

# A Review of Recent Corpus-based Methods for Evaluating Information Ordering in Text Production

**Nikiforos Karamanis**

Computational Linguistics Research Group  
University of Wolverhampton, UK  
N.Karamanis@wlv.ac.uk

**Chris Mellish**

Department of Computing Science  
University of Aberdeen, UK  
cmellish@csd.abdn.ac.uk

## Abstract

This paper surveys the corpus-based methods for evaluating Information Ordering (IO) which emerged recently in the literature on text production. First, we discuss how different assumptions about the input to IO make the preparation of corpora suitable for evaluation more challenging in Natural Language Generation than in Automatic Multidocument Summarisation. Then, we present the types of corpora and performance measures employed by the reviewed work emphasising the considerable consensus that has emerged in these two aspects of automatic evaluation of IO.

## 1 Introduction

Although evaluating text production systems is much more complicated than evaluating systems analysing text [Hirschman and Mani, 2003], breaking down the tasks that text production consists of may clarify the questions that need to be asked [Mellish and Dale, 1998]. Information Ordering (henceforth, IO), i.e. deciding in which sequence to present a set of preselected information-bearing items is an important issue in Natural Language Generation (NLG) [Reiter and Dale, 2000] and related areas such as Multi-Document Summarisation (MDS) [Barzilay *et al.*, 2002].

Because investigating IO extensively by employing human informants in psycholinguistic experiments is often unfeasible [Karamanis, 2003; Barzilay and Lee, 2004], empirical work on the evaluation of IO has recently become increasingly automatic and corpus-based, which in turn enables large-scale experimentation to take place more easily and researchers to start sharing data between them. Hence, although the corpus-based evaluation of IO is far from mature yet, it seems to represent a good case study on the feasibility of using corpora for testing a specific task in NLG and MDS.

This short paper reviews the corpus-based methods for evaluating IO presented recently in [Duboue and McKeown, 2002], [Dimitromanolaki and Androutsopoulos, 2003], [Lapata, 2003], [Barzilay and Lee, 2004], [Karamanis *et al.*, 2004a; 2004b], [Karamanis and Mellish, 2005] and [Barzilay

and Lapata, 2005].<sup>1</sup> Following [Bangalore *et al.*, 2000] who evaluated sentence planning in NLG, all reviewed papers share the assumption that a text production method can be evaluated, albeit approximately, by automatically comparing its output with human-defined solutions attested in a corpus of texts each of which is viewed as a “gold standard”. Although we agree (as the other reviewed authors do as well) with [Reiter and Sripada, 2002] that the results of corpus-based evaluation are best treated as hypotheses which eventually need to be integrated with other, perhaps even more time-consuming, evaluation efforts, automatic evaluation might still be particularly useful during development. Thus, reviewing the existing attempts to automatically evaluate IO seems to be justified as it discusses issues worthy of the attention of our colleagues in the text production community.

In the next section, we discuss how the reviewed evaluation efforts differ in the assumptions made about the inputs and outputs of IO depending on whether IO is related to NLG or MDS. We then explain how the different input assumptions introduce additional complications in the preparation of a corpus suitable for NLG-related evaluation. We continue by distinguishing between *multi-authored* and *parallel-authored* corpora employed by the reviewed research and assess which type best meets the *representativeness* requirement [McEnery, 2003]. Last but not least, we present what appears to be the main novelty of the reviewed evaluation efforts, that is, the performance measures they introduce to automatically compare machine-generated orderings of information-bearing items with human-defined orderings from a corpus. We distinguish between *search-oriented* and *distance-based* evaluation measures and discuss which type of measure is more suitable to evaluate certain types of IO methods.<sup>2</sup> The review is completed with a summary of our main conclusions.

<sup>1</sup>[Karamanis *et al.*, 2004a; 2004b; Karamanis and Mellish, 2005] are based on [Karamanis, 2003]. We do not consider work which is primarily based on human judgments such as [Walker *et al.*, 2002].

<sup>2</sup>Although we distinguish between *deterministic* and *non-deterministic* ways (in our terms *methods*) of producing an ordering, we refrain from explaining in detail how the machine-generated orderings which are subject to the reviewed evaluation experiments (in our terms *efforts*) are actually produced.

Inputs to IO:	Surface text	Database facts
Evaluation effort:	MDS-related	NLG-related
Papers:	[Lapata, 2003], [Barzilay and Lee, 2004], [Karamanis <i>et al.</i> , 2004b], [Barzilay and Lapata, 2005]	[Duboue and McKeown, 2002], [Dimitrom. and Andr. 2003], [Karamanis <i>et al.</i> , 2004a], [Karamanis and Mellish, 2005]

Table 1: Classification of reviewed papers in terms of the assumed input to Information Ordering (IO) and the corresponding type of evaluation effort

## 2 Information ordering: inputs and outputs

Typically, IO in MDS constitutes a separate module which receives words already organised as sentences [Lapata, 2003; Barzilay and Lee, 2004; Barzilay and Lapata, 2005] (or, less typically, finite clauses [Karamanis *et al.*, 2004b]) as its input. The output of this module is simply an ordering of the input set of information-bearing items. We will subsequently refer to the evaluation efforts in which surface text chunks are assumed to be the input to IO as “MDS-related” (displayed in the second column of Table 1).

The remaining reviewed evaluation efforts (summarised in the third column of Table 1) are characterised as “NLG-related” because they hypothesise an input to IO similar to the representation typically used during an early NLG stage called document planning [Reiter and Dale, 2000]. This input is a system-specific collection of database facts (also known as messages) which (although they are often seen as corresponding to sentences or clauses) are not yet realised as surface text when document planning is performed.

Notably, organising IO as a separate module is not the mainstream NLG view. Rather, in standard NLG, IO is the result of organising the database facts to a tree-like structure during document planning. Crucially, even [Duboue and McKeown, 2002], the only paper in our review that evaluates a hierarchical document planning method, present a corpus-based evaluation measure dealing with orderings rather than tree-like representations. This suggests that traditional document planning methods could also be evaluated in terms of IO, especially since it seems easier to annotate a text for the ordering of the information-bearing items it contains (than the way these items are organised into a hierarchical structure).

However, using database facts as the input to IO introduces additional complications in the preparation of data suitable for NLG-related evaluation, as discussed in the next section in more detail.

### 2.1 Issues in corpus annotation

The MDS-related evaluation efforts typically test IO methods which rely on syntactic and semantic features that can normally be annotated directly on the input text giving rise to representations similar to the ones employed in other corpus-based research. While [Barzilay and Lapata, 2005] annotate a large collection of texts automatically using methods reported to suffer from an at least 10% error rate, [Karamanis *et al.*, 2004b] make use of a much smaller corpus manually yet very reliably annotated [Poesio *et al.*, 2004]. [Barzilay and Lee, 2004] rely on a knowledge-lean IO method which is applied directly to many un-annotated texts in five different domains.

On the other hand, for a corpus to be used for research

Database fact	Sentence
subclass(ex1, amph)	→ This exhibit is an amphora.
painted-by(ex1, p-Kleo)	→ This exhibit was decorated by the Painter of Kleofrades.
painter-story(p-Kleo, en4049)	→ The Painter of Kleofrades used to decorate big vases.
exhibit-depicts(ex1, en914)	→ This exhibit depicts a warrior performing splachnoscropy before leaving for the battle.
current-location(ex1, wag-mus)	→ This exhibit is currently displayed in the Martin von Wagner Museum.
museum-country(wag-mus, ger)	→ The Martin von Wagner Museum is in Germany.

Figure 1: Database facts corresponding to sentences

in NLG, the surface text which represents the system’s target output has to be mapped with representations of the input data that the text is supposed to communicate [Reiter and Dale, 2000; Reiter and Sripada, 2002]. This requirement makes the corpora used for the NLG-related evaluation of IO quite different from the corpora typically employed in the MDS-related evaluation efforts.

[Duboue and McKeown, 2002] responded to this requirement by manually mapping each of the texts in their corpus to database facts from their NLG system. Manual annotation is not a trivial effort and, if not properly controlled, the resulting representation might represent several biases introduced by the annotator [Mellish and Dale, 1998].

As the texts of [Duboue and McKeown, 2002] were produced by humans who did not have the system’s input in mind, they were found to express information not contained in the system’s database while information that the system was supposed to express in a certain text was not verbalised. Moreover, a close examination of their data reveals that e.g. the phrase “John Doe, a 41-year old patient of Dr Smith” (taken from Figure 6 in [Duboue and McKeown, 2002]) maps to at least three facts from the database in a way that makes it difficult to define an order between them.

Problems like these surface quite commonly in NLG research and can be addressed by revising the human text to a more simplified, NLG-tuned (yet usually more stilted) document [Reiter and Dale, 2000]. Instead of that, [Duboue and McKeown, 2002] adopted an evaluation measure which appears to be flexible to account for these discrepancies, at least to a certain extent (see section 4.2). The case remains, however, that extensive misalignments of this kind may severely compromise the NLG-related corpus-based evaluation of IO.

An innovative solution to the aforementioned problems comes from [Dimitromanolaki and Androutopoulos, 2003] who not only made the human authors aware of the assumed input to IO but also exercised extensive control over the way this input is realised as surface text. First, they derived many sets of facts from the database of their NLG system representing a hypothetical input to IO in their domain. After each fact has been realised as a stand-alone sentence (i.e. without performing any pronominalisation or aggregation), a domain expert was asked to order the sentences in each set and these orderings served as the basis of the evaluation of their IO method. A subset of their data was also employed in the experiments of [Karamanis *et al.*, 2004a] who evaluated different methods for IO, albeit in the same domain. Figure 1

<b>Authoring:</b>	Multi-authored	Parallel-authored
<b>Representativeness:</b>	Assumed	Verified
<b>Papers:</b>	[Duboue and McKeown, 2002], [Barzilay and Lee, 2004], [Karamanis <i>et al.</i> , 2004b], [Barzilay and Lapata, 2005]	[Lapata, 2003], [Karamanis and Mellish, 2005]

Table 2: Classification of reviewed papers in terms of authoring methods and representativeness of the employed corpus

(from [Karamanis *et al.*, 2004a]) shows an example set of sentences corresponding to database facts in the order defined by the expert.

Translating facts to sentences makes it feasible to produce a corpus appropriate for NLG-related evaluation by consulting subjects who do not have to become familiar with the internal representation used in a specific system. Moreover, the ordering of sentences defined by the humans can be compared with the output of various IO methods operating on the corresponding facts. However, due to the lack of aggregation and pronominalisation, although this unconventional corpus may consist of well-ordered sentences, it could still be far from fluent and will probably be of limited use for other corpus-based research.

Overall, it seems that using corpora for the MDS-related evaluation of IO is rather straightforward. This is not the case for the NLG-related evaluation efforts due to the different assumptions made about the input to IO. Hence, producing resources which are suitable for the NLG-related corpus-based evaluation requires considerable effort and may give rise to a collection of rather unconventional data.

### 3 Corpus authoring and representativeness

In addition to the issues related to corpus annotation, another well known problem in the evaluation of text production systems is that the same information may be communicated successfully in many ways [Mellish and Dale, 1998]. As far as IO is concerned it is clear that the ordering attested in the corpus might not be the only good way of organising the underlying information. Although this appears to severely challenge the view of the corpus text as a gold standard, it is also widely acknowledged that presenting information randomly gives rise to a vast number of inappropriate orderings which are very unlikely to manifest in a corpus. Hence, a corpus can provide useful information, especially if it is *representative* [McEnery, 2003], i.e. manifests a range of solutions available to an author. Crucially, evaluation results produced by averaging over a representative corpus are less likely to be biased in favour of a particular IO strategy [Duboue and McKeown, 2002].

Most reviewed studies make use of a *multi-authored* corpus, i.e. texts which are not all written by the same author. The larger the multi-authored corpus, the more representative it is likely to be. Hence, the corpora used for evaluation in [Barzilay and Lee, 2004] and [Barzilay and Lapata, 2005], each of which consists of 100 texts in more than one domain, are expected to provide more coverage than the corpora of [Duboue and McKeown, 2002] and [Karamanis *et al.*,

2004b] (consisting of just 23 and 20 texts respectively).<sup>3</sup> By contrast, although the corpus collected by [Dimitromanolaki and Androutsopoulos, 2003] is large (880 sets of ordered sentences), it is authored by one domain expert only (whom we will henceforth refer to as E0), raising the question whether the ordering preferences on which their evaluation is based on are shared by other experts.

[Karamanis and Mellish, 2005] attempt to verify the generality of the data created by E0. Similarly to [Lapata, 2003], they constructed a *parallel-authored* corpus by asking three more experts to produce additional orderings using the same materials as E0. The measure discussed in section 4.2 was used to investigate the extent to which the experts agree with each other. [Karamanis and Mellish, 2005] report that two out of the three experts they consulted agreed with each other as well as with E0 to a great extent, thus verifying the reliability of the corpus initially collected by [Dimitromanolaki and Androutsopoulos, 2003].

Although disagreement between authors cannot (and in fact should not) always be eliminated [Mellish and Dale, 1998], a well designed study like the one carried out by [Karamanis and Mellish, 2005] is likely to control conditions so that a basis for agreement is found amongst the authors. Like [Lapata, 2003], this study uses human agreement as the upper bound in the corpus-based evaluation of automatic IO methods which again provides an alternative to the view of the corpus text as an absolute gold standard.

Table 2 presents a classification of the reviewed work in terms of the authoring method and representativeness of the employed corpora. Clearly, a parallel-authored corpus represents the range of choices available to an author in a much more straightforward way than a multi-authored corpus, which presumably explains why representativeness has been subject to experimental investigation in the former (but not the latter) case as far as IO is concerned.<sup>4</sup> However, similarly to standard psycholinguistic experiments, consulting many informants can be easily done on a small scale but is more difficult to extend to a larger collection of texts.

Thus, [Karamanis and Mellish, 2005], who had to rely on expert advice only, collected three additional orderings for just 16 sets of sentences from the initial collection of [Dimitromanolaki and Androutsopoulos, 2003] (that is, less than 2% of total). Similarly [Lapata, 2003], who used non-expert informants and the web-based interface of [Barzilay *et al.*, 2002], collected a larger number of parallel sentence orderings (approximately 33 per text) for 12 texts in total which were randomly selected from a corpus consisting of more than 98,000 texts.<sup>5</sup>

<sup>3</sup>However, as already suggested, there is a tradeoff between corpus size and the reliability of the employed annotation methods.

<sup>4</sup>Clearly, the representativeness of a multi-authored corpus can also be investigated e.g. by comparing the choices different authors make in similar situations [Reiter and Sripada, 2002].

<sup>5</sup>[Lapata, 2003] also made use of the parallel-authored corpus collected by [Barzilay *et al.*, 2002] which consists of 10 parallel orderings for a total of 10 texts.

<b>Perf. measure:</b>	search-oriented	distance-based
<b>IO method:</b>	non-deterministic	any method
<b>Papers:</b>	[Karamanis <i>et al.</i> , 2004a; 2004b], [Barzilay and Lee, 2004], [Barzilay and Lapata, 2005],	[Duboue and McKeown, 2002], [Lapata, 2003], [Karamanis and Mellish, 2005],

Table 3: Classification of reviewed papers in terms of the employed performance measure and the IO method it may evaluate

## 4 Devising performance measures

As mentioned in the introduction, perhaps the most novel aspect of the reviewed work is the attempt to evaluate IO quantitatively using a performance measure. To discuss these measures in detail, we first need to distinguish between non-deterministic and deterministic methods for performing IO. Non-deterministic IO methods, assumed by some reviewed evaluation efforts, select the best ordering of information-bearing items among various alternatives on the basis of scores assigned by a metric of coherence [Karamanis *et al.*, 2004a; 2004b] or a probabilistic language model [Barzilay and Lee, 2004; Barzilay and Lapata, 2005]. The aim of the evaluation in this case is to estimate how well the metric (or model) will perform for the purposes of IO even before it is used to produce any actual output. This gives rise to the first type of measures which are search-oriented and estimate *how likely* a non-deterministic IO method is to produce the attested ordering in the corpus as its preferred output.

Most of the remaining evaluation efforts test deterministic IO methods which produce an ordering without explicitly comparing it with its alternatives. Although these cannot be evaluated in terms of the search-oriented measures, their performance can be estimated in terms of distance-based measures which signify *how close* the generated ordering stands when compared with a human defined ordering. Notably, the second set of measures can be applied to the evaluation of non-deterministic methods too as shown in [Karamanis and Mellish, 2005]. Table 3 classifies the reviewed papers in terms of the performance measure employed and the IO method this measure may evaluate.

### 4.1 Search-oriented evaluation measures

As already mentioned, the corpus-based experiments of [Karamanis *et al.*, 2004a; 2004b] take place prior to the actual generation of an ordering of information-bearing items. Their aim is to identify the most suitable metrics among the many candidates that can be used for the non-deterministic IO method they assume. To distinguish between the metrics, they introduce a performance measure called the *classification rate* of a metric of coherence  $M$  on a gold standard ordering (abbreviated here as GSO) which estimates how likely  $M$  is to produce the GSO as the output of IO.

To calculate the classification rate (signified as  $v$ ),  $M$  first assigns a score to the GSO. Then, the space of alternative orderings (defined by the permutations of the sentences that the GSO consists of) is searched. Each alternative ordering is scored and its score is compared with the score given to the GSO.  $\text{Better}(M, \text{GSO})$  stands for the percentage of orderings that are found to score better than the GSO according to  $M$ , whilst  $\text{Equal}(M, \text{GSO})$  is the percentage of orderings that score equal to the GSO according to  $M$ . The formula:

$$v = \text{Better}(M, \text{GSO}) + \frac{\text{Equal}(M, \text{GSO})}{2}$$

expresses the classification rate of  $M$  on the GSO as the expected percentage of alternative orderings which will have a higher chance than the GSO to be selected as the output of IO should  $M$  be used to drive this process under the scenario assumed by [Karamanis *et al.*, 2004a; 2004b].<sup>6</sup> The lower the classification rate, the better  $M$  is found to perform.

[Karamanis *et al.*, 2004a; 2004b] discuss how the performance of a metric can be estimated using many GSOs from a multi-authored corpus and show how different metrics of coherence can be compared with each other using the classification rate as their performance measure. An experimental methodology is detailed which consists of a set of pairwise comparisons that employ the Sign Test to test significant differences between the metrics.

[Barzilay and Lapata, 2005] and [Barzilay and Lee, 2004] also search the space of alternative orderings of the sentences that the GSO consists of. [Barzilay and Lapata, 2005] introduce a probabilistic ranking model which, like the metrics of [Karamanis *et al.*, 2004a; 2004b], compares the GSO with each of these orderings. The performance measure they introduce is called *ranking accuracy* and expresses the percentage of alternative orderings that are ranked lower than the GSO by their model. In terms of [Karamanis *et al.*, 2004a; 2004b], the ranking accuracy equals  $100\% - \text{Better}(M, \text{GSO})$ .<sup>7</sup>

A different stochastic model is used in [Barzilay and Lee, 2004] to compute the probability of generating the GSO and each alternative ordering in the explored search space. Then, all orderings are ranked according to this probability and the rank given to the GSO is retrieved. The *average GSO rank*, i.e. the average rank given to the GSO across the documents in a certain domain, serves as the first performance measure of their model. The GSO rank reports the mere number of alternative orderings that appear between the GSO and the top ranked ordering for each text without considering how many orderings stand beneath the GSO (which is a variable taken into account in the computation of the classification rate).

For instance, the worst average rank reported in [Barzilay and Lee, 2004] is 15.38 in the “drugs” domain which consists of texts made of 10.3 sentences on average, while the standard deviation of the number of sentences per text in this domain is 7.5 sentences. Assuming that no alternative ordering is assigned the same probability as the GSO, let us first consider the case in which the GSO is ranked as the 15th best permutation in a text that consists of e.g. 5 sentences. The population of alternative orderings in that case is 120 and the resulting classification rate is 12.5%.

However, if the GSO comes from a text consisting of e.g. 10 sentences, the same rank corresponds to a classification rate of less than  $2 \cdot 10^{-7}\%$ . Hence, using the classification rate as the performance measure makes it possible to state quite safely that in the second case the relative amount of

<sup>6</sup>A detailed justification of the cited formula in relation to the assumed scenario is presented in chapter 5 of [Karamanis, 2003].

<sup>7</sup>In contrast to [Karamanis *et al.*, 2004a; 2004b], neither [Barzilay and Lapata, 2005] nor [Barzilay and Lee, 2004] consider the possibility of the GSO scoring the same as another ordering.

alternative orderings which score higher than the GSO is much smaller than in the first example (perhaps giving rise to an average performance value across both documents). By contrast, using the GSO rank as the only performance measure does not distinguish between the two cases at all.

While [Barzilay and Lee, 2004] attempt to exhaustively enumerate billions of alternative orderings, the other experimenters deal with much smaller search spaces. As the search space grows factorially, exhaustive enumeration of orderings will unavoidably become impractical. However, [Karamanis *et al.*, 2004a] present detailed arguments that search-oriented evaluation can be practically limited to a sample of  $10^6$  alternative orderings irrespective of the actual size of the search space.

So far, there has not been any attempt to apply the search-oriented evaluation measures to a parallel-authored corpus. Hence, the generality of some results might be questioned especially when they are based on a small corpus as e.g. in [Karamanis *et al.*, 2004b]. Equally important is the fact that, although [Barzilay and Lee, 2004] and [Barzilay and Lapata, 2005] experiment on reasonably large corpora, they do not report any statistical tests to verify that the observed differences are significant.

To compare their model with the deterministic IO method of [Lapata, 2003], [Barzilay and Lee, 2004] additionally introduce the *GSO prediction rate*, i.e. the percentage of documents in the corpus in which the GSO is ranked first. This measure expresses how often their model reproduces the GSO across the texts in a corpus and does not actually require looking at any other ordering in the search space. However, as the GSO does not always get the highest rank, three rank ranges are also introduced to report the percentage of documents in which the GSO is ranked a) between 1st and 4th b) between 5th and 10th and c) below the 10th position.

In essence, all four search-oriented measures reviewed in this section (classification rate, ranking accuracy, average GSO rank, rank range) are trying to account for the possibility that an IO method might produce an ordering different from the GSO, which can be tested the method is actually used to produce an ordering due to its non-deterministic nature. By penalising the IO method proportionally to its failure to promote the GSO as the best scoring ordering, its performance can be compared with other ways of performing IO within and across domains.

## 4.2 Distance-based evaluation measures

Deterministic IO methods cannot be evaluated using a search-oriented measure. To evaluate their classification-based IO method [Dimitromanolaki and Androutsopoulos, 2003] measure the percentage of correct selections at each position made by their program compared to the selection made by the human author. However, as they acknowledge, ordering a sequence of sentences is not a series of independent decisions. Hence, calculating precision at each position does not appear to be the most appropriate evaluation measure. By contrast, measures which estimate automatically *how close* the ordering of information-bearing items actually produced by an IO system stands in comparison to the ordering attested in a corpus represent a better solution.

[Duboue and McKeown, 2002] employ a computationally intensive global alignment measure (typically used to compare DNA sequences) which is computed according to the Needleman-Wunsch Algorithm (NWA) as defined in [Durbin *et al.*, 1998]. This measure allows them to compare a computer-generated ordering of facts with the ordering of facts in the corresponding human text even if the latter consists of fewer, more, or even distinct facts. The main problem with the NWA is that it relies on predetermined scores for aligning pairs of facts (or aligning a fact with a gap). As [Duboue and McKeown, 2002] do not discuss how these scores are actually specified, we are not aware of any indication of how this can be done.

[Lapata, 2003] was the first to present an experimental setting which employs the *distance between two orderings* to estimate automatically how close a sentence ordering produced by her probabilistic IO method stands in comparison to an ordering provided by a human judge. Based on the argumentation in [Howell, 2002], [Lapata, 2003] selects Kendall's  $\tau$  as the most appropriate measure to estimate this distance. Kendall's  $\tau$  is based on the number of *inversions* between the two orderings and is calculated as follows:

$$\tau = 1 - \frac{2I}{P_N} = 1 - \frac{2I}{N(N-1)/2}$$

$P_N$  stands for the number of pairs of sentences and  $N$  is the number of sentences to be ordered.<sup>8</sup>  $I$  stands for the number of inversions, that is, the number of adjacent transpositions necessary to bring one ordering to another. Kendall's  $\tau$  ranges from  $-1$  (inverse ranks) to  $1$  (identical ranks). The higher the  $\tau$  value, the smaller the distance between the two orderings.

In the simplest case, [Lapata, 2003] computes the *average*  $\tau$  score (denoted as  $T$ ) to compare orderings produced by different versions of her system with the orderings found in a multi-authored corpus and employs the Tukey test to investigate significant differences between the scores.<sup>9</sup>

However, what makes the measure in [Lapata, 2003] particularly appealing is that it has also been applied to parallel-authored corpora which enable one to investigate the range of solutions for IO much more straightforwardly than multi-authored ones. [Lapata, 2003] first computes the average distance between all human defined orderings which represents the upper bound of the evaluation. This score is then compared with the distance between the orderings produced by (different versions of) her model and the human orderings.

Perhaps due to space restrictions, the way that the average  $T$  scores are computed is somehow glossed over in [Lapata, 2003]. A more detailed account of this evaluation procedure is presented in [Karamanis and Mellish, 2005] who applied her methodology to verify the generality of the data collected by [Dimitromanolaki and Androutsopoulos, 2003] as already mentioned in section 3. Additionally, [Karamanis and

<sup>8</sup>Like all other reviewed measures (except for the one employed in [Duboue and McKeown, 2002]), Kendall's  $\tau$  is meant to compare different orderings of identical items.

<sup>9</sup>Table 2 of [Barzilay and Lee, 2004] also reports evaluation results using Lapata's measure, but these are not discussed at all in their "Results" section.

Mellish, 2005] employ T measures for the evaluation of coherence metrics used under a non-deterministic IO method, building upon [Karamanis *et al.*, 2004a].<sup>10</sup>

## 5 Conclusion

To sum up, the review of the emerging corpus-based evaluations for IO in MDS and NLG leads us to the following conclusions: First, producing a corpus suitable for the NLG-related evaluation of an IO method is considerably demanding because of the need to represent unconventional input data. This is not the case for MDS-related evaluations which rely on features that can be easily annotated.

Despite the differences in input representations, in this review we were able to identify two general types of corpora as well as two types of evaluation measures which appear to be particularly helpful for the automatic evaluation of IO irrespective of whether it takes place within NLG or MDS. We treat this consensus as an indication that the corpus-based evaluation of IO is indeed feasible and might enjoy considerable popularity in future work.

More specifically, it is reasonable to expect that the distinctions and measures discussed in sections 3 and 4 might become standard in the MDS-related corpus-based evaluation of IO. Provided that the problems discussed in section 2.1 are looked at more closely, the same thing could happen in NLG-related evaluation as well.

## Acknowledgments

Many thanks to Laura Hasler, Constantin Orasan and Victor Pekar for helpful discussions and to Mirella Lapata for providing us with her latest paper.

## References

- [Bangalore *et al.*, 2000] Srinivas Bangalore, Owen Rambow, and Steven Whittaker. Evaluation metrics for generation. In *Proceedings of INLG 2000*, pages 1–8, Israel, 2000.
- [Barzilay and Lapata, 2005] Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. In *Proceedings of ACL 2005*, 2005.
- [Barzilay and Lee, 2004] Regina Barzilay and Lillian Lee. Catching the drift: Probabilistic content models with applications to generation and summarization. In *Proceedings of HLT-NAACL 2004*, pages 113–120, 2004.
- [Barzilay *et al.*, 2002] Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55, 2002.
- [Dimitromanolaki and Androutsopoulos, 2003] Aggeliki Dimitromanolaki and Ion Androutsopoulos. Learning to order facts for discourse planning in natural language generation. In *Proceedings of the 9th European Workshop on Natural Language Generation*, Budapest, Hungary, 2003.
- [Duboue and McKeown, 2002] Pablo Duboue and Kathleen McKeown. Content planner construction via evolutionary algorithms and a corpus-based fitness function. In *Proceedings of INLG 2002*, pages 89–96, Harriman, NY, USA, July 2002.
- [Durbin *et al.*, 1998] Richard Durbin, Sean Eddy, Anders Krogh, and Greame Mitchinson. *Biological Sequence Analysis*, pages 17–28. Cambridge University Press, 1998.
- [Gaizauskas, 1998] Robert Gaizauskas. Evaluation in language and speech technology. *Computer Speech and Language*, 12(4):249–262, 1998.
- [Hirschman and Mani, 2003] Linette Hirschman and Inderjeet Mani. Evaluation. In Mitkov [2003], chapter 22, pages 414–429.
- [Howell, 2002] David C. Howell. *Statistical Methods for Psychology*. Duxbury, Pacific Grove, CA, 5th edition, 2002.
- [Karamanis and Mellish, 2005] Nikiforos Karamanis and Chris Mellish. Using a corpus of sentence orderings defined by many experts to evaluate metrics of coherence for text structuring. In *Proceedings of ENLG05*, Aberdeen, UK, 2005. Poster session.
- [Karamanis *et al.*, 2004a] Nikiforos Karamanis, Chris Mellish, Jon Oberlander, and Massimo Poesio. A corpus-based methodology for evaluating metrics of coherence for text structuring. In *Proceedings of INLG04*, pages 90–99, Brockenhurst, UK, 2004.
- [Karamanis *et al.*, 2004b] Nikiforos Karamanis, Massimo Poesio, Chris Mellish, and Jon Oberlander. Evaluating centering-based metrics of coherence using a reliably annotated corpus. In *Proceedings of ACL04*, pages 391–398, Barcelona, Spain, 2004.
- [Karamanis, 2003] Nikiforos Karamanis. *Entity Coherence for Descriptive Text Structuring*. PhD thesis, Division of Informatics, University of Edinburgh, 2003.
- [Lapata, 2003] Mirella Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL 2003*, pages 545–552, Sapporo, Japan, July 2003.
- [McEnery, 2003] Tony McEnery. Corpus linguistics. In Mitkov [2003], chapter 24, pages 448–463.
- [Mellish and Dale, 1998] Chris Mellish and Robert Dale. Evaluation in the context of natural language generation. *Computer Speech and Language*, 12(4):349–373, 1998.
- [Mitkov, 2003] Ruslan Mitkov, editor. *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2003.
- [Poesio *et al.*, 2004] Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. Centering: a parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363, 2004.
- [Reiter and Dale, 2000] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Cambridge University Press, 2000.
- [Reiter and Sripada, 2002] Ehud Reiter and Somayajulu Sripada. Should corpora texts be gold standards for NLG? In *Proceedings of INLG 2002*, pages 97–104, Harriman, NY, USA, July 2002.
- [Walker *et al.*, 2002] Marilyn A. Walker, Owen Rambow, and Monica Rogati. Training a sentence planner for spoken dialogue using boosting. *Computer Speech and Language*, 16:409–433, 2002.

<sup>10</sup>Notably, this is the only domain in which both average T scores on a parallel-authored corpus as well as the classification rate, albeit on single-authored corpus, have been reported.