times.*

## 1. Introduction

The systematic evaluation of natural language processing (NLP) systems is an important prerequisite for assessing their quality and improving their performance. Traditionally, human involvement is called for in evaluating systems that generate textual output. Examples include text generation, summarization, and, notably, machine translation. Human evaluations consider many aspects of automatically generated texts ranging from grammaticality to content selection, fluency, and readability (Teufel and van Halteren 2004; Nenkova 2005; Mani 2001; White and O'Connell 1994).

The relatively high cost of producing human judgments, especially when evaluations must be performed quickly and frequently, has encouraged many researchers to seek ways of evaluating system output automatically. Papineni et al. (2002) proposed BLEU, a method for evaluating candidate translations by comparing them against reference translations (using *n*-gram co-occurrence overlap). Along the same lines, the content of a system summary can be assessed by measuring its similarity to one or more manual summaries (Hovy and Lin 2003). Bangalore, Rambow, and Whittaker (2000) introduce a variety of quantitative measures for evaluating the accuracy of an automatically generated sentence against a reference corpus string.

Despite differences in application and form, automatic evaluation methods usually involve the following desiderata. First, they measure numeric similarity or closeness of system output to one or several gold standards. Second, they are inexpensive, robust, and ideally language independent. Third, correlation with human judgments is an important part of creating and testing an automated metric. For instance, several studies have shown that BLEU correlates with human ratings on machine translation quality (Papineni et al. 2002; Doddington 2002; Coughlin 2003). Bangalore, Rambow, and Whittaker (2000) demonstrate that tree-based evaluation metrics for

* School of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK.
E-mail: mlap@inf.ed.ac.uk

Submission received: 28 December 2005; accepted for publication: 6 May 2006.

surface generation correlate significantly with human judgments on sentence quality and understandability. Given their simplicity, automatic evaluation methods cannot be considered as a direct replacement for human evaluations (see Callison-Burch, Osborne, and Koehn [2006] for discussion on some problematic aspects of BLEU). However, they can be usefully employed during system development, for example, for quickly assessing modeling ideas or for comparing across different system configurations (Papineni et al. 2002; Bangalore, Rambow, and Whittaker 2000).

Automatic methods have concentrated on evaluation aspects concerning lexical choice (e.g., words or phrases shared between reference and system translations), content selection (e.g., document units shared between reference and system summaries), and grammaticality (e.g., how many insertions, substitutions, or deletions are required to transform a generated sentence to a reference string). Another promising, but, less studied, avenue for automatic evaluation is information ordering. The task concerns finding an acceptable ordering for a set of preselected information-bearing items (Lapata 2003; Barzilay and Lee 2004). It is an essential step in concept-to-text generation, multidocument summarization, and other text synthesis problems. Depending on the application and domain at hand, the items to be ordered may vary greatly from propositions (Karamanis 2003; Dimitromanolaki and Androutsopoulos 2003) to trees (Mellish et al. 1998) or sentences (Lapata 2003; Barzilay and Lee 2004). It is therefore not surprising that evaluation methods have concentrated primarily on the generated orders, thus abstracting away from the items themselves.

More concretely, Lapata (2003) proposed the use of Kendall's $\tau$, a measure of rank correlation, as a means of estimating the distance between a system-generated and a human-generated gold-standard order. Rank correlation is an appealing way of evaluating information ordering: It is a well-understood and widely used measure of the strength of association between two variables; it is computed straightforwardly and can operate over distinct linguistic units (e.g., sentences, trees, or propositions). Indeed, several studies have adopted Kendall's $\tau$ as a performance measure for evaluating the output of information-ordering components both in the context of concept-to-text generation (Karamanis and Mellish 2005; Karamanis 2003) and summarization (Lapata 2003; Barzilay and Lee 2004; Okazaki, Matsuo, and Ishizuka 2004).

Despite its growing popularity, no study to date has investigated whether Kendall's $\tau$ correlates with human judgments on the information-ordering task. This is in marked contrast with other automatic evaluation methods that have been shown to correlate with human assessments. In this article, we aim to rectify this and undertake two studies that examine whether there is indeed a relationship between $\tau$ and behavioral data. We first briefly introduce Kendall's $\tau$ and explain how it can be employed for evaluating information ordering (Section 2). Next, we present a controlled experimental study that examines whether Kendall's $\tau$ is correlated with human ratings (Section 3).

A commonly raised criticism of the judgment elicitation methodology is that it is not fine-grained enough to rule out possible confounds. In the information-ordering task, for example, we cannot be certain that subjects rate a document low because it is genuinely badly organized and, therefore, difficult to comprehend or because they are unfamiliar with its content or simply disinterested or distracted. Similar confounds also arise in the evaluation of the output of MT systems, where it may be difficult to tease apart whether subjects' ratings reflect their assessment of the quality of the translated text or its subject matter and structure. To eliminate such confounds, we follow our judgment elicitation study with an on-line reading experiment and demonstrate that $\tau$ is also correlated with processing time (Section 4). Our second experiment provides

additional evidence for the validity of τ as a measure of text well-formedness. Discussion of our results concludes the article.

## 2. Kendall's Measure

In common with other automatic evaluation methods, we assume that we have access to a reference output that in most cases will be created by one or several humans. Our task is to compare a system-produced ordering of items against a reference order. For ease of exposition, let us assume that our information-ordering component is part of a generation application whose ultimate goal is to generate coherent and understandable text. It is not crucially important how the items to be ordered are represented. They can be facts in a database (Duboue and McKeown 2001), propositions (Karamanis 2003), discourse trees (Mellish et al. 1998), or sentences (Lapata 2003; Barzilay and Lee 2004).

Now, we can think of the items as objects for which a ranking must be produced. Table 1 gives an example of a reference text containing 10 items (A–J) and the orders (i.e., rankings) produced by two hypothetical systems. We can then calculate how much the system orders *differ* from the reference order, the underlying assumption being that acceptable orders should be fairly similar to the reference. A number of metrics can be employed for this purpose, such as Spearman's correlation coefficient ($r_s$) for ranked data, Cayley distance, or Kendall's τ (see Lebanon and Lafferty [2002] for an overview). Here we describe Kendall's τ (Kendall 1938) and explain why it is an appropriate choice for information-ordering tasks.

Let $\mathcal{Y} = y_1 \ldots y_n$ be a set of items to be ranked. Let π and σ denote two distinct orderings of $\mathcal{Y}$, and $S(\pi, \sigma)$ the minimum number of adjacent transpositions needed to bring π to σ. Kendall's τ is defined as:

$$\tau = 1 - \frac{2S(\pi, \sigma)}{N(N-1)/2} \tag{1}$$

where $N$ is the number of objects (i.e., items) being ranked. As can be seen, Kendall's τ is based on the number transpositions, that is, interchanges of consecutive elements, necessary to rearrange π into σ. In Table 1 the number of transpositions can be calculated by counting the number of intersections of the lines. The τ between the Reference and System 1 is 0.82, between the Reference and System 2 is 0.24, and between the two systems is 0.15. The metric ranges from −1 (inverse ranks) to 1 (identical ranks). The calculation of τ must be appropriately modified when there are tied rankings (Hays 1994; Siegel and Castellan 1988).

Kendall's τ seems particularly appropriate for the information-ordering tasks considered in this article. The metric is sensitive to the fact that some items may be always

**Table 1**
Example of reference order and system orders for a text consisting of 10 items.

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Reference | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| System 1 | 2 | 1 | 5 | 3 | 4 | 6 | 7 | 9 | 8 | 10 |
| System 2 | 10 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 1 |

ordered next to each other even though their absolute orders might differ. It also penalizes inverse rankings. Comparison between the Reference and System 2 gives a $\tau$ of 0.24 even though the orders between the two models are identical modulo the beginning and the end. This seems appropriate given that flipping the introduction in a document with the conclusions seriously disrupts coherence.

Kendall's $\tau$ is less widely used than Spearman's rank correlation coefficient ($r_s$). Both coefficients use the same amount of information in the data, and thus both have the same sensitivity to detect the existence of association. This means that for a given data set, both measures will lead to rejection of the null hypothesis at the same level of significance. However, the two measures have different underlying scales, and, numerically, they are not directly comparable to each other. Siegel and Castellan (1988) express the relationship of the two measures in terms of the inequality:

$$-1 \leq 3\tau - 2r_s \leq 1 \tag{2}$$

More importantly, Kendall's $\tau$ and $r_s$ have different interpretations. Kendall's $\tau$ can be interpreted as a simple function of the probability of observing concordant and discordant pairs (Kerridge 1975). In other words, it is the difference between the probability that in the observed data two variables are in the same order versus the probability that they are in different orders (the probability is rescaled to range from $-1$ to 1 as is customary for correlation; see equation (1)). Unfortunately, no simple meaning can be attributed to $r_s$. The latter is the same as a Pearson product–moment correlation coefficient ($r_p$) computed for values consisting of ranks. Although $r^2$ represents the percent of variance shared by two variables in the case of $r_p$, its interpretation is less straightforward for $r_s$, where it refers to the percent of variance of two ranks. It is difficult to draw any meaningful conclusions with regard to information ordering based on the variance of ranks.

In practice, while both correlations frequently provide similar answers, there are situations where they diverge. For example, the statistical distribution of $\tau$ approaches the normal distribution faster than $r_s$ (Kendall and Gibbons 1990), thus offering an advantage for small to moderate sample studies with 30 or fewer data points. This is crucial when experiments are conducted with a small number of subjects (a situation common in NLP) or test items. Another related issue concerns sample size. Spearman's rank correlation coefficient is a biased statistic (Kendall and Gibbons 1990). The smaller the sample the more $r_s$ diverges from the true population value, usually underestimating it. In contrast, Kendall's $\tau$ does not provide a biased estimate of the true correlation. Furthermore, $\tau$ maintains good control of type I error rates (i.e., rejecting the null hypothesis when it is actually true). Arndt, Turvey, and Andreasen (1999) undertake an extensive empirical study and show that the number of times $\tau$ incorrectly signals a significant correlation when there is none is close to the nominal 5% using a $p < 0.05$ significance criterion. For a more detailed discussion of the advantages of $\tau$ over $r_s$, we refer the interested reader to Kendall and Gibbons (1990) and Arndt, Turvey, and Andreasen (1999).

## 3. Experiment 1: Judgment Elicitation

To assess whether Kendall's $\tau$ reliably correlates with human ratings, it is necessary to have access to several different orderings of the same input. In what follows we

describe our method for assembling a set of experimental materials and collecting human judgments.

### 3.1 Method

**3.1.1 Design and Materials.** Our goal here is to establish whether $\tau$ correlates with human judgments on overall text understandability and coherence. A system that randomly pastes together sentences or facts from a database will ultimately produce a badly organized document lacking coherence. A good automatic evaluation method should assign low values to such documents and higher values to documents that are easy for humans to read and understand.

We could elicit judgments by asking humans to rate the output of an information-ordering component. The ratings could be then correlated with $\tau$ values representing the difference between system and reference orders. Such a comparison is, however, undesirable for a number of reasons. First, the system may be biased toward very good or very bad orders. This means that our hypothetical study would only examine a restricted and potentially skewed range of $\tau$ values. Furthermore, in concept-to-text generation applications, information ordering typically operates over symbolic representations that will be unfamiliar to naive informants and could potentially distort their judgments. A related issue arises in text-to-text generation applications where the produced documents are not necessarily grammatical, for example, when a summary is the output of an information fusion component (Barzilay 2003; Radev and McKeown 1998). Again, it is difficult to control whether informants judge the ordering or the grammaticality of the texts.

To make the judgment task easier, we concentrated on a document representation familiar to our participants, namely, sentences. We simulated the output of an information-ordering component by randomly generating different sentence orders for a reference text. We elicited judgments for eight texts of the same length (eight sentences). The texts were randomly sampled from a corpus collected by Barzilay and Lee (2004) (sampling took place over eight-sentence-length documents only). The corpus consists of Associated Press articles on the topic of natural disasters, drug-related criminal offenses, clashes between armies and rebel groups, and narratives from the U.S. National Transportation Safety Board database.[1]

A document consisting of eight sentences can be sequenced in 8! ways. We exhaustively enumerated all possible orderings and calculated their $\tau$ value against the reference order found in the corpus.[2] Figure 1 shows how many different orders correspond to a given $\tau$ value. For example, there is only one order with a $\tau$ of 1 or $-1$, whereas there are 3,736 orders with $\tau$ 0.07 or $-0.07$.

Ideally, we should elicit judgments on orders corresponding to all 29 values from Figure 1. Unfortunately, this would render our experimental design unwieldy. Assuming we randomly select one order for each value, our participants would have to judge $29 \times 8 = 232$ texts. In order to strike a balance between a manageable design and a wide range of $\tau$ values, we split the $\tau$ range into eight bins (see Figure 2). For each text, an order was randomly sampled from each bin. Thus, our set of materials consisted of $8 \times 8 = 64$ texts. Pronouns that could not be resolved intra-sententially were substituted by their referents to avoid creating coherence violation artifacts. For

1 The corpus is available from http://people.csail.mit.edu/regina/struct/.

2 Notice that the number of permutations and range of $\tau$ values is the same for all our texts, since they all have the same length.

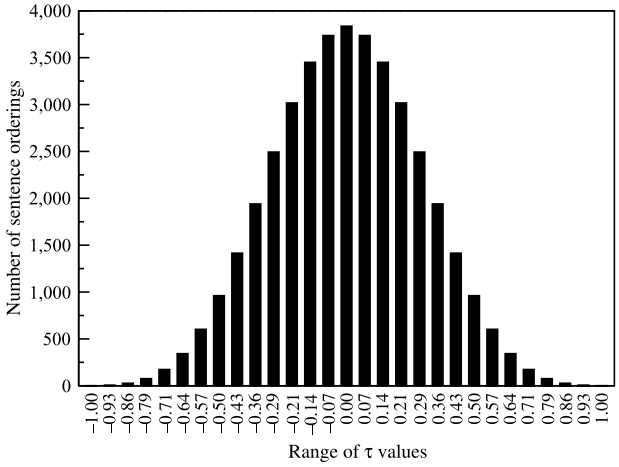


**Figure 1**
Range of τ values for a document consisting of eight sentences.

the same reason, we excluded from our materials texts containing discourse connectives (e.g., *but*, *therefore*).

**3.1.2 Procedure.** During the elicitation study, participants were presented with texts and asked to judge how comprehensible they were on a seven-point scale. They were told that some texts would be perfectly understandable, whereas others would be fairly incoherent and the order of the sentences might seem scrambled.

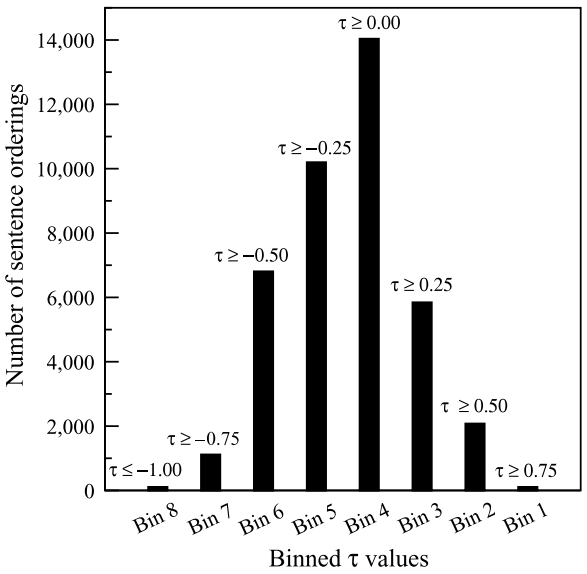


**Figure 2**
Range of τ values when collapsed across eight bins.

The study was conducted remotely over the Internet. Participants first saw a set of instructions that explained the task and provided several examples of well- and badly organized texts, together with examples of numerical estimates. From our set of materials we generated 8 lists (each consisting of 8 texts) following a Latin square design. Each subject was randomly assigned to one list. The procedure ensured that no two texts in a given list corresponded to the same reference text. It was emphasized that there were no correct answers and that subjects should base their judgments on first impressions, not spending too much time on any one text. Example stimuli are shown in Table 2.

The subjects accessed the experiment using their Web browser. Experimental instructions and materials were administered via CGI scripts. A number of safeguards were put in place to ensure the authenticity of the subjects taking part. Participants had to provide their e-mail address and were asked to fill in a short questionnaire including basic demographic information (name, age, sex, handedness, and language background). Subjects' e-mail addresses were automatically checked for plausibility and subjects with fake addresses were removed. The elicited responses were also screened to identify (and eliminate) subjects taking part in the experiment more than once.

**3.1.3 Subjects.** The experiment was completed by 189 unpaid volunteers, all self-reported native speakers of English. Subjects were recruited by postings to local e-mail lists; they had to be linguistically naive, neither linguists nor students of linguistics were allowed to participate. Four subjects were eliminated because they were non-native English speakers. The data of six subjects were excluded after inspection of their responses revealed anomalies in their ratings. For example, they either provided ratings outside the prespecified scale (1–7) or rated all documents uniformly. This left 179 subjects for analysis (approximately 22 per text). Forty-nine of our participants were

**Table 2**
Example stimuli representing a well- (top) and badly- (bottom) organized document.

| |
|---|
| Police arrested 18 people Saturday in an alleged international ring that smuggled hashish in from Morocco for distribution in Europe. The group allegedly smuggled the hashish to Cadiz, on Spain's southern coast, and then used trains to transport it to Barcelona and Italy. The group, based in Seville with ties in Las Palmas, Barcelona, Morocco and Italy, was headed by the Rufos family, police said. Police seized 100 kilograms (220 pounds) of hashish, 10 million pesetas (dlrs 80,000), nine vehicles, rifles, computers, mobile phones, video cameras and false identification papers. Arrests were made in Seville, Las Palmas and Barcelona. Police did not provide names of suspects, or nationalities of those arrested. Southern Spain is a main gateway for hashish being smuggled into Europe from northern Africa. Hundreds of kilograms (pounds) are seized each week. |
| The group allegedly smuggled the hashish to Cadiz, on Spain's southern coast, and then used trains to transport it to Barcelona and Italy. Hundreds of kilograms (pounds) are seized each week. Southern Spain is a main gateway for hashish being smuggled into Europe from northern Africa. Arrests were made in Seville, Las Palmas and Barcelona. Police did not provide names of suspects, or nationalities of those arrested. Police seized 100 kilograms (220 pounds) of hashish, 10 million pesetas (dlrs 80,000), nine vehicles, rifles, computers, mobile phones, video cameras and false identification papers. The group, based in Seville with ties in Las Palmas, Barcelona, Morocco and Italy, was headed by the Rufos family, police said. Police arrested 18 people Saturday in an alleged international ring that smuggled hashish in from Morocco for distribution in Europe. |

female and 42 male. The age of the subjects ranged from 18 to 60 years. The mean was 28.5 years.

### 3.2 Results

The judgments were averaged to provide a single rating per text. We first analyzed the correspondence of human ratings and $\tau$ values by performing an analysis of variance (ANOVA). Recall that $\tau$ represents the degree of similarity between a synthetically generated text and a reference text. In our case, the reference texts are the original human-authored documents from our corpus. Our participants judge how well a document is organized without having access to the original reference.

Our ANOVA analysis had one factor (i.e., $\tau$ value) with eight levels corresponding to the eight bins discussed in Section 3.1 (see Figure 2). The ANOVA showed that this factor was significant in both by-subjects and by-items analyses: $F_1(7, 1239) = 42.60, p < 0.01$; $F_2(7, 56) = 2.77, p < 0.01$. Table 3 shows the average subject ratings and descriptive statistics for each of the eight bins. Post hoc Tukey tests indicated that the ratings for texts with $\tau$ values from Bin 1 were significantly different from the ratings assigned to all other bins ($\alpha = 0.01$). Although ratings for Bins 2, 3, 4, and 5 did not significantly differ from each other, they all differed from Bins 6, 7, and 8. The results of the ANOVA show that our participants tended to give high scores to texts with high $\tau$ values and low scores to texts with low $\tau$ values.

We next used correlation analysis to explore the linear relationship between subjects' ratings and Kendall's $\tau$. The comparison yielded a Pearson correlation coefficient of $r = 0.45$ ($p < 0.01$, $N = 64$). Figure 3 plots the relationship between judgments and $\tau$ values. To get a better understanding of how this automatic evaluation method compares with human judgments, we examined how well our raters agreed in their assessment. To calculate intersubject agreement we used leave-one-out resampling. The technique is a special case of $n$-fold cross-validation (Weiss and Kulikowski 1991) and has been previously used for measuring how well humans agree on judging semantic similarity (Resnik and Diab 2000; Resnik 1999), adjective plausibility (Lapata and Lascarides 2003), and text coherence (Barzilay and Lapata 2005).

The set of $m$ subjects' responses was divided into two sets: a set of size $m - 1$ (i.e., the response data of all but one subject) and a set of size one (i.e., the response data of a single subject). We then correlated the mean ratings of the former set with the ratings of the latter. This was repeated $m$ times. Since we had 179 subjects, we performed

**Table 3**
Average subject ratings for binned $\tau$ values and descriptive statistics.

| Bins | Mean | Min | Max | SD |
|---|---|---|---|---|
| 1 | 5.348 | 3.000 | 7.000 | 1.236 |
| 2 | 4.916 | 1.000 | 7.000 | 1.612 |
| 3 | 4.927 | 2.000 | 7.000 | 1.580 |
| 4 | 4.470 | 1.000 | 7.000 | 1.489 |
| 5 | 4.382 | 1.000 | 7.000 | 1.559 |
| 6 | 4.208 | 1.000 | 7.000 | 1.600 |
| 7 | 4.028 | 1.000 | 7.000 | 1.702 |
| 8 | 3.966 | 1.000 | 7.000 | 1.558 |

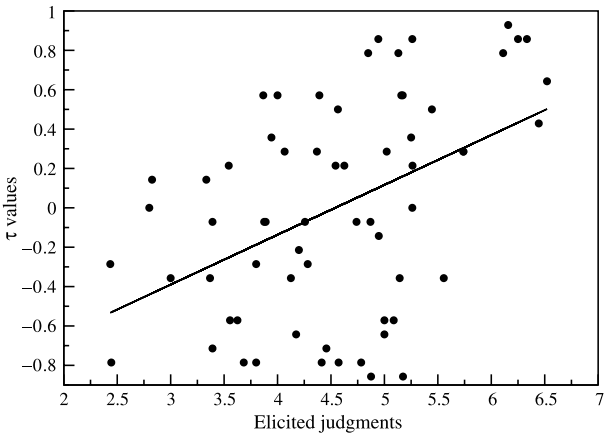


**Figure 3**
Correlation of elicited judgments and τ values.

178 correlation analyses and report their mean.[3] The average intersubject agreement was $r = 0.56$ (min $= 0.001$, max $= 0.94$, SD $= 0.25$), thus indicating that τ's agreement with the human data is not far from the average human agreement.

## 4. Experiment 2: Kendall's Tau and Processing Effort

A potential criticism of our previous study is that it is based solely on ratings. The problem with this off-line measure is that it indicates whether participants find a text easy or difficult to comprehend, without, however, isolating the causes for this difficulty. For example, the ratings may reflect not only what subjects think about how a text is organized but also their (un)familiarity with its genre or style, their lack of attention, or disinterest in the subject matter. To ascertain that this is not the case, we conducted a follow-up experiment whose aim was to explore the relationship between Kendall's τ and processing effort. Much work in psychology (McKoon and Ratcliff 1992; Britton 1991) indicates that low-coherence texts require more inferences and therefore take longer to read. If Kendall's τ does indeed capture aspects of overall document organization and coherence, then documents assigned a high τ value should take less time to read than documents with low τ values. Unlike ratings, reading times are an immediate measure of processing effort that participants cannot consciously control or modulate.

### 4.1 Method

**4.1.1 Design and Materials.** The experiment was designed to assess the relation of Kendall's τ with processing effort. Our selection of materials was informed by the ANOVA results presented in Experiment 1. We used the same eight reference texts from the previous experiment. For each text we randomly selected three synthetically generated orders, each from Bin 1 (high τ value), Bins 2–4 (medium τ value), and Bins 5–8 (low τ value). In other words, we collapsed Bins 2–4 and Bins 5–8, since the ANOVA

3 We cannot apply the commonly used kappa statistic for measuring intersubject agreement since it is appropriate for nominal scales, whereas our texts are rated on an ordinal scale.

**Table 4**
Mean reading times (in milliseconds) for three experimental conditions.

| | Mean | Min | Max | SD |
|---|---|---|---|---|
| High | 5762.6 | 1963.3 | 9111.1 | 1429.0 |
| Medium | 6499.0 | 1574.0 | 10344.5 | 2017.5 |
| Low | 7250.4 | 1428.0 | 15000.0 | 3121.3 |

revealed that ratings for these bins were not significantly different. Our set of materials consisted of $8 \times 3 = 24$ texts.

**4.1.2 Procedure.** The presentation of stimuli and collection of responses was controlled by E-Prime software[4] (version 1.1) running on a Dell Optiplex GX270 with an Intel Pentium 4 processor and 512 MB memory. The experiment started with a practice session comprising two texts, each eight sentences long. Then eight texts were presented; the presentation followed a Latin square design, thus ensuring that no subject saw the same text twice.

The texts were presented one sentence at a time. The participant pressed the space bar to proceed from one sentence to the next. Participants were instructed to read the texts at their own pace and to press the space bar after each sentence once they were certain that they understood it. Participants' reading time was recorded for each sentence. After the final sentence was displayed, subjects were asked a comprehension yes/no question to make sure that they were actually reading the texts rather than pressing the space bar randomly.

**4.1.3 Subjects.** The experiment was completed by 32 volunteers, all self-reported native speakers of English. The experiment was administered in the laboratory and subjects were paid £5 for their participation. None of the subjects had previously participated in Experiment 1.

### 4.2 Results

Sentence reading times were averaged to provide reading times for each text. As a first step, the reading time data were screened to remove errors and outliers. Errors consisted of items where the subjects had incorrectly answered the comprehension question. This affected 12.3% of the data. Reading times beyond 2.5 standard deviations above or below the mean for a particular participant were replaced with the mean plus this cut-off value. This adjustment of outliers affected 9.7% of the data. Mean reading times for each experimental condition (high, medium, low) are shown in Table 4.

ANOVA showed significant differences in reading times [$F_1(2, 62) = 6.39$, $p < 0.01$; $F_2(2, 14) = 4.23$, $p < 0.05$]. Post hoc Tukey tests revealed that high-$\tau$ texts were read significantly faster than medium- and low-$\tau$ texts ($\alpha = 0.01$). Reading times for medium-$\tau$ texts were not significantly different from low-$\tau$ texts.

4 E-prime is a suite of tools for creating and running experiments while allowing for millisecond precision data collection. For more information see http://www.pstnet.com/products/e-prime/.

We next examine through correlation analysis whether there is a linear relationship between reading times and $\tau$ values. We regressed $\tau$ values and reading times following the procedure[5] recommended in Lorch and Myers (1990). The regression yielded a Pearson correlation coefficient of $r = -0.48$ ($p < 0.01$). Expectedly, reading times are also significantly correlated with human ratings: Pearson's $r = -0.47$ ($p < 0.01$).[6]

To summarize, the results of our second experiment provide additional evidence for the use of Kendall's $\tau$ as a measure of text well-formedness. It correlates not only with human ratings but also with reading times. The latter constitute much more fine-grained behavioral data, directly associated with processing effort: Less well-structured documents tend to have low $\tau$ values and cause longer reading times, whereas documents with high $\tau$ values tend to be better organized and cause shorter reading times.

## 5. Discussion

In this article, we argue that Kendall's $\tau$ can be used as an automatic evaluation method for information-ordering tasks. We have undertaken a judgment elicitation study demonstrating that $\tau$ correlates reliably with human judgments. We have also shown that $\tau$ correlates with processing effort—texts with high $\tau$ values take less time to read than texts with low $\tau$ values. We have presented behavioral evidence collected via two distinct experimental paradigms suggesting that Kendall's $\tau$ is an ecologically valid measure of document well-formedness and structure.

An attractive feature of the $\tau$ evaluation method is that it is representation independent. It can therefore be used to evaluate both symbolic and statistical generation systems. We do not view $\tau$ as an alternative to human evaluations; rather we consider its role complementary. It can be used during system development for tracking incremental progress or as an easy way of assessing whether an idea is promising. It can also be used to compare systems that employ comparable information-ordering strategies and operate over the same input. Furthermore, statistical generation systems (Lapata 2003; Barzilay and Lee 2004; Karamanis and Manurung 2002; Mellish et al. 1998) could use $\tau$ as a means of directly optimizing information ordering, much in the same way MT systems optimize model parameters using BLEU as a measure of translation quality (Och 2003).

The $\tau$ evaluations presented in this article used a single reference text. Previous work (Barzilay, Elhadad, and McKeown 2002; Lapata 2003; Karamanis and Mellish 2005) has shown that there may be many acceptable orders for a set of information-bearing items, although topically related sentences seem to appear together (Barzilay, Elhadad, and McKeown 2002). A straightforward way to incorporate multiple references in the evaluation paradigm discussed here is to compute the $\tau$ statistic $N$ times for every reference–system output pair and report the mean. A more interesting future direction is to weight transpositions (see Section 2) according to agreements or disagreements in the set of multiple references. A possible implementation of this idea would

5 Lorch and Myers (1990) argue that it is not appropriate to average over subjects when dealing with repeated measures designs. Instead they propose three methods that effect regression analysis on reading times collected from individual subjects. We refer the interested reader to Lorch and Myers (1990) and Baayen (2004) for further discussion.

6 The correlation coefficients are negative since longer reading times correspond to lower ratings and $\tau$ values.

be to compute $\tau$ against one (randomly selected) reference, but change the metric so as to give fractional counts (i.e., less than one) to transpositions that are not uniformly attested in the reference set.

Naturally, Kendall's $\tau$ is not the only automatic evaluation method that can be employed to assess information ordering. Barzilay and Lee (2004) and Barzilay and Lapata (2005) measure accuracy as the percentage of test items for which the system gives preference to the gold-standard reference order. This measure allows us to compare the output of different systems; however, it only rewards orders identical to the gold standard, and considers all other orders deviating from it deficient. Barzilay and Lee (2004) propose an additional evaluation measure based on ranks. Assuming that a system can exhaustively generate all possible orders for a set of items (with a certain probability), they report the rank given to the reference order when all possible orders are sorted by their probability. The best possible rank is 0 and the worst rank is $N! - 1$. A system that gives a high rank to the reference order is considered worse than a system that gives it a low rank. However, not all systems are designed to exhaustively enumerate all possible permutations for a given document or have indeed a scoring mechanism that can rank alternative document renderings. Duboue and McKeown (2002) employ an alignment algorithm that allows them to compare the output of their algorithm with a gold-standard order. The alignment algorithm works by considering the similarity between system-generated and gold-standard facts. The similarity function is domain dependent (Duboue and McKeown [2002] generate postcardiac surgery medical briefings) and would presumably have to be redefined for a different set of facts in another domain.

Kendall's $\tau$ can be easily used to evaluate the output of automatic systems, irrespectively of the domain or application at hand. It requires no additional tuning and correlates reliably with behavioral data. Since it is a similarity measure, it can be used to evaluate system output that is not necessarily identical to the gold standard. Also note that $\tau$ could be used to compare across systems operating over similar input/output even if reference texts are not available. For example, $\tau$ could identify outlier systems with output radically different from the mean.

**Acknowledgments**

The author acknowledges the support of EPSRC (grant GR/T04540/01). Thanks to Frank Keller, Nikiforos Karamanis, Scott McDonald, and two anonymous reviewers for helpful comments and suggestions.

**References**

Arndt, Stephan, Carolyn Turvey, and Nancy C. Andreasen. 1999. Correlating and predicting psychiatric symptom ratings: Spearman's *r* versus Kendall's tau correlation. *Journal of Psychiatric Research*, 33:97–104.

Baayen, Harald R. 2004. Statistics in psycholinguistics: A critique of some current gold standards. In *Mental Lexicon Working Papers 1*. University of Alberta, Edmonton, pages 1–45.

Bangalore, Srinivas, Owen Rambow, and Steven Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of the INLG*, pages 1–8, Mitzpe Ramon, Israel.

Barzilay, Regina. 2003. *Information Fusion for Multi-Document Summarization: Paraphrasing and Generation*. Ph.D. thesis, Columbia University.

Barzilay, Regina, Noemie Elhadad, and Kathleen R. McKeown. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

Barzilay, Regina and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 141–148, Ann Arbor.

Barzilay, Regina and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the 2nd Human Language Technology Conference*

*and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 113–120, Boston, MA.

Britton, Bruce K. 1991. Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology*, 83(3):329–345.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics*, pages 249–256, Trento, Italy.

Coughlin, Deborah. 2003. Correlating automated and human assessments of machine translation quality. In *Proceedings of MT Summit IX*, pages 63–70, New Orleans.

Dimitromanolaki, Aggeliki and Ion Androutsopoulos. 2003. Learning to order facts for discourse planning in natural language generation. In *Proceedings of the 9th European Workshop on Natural Language Generation*, pages 113–120, Budapest, Hungary.

Doddington, George. 2002. Automatic evaluation of machine translation quality using *n*-gram cooccurrence statistics. In *Human Language Technology: Notebook Proceedings*, pages 128–132, San Diego.

Duboue, Pablo and Kathleen R. McKeown. 2002. Content planner construction via evolutionary algorithms and a corpus-based fitness function. In *Proceedings of INLG 2002*, pages 89–96, New York.

Duboue, Pablo A. and Kathleen R. McKeown. 2001. Empirically estimating order constraints for content planning in generation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 172–179, Toulouse, France.

Hays, William L. 1994. *Statistics*. Harcourt Brace College Publishers, New York, 3rd edition.

Hovy, Eduard and Chin-Yew Lin. 2003. Automatic evaluation of summaries using *N*-gram co-occurrence statistics. In *Proceedings of the 1st Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 71–78, Edmonton, Canada.

Karamanis, Nikiforos. 2003. *Entity Coherence for Descriptive Text Structuring*. Ph.D. thesis, University of Edinburgh.

Karamanis, Nikiforos and Hisar Maruli Manurung. 2002. Stochastic text structuring using the principle of continuity. In *Proceedings of the 2nd International Conference on Natural Language Generation*, pages 81–88, New York.

Karamanis, Nikiforos and Chris Mellish. 2005. Using a corpus of sentence orderings defined by many experts to evaluate metrics of coherence for text structuring. In *Proceedings of the 10th European Workshop on Natural Language Generation*, pages 174–179, Aberdeen, Scotland.

Kendall, Maurice G. 1938. A new measure of rank correlation. *Biometrika*, 30:81–93.

Kendall, Maurice G. and Jean Dickinson Gibbons. 1990. *Rank Correlation Methods*. Oxford University Press, New York.

Kerridge, D. 1975. The interpretation of rank correlations. *Applied Statistics*, 24(2):257–258.

Lapata, Maria and Alex Lascarides. 2003. A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2):263–317.

Lapata, Mirella. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 545–552, Sapporo, Japan.

Lebanon, Guy and John Lafferty. 2002. Combining rankings using conditional probability models on permutations. In *Proceedings of the 19th International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers, pages 363–370.

Lorch, Robert F. and Jerome L. Myers. 1990. Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(1):149–157.

Mani, Inderjeet. 2001. *Automatic Summarization*. John Benjamins Pub Co., Amsterdam; Philadelphia.

McKoon, Gail and Roger Ratcliff. 1992. Inference during reading. *Psychological Review*, 99(3):440–446.

Mellish, Chris, Alistair Knott, Jon Oberlander, and Mick O' Donnell. 1998. Experiments using stochastic search for text planning. In *Proceedings of the 9th International Workshop on Natural Language Generation*, pages 98–107, Ontario, Canada.

Nenkova, Ani. 2005. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *Proceedings of the 20th National Conference on Artificial Intelligence*, pages 1436–1441, Pittsburgh, PA.

Och, Franz Joseph. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Okazaki, Naoaki, Yutaka Matsuo, and Mitsuru Ishizuka. 2004. Improving chronological sentence ordering by precedence relation. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, pages 750–756.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Radev, Dragomir and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.

Resnik, Philip. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, pages 95–130.

Resnik, Philip and Mona Diab. 2000. Measuring verb similarity. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum Associates, Mahwah, NJ, pages 399–404.

Siegel, Sidney and N. John Castellan. 1988. *Non Parametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York.

Teufel, Simone and Hans van Halteren. 2004. Evaluating information content by factoid analysis: Human annotation and stability. In Dekang Lin and Dekai Wu, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 419–426, Barcelona.

Weiss, Sholom M. and Casimir A. Kulikowski. 1991. *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann, San Mateo, CA.

White, John S. and T. O'Connell. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 193–205, Columbia, MD.