Intermixing Multiple Discourse Strategies for Automatic Text Composition

Mick O'Donnell
University of Edinburgh

Abstract: How we compose a text is not a well-understood process. This paper explores Enkvist's notion of text composition through the application of multiple discourse strategies. To write a text, we don't follow a single strategy, but rather intermix a number of strategies as required by the context of writing. This approach to text composition is then applied to automatic text composition, a system for producing on-line descriptions of museum artefacts on the fly.

1. Introduction

This paper addresses the question of how we compose a text. The initial stimulus for the work arose in the ILEX project, [1] at the University of Edinburgh. The goal of this project was to provide an on-line museum, allowing people to access descriptions of museum exhibits via the web. These descriptions were not to be written by hand, but rather composed by the computer using the museum's database of information about these exhibits. ILEX can be seen at: *http://www.cstr.ed.ac.uk/cgi-bin/ilex.cgi*.

Exactly how humans compose text is not a well understood process. Systemics includes analyses of many aspects of text structure including: Generic Structure Potential (GSP -- Hasan 1978); Conjunctive Relations (Halliday and Hasan 1976; Martin 1983, 1992: etc.); Rhetorical Structure Theory (RST -- Mann and Thompson 1987) or Focal Progression (Dane_ 1974; Fries 1995; etc.). However, none of these analyses by themselves tell us much about *how* we compose a text.[2]

In this paper, I wish to describe the approach we took to composing one type of text (the above-mentioned museum artefact descriptions), with the belief that the approach is extendible to other text-types.

While not used in the actual implementation, Enkvist's (1987) discussion of "text production strategies" will be used to structure the description of our approach to text composition. Enkvist (1987) proposes that to understand the text composition process we need to understand that, for many texts, there is not a single discourse strategy at play, but that a writer will draw upon a number of different strategies to produce the text. He provides a number of examples of text-types which use a blend of different text-arranging strategies. For instance:

> A guidebook is arranged by place and sight, and is thus locative-dominated; but within the divisions of the text signalled by shift of place or sight, a guidebook can provide information of different kinds, for instance in the form of chronologically arranged narrative. (Enkvist 206)

Enkvist is talking about more than genre-mixing here, but rather about the existence in a genre of a number of component strategies which together compose the genre. The skill in applying the genre well is in the alternation between the multiple strategies, avoiding the conflicts which may arise out of the different strategies, and using the compatibilities.

This description of the text composition task fits well to the process we developed in the ILEX project, better than the descriptions we have previously provided ourselves. I will thus attempt to re-package our work in terms of a multiple strategy text production algorithm.

In section 2, I will review our corpus of museum artefact descriptions, extracting out some of the repeating discourse strategies, and how these strategies can work together. A sub-set of these strategies are used by ILEX to generate our own descriptions.

Before detailing the multi-strategy text composition process, I will outline how we represent the potential content of descriptions, the *ideational potential* (section 3). Much of the ensuing discussion will take this knowledge as a starting point.

Sections 4 and 5 will begin the discussion of text composition, showing how the text generator can make use of multiple discourse strategies, intermixing them, to produce text which is reasonably natural for the genre.

Section 6 will then discuss how the discourse level is realised into text, via rhetorical structure trees and syntactic structure.

2. Composition Strategies in Museum Artefact Descriptions

To explore the set of strategies commonly used in the museum artefact description genre, we collected a corpus of descriptions from several sources, including recording a curator as she gave a guided tour; from books describing museum exhibits; and from virtual museums (museums on the web).

A typical description of the genre is as follows:

> **The 98.6 Carat Bismark Sapphire**: A spectacular close-up of the Bismark Sapphire, one of the world's largest. This 98.6 carat gem is exceptionally large and well coloured. Originally from Sri Lanka, it is part of the collection of the National Museum of Natural History. Also shown are many of the diamonds which encrust both the stone and it's necklace. (Smithsonian Institution)

Analysing this corpus, we recognised a number of repeating strategies. Some of these are listed below:

1. **List**: the basic structure of the genre. The strategy is to provide a list of the facts available about the entity being described, one after another. The entity itself is in generally in Subject position, although fronted Circumstances may act as Theme. These facts range over a number of possible details, such as:

   - **Classify**: provide the class of the entity, e.g., *This item is a brooch.* (The class is often given in the title of the exhibit, rather than the text, or within a nominal reference, e.g., *this 98.6 carat gem*).
   - **Define**: provide details which uniquely distinguish the entity, e.g.,

     > A Communion Token is a simple metal ticket which permitted the holder to partake of Communion in the Church of Scotland and other Presbyterian Churches. (Hunterian Museum)

   - **Describe**: provide details of the physical appearance of the entity, or its materials, etc. For example:

     > This 98.6 carat gem is exceptionally large and well coloured. (Smithsonian Institution)

     > It's bits of cut-off razor blade, biro, knitting needles, inlaid into layer after layer of resin. (Goring)

   The strategy usually entails grouping similar types of information, so that classification/definition tends to come first, followed by physical description, and after that other aspects such as function, etc.

2. **Digress**: shift the focus to an entity introduced in one of the facts in a list, e.g., in the following, the focus shifts from the necklace to Liberty:

   > This is one of the necklaces in this case which was made for Liberty & Co. Liberty are a company based in London, in Regent St, who were really at the interface between mass-produced jewellery and `craft' jewellery; one-offs. They used the very best designers to design jewels for them, which were then produced in fairly limited quantity, but in quantity rather than as one-offs. (Goring)

3. **Aggregate**: express multiple elements of a list as a single sentence, e.g.,

   > Originally from Sri Lanka, it is part of the collection of the National Museum of Natural History. (Smithsonian Institution)

4. **Generalise**: shift from the current focus to discussing the class of entities which the current fact somehow introduces. There are two common methods for generalising in this way. Firstly, after classifying an item (e.g., *This item is a brooch*), shift to discussion of the class (e.g., *Brooches are …*). The second case is a bit more complex. After providing a fact, we can shift focus to the class of entities for which the fact is also true. For instance: *This item is in the Art-Deco style. Art-Deco jewels are often made using enamel.* The first sentence creates a general class, those jewels for which the fact is true: *Art-Deco jewels*. The second sentence picks up on this general class and expands on it.[3]

   This strategy is similar to a digression, except in a digression, the new focus was introduced in the original fact, while in a generalisation, the new focus is derived from the original fact as a whole.

   Generalisation is an important strategy for the museum label genre, because, for most curators, artefacts are not important in themselves, but rather as places to attach more general cultural information, regarding society, technology, etc. The artefact is the launching point for the generalisation.

5. **Compare**: compare and contrast the artefact with other similar artefacts.

   > The four pieces here actually show four quite distinct aspects of Jessie King's jewellery work. For instance, number 3: the very strong blue enamel colour, and the cloud forms; the festoons on number 1, that very feminine style of jewellery, that's beautiful work. And number 2, the very subtle colourings. (Goring)

Many other strategies are observed in the corpus, but most of them do not appear on a regular basis, e.g., contrast, exemplification, restatement, etc. We won't mention them here, as the selection of strategies listed above will be enough to demonstrate the ideas of this paper. This selection of strategies still allows the production of text of reasonable complexity, and demonstrates the notion of intermixing these strategies as a means of text composition.

None of the strategies outlined above are unique to the museum artefact description genre. Each strategy can be used in a wide range of genres. However, a particular genre will use these strategies in different ways: in one genre the strategy may be dominant, providing the overall shape of the text, while in another, the strategy is secondary, only coming into play when the opportunity permits. Returning to Enkvist's example, a guide-book is primarily organised by a spatial strategy, but can make use of a temporal (narrative) strategy within the spatial constraint. A different genre, say a chronicle, uses a temporal sequencing strategy, but may use a spatial organising strategy within the sequencing constraint.

We could say that a genre is defined in this way, by the set of discourse-forming strategies that it regularly uses, the way in which they can combine, and the order of dominance between them.

Discourse strategies, as I define them, sometimes involve the addition of a single element to a text, connected via a rhetorical relation (as used in RST). However, a

strategy may be more complex, involving a complex of rhetorical relations (see the Generalise/Exemplify strategy later), or involving a multi-part structure such as used in GSP. This use of strategies thus includes both RST and GSP, and possibly other text-forming strategies as well. Also, a strategy is not restricted to a single stratum, potentially involving constraints at discourse, syntactic and other levels.

3. The Ideation Potential

Before I enter into discussion of how these strategies are used to compose a text, I will introduce a means of representing the *ideational potential*. We will use this representation to exemplify the discussions that follow.

3.1 The Ideational Potential

The i*deation potential* represents the information we have to express about the various exhibits in the museum, their appearance, their designers and makers, the materials they are made of, the places they are made in, etc. We need to keep in mind that what we have to say is a major constraint on the text we compose.

For ILEX's first application, we decided to build a virtual gallery based on a single gallery in Edinburgh´s Royal Museum of Scotland. The *Modern Jewellery* gallery was selected. The museum provided us with their database for the gallery.

For our purposes, a pure Systemic implementation would represent this information in terms of processes, participants, and circumstances.[4] However, the data we received from the museum was a relational database, with information on each artefact, such as:

```
item:       j-990656
class:      necklace
designer:   King01
date:       1906
style:      "Arts and Crafts"
material:   gold
material:   enamel
```

To simplify the task of expressing this kind of information, we elected to represent our ideational information in terms of relations between two entities, such as in:

```
class(j-990656, necklace)
designer(j-990656, King01)
etc.
```

The museum databases also provide information on designers. We have also enriched the database with information we have gathered from other references, such as books on modern jewellery styles and techniques.

We call each of these relations a *fact.* These facts, when taken together, form an interconnecting structure of facts, what we could call the *ideational potential* (being the set of facts that we could express). See Figure 1.

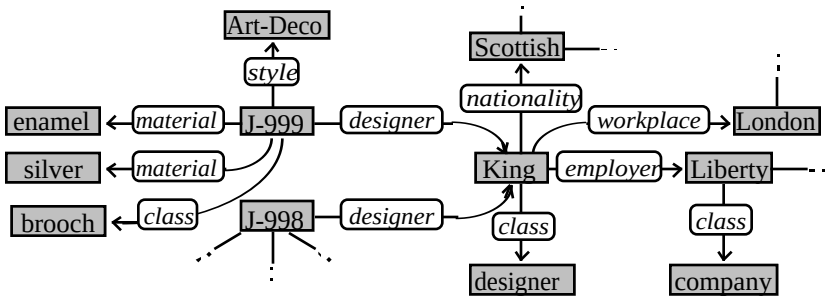

Figure 1: The Ideational Potential.

The square/grey boxes represent entities of the domain (jewels, people, nationalities, materials, etc.). J-998 and J-999 represent two jewels in this gallery. Each rounded box represents a fact, a relation between two entities.[5]

3.2 Adding Generalisations

From various sources, we have gained extra information about classes of entities, such as *Brooches are generally worn attached to the front*; or *Art-Deco jewellery tends to be made using enamel*.

We add this information into the ideational potential. We first create entity nodes for the general classes, and then create facts for the generalisations we know. For instance, see Figure 2, which introduces two general classes (`ADJewels` and `Brooch`), and one fact about each of them. Note that the rounded box around the facts are dotted, to represent that the facts are tendencies only: to be expressed via "usually" or "tends to". Most of our generalisations are of this kind.

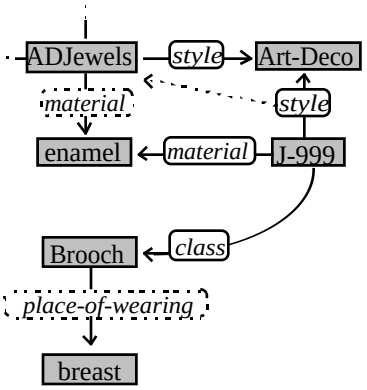

Figure 2: Adding Generalisations.

As stated in the discussion of generalisation in section 2, we have a link between a fact such as *This jewel is in the Art-Deco style*, and the general class of entities for which this fact is true. We thus show graphically a generalisation link from the instantial fact to the `ADJewels` entity (the dotted line).

3.3 Relations between Facts

Sometimes, we can deduce relations between facts. For instance, if we know that a particular jewel is in the Art-Deco style, and that it uses enamel, then we can assert an `Instantiation` relation between the generalisation and the fact that this jewel uses enamel. We can also assert a `Generalisation` relation in the reverse direction. This allows us to add an additional level to our ideational potential, connecting facts together

via potential relations. See Figure 3, where the relations between facts are indicated by bold labels.

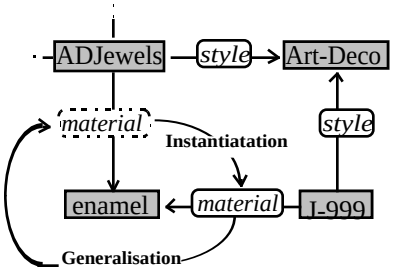

Figure 3: Adding Relations.

Note that these relations are to be taken as relations between *ideational units* not between *units of text*. Rhetorical relations, to be discussed below, are relations between units of text. Rhetorical relations may realise ideational relations however. For instance, a rhetorical relation such as `RST-Exemplify` can be used to realise the ideational `instantiation` relation.

4. Strategies for Generating Discourse Artefact Descriptions

To automatically generate museum artefact descriptions, we selected a subset of the discourse strategies observed in the corpus, those which were easiest to implement given the database information we had at hand.

We chose to use the *list* strategy to provide the basic structure of the description, and allow strategic modifications of this basic form via the applications of *digressions*, *generalisations*, *exemplifications*, *exceptions* and *aggregations*. An assortment of other strategies were also catered to.

This section will discuss each of these strategies, and show how the strategy relates to *movements through the ideational potential*. Figure 4, a small sample ideational potential, will be used to exemplify this discussion. Note that this potential represents a very small part of the actual ideational potential of our museum domain.

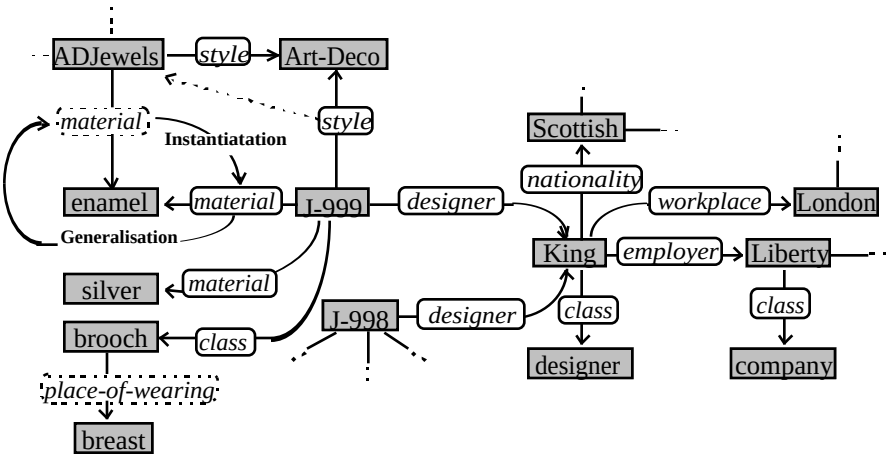

Figure 4: A larger subnet of the Ideational Potential.

4.1 List

The basic structure of our texts is a list of facts involving an entity. We call the object to be described the *page-focus.* Centering on this object in the ideational potential (Figure 4), we select all facts which begin at that entity. We select some of these to include in our list. These are then ordered in terms of a stated preference (class first, then facts of a describing nature, etc.). Using this strategy by itself would result in an object description for `J-999` like:

> *This item is a brooch. It is made of enamel. It is made of silver. It was designed by Jessie King. It is in the Art-Deco style.*

Apart from page-focus, we also represent a *local-focus*, the focus of the individual sentence. We realise local-focus by placing the entity in Subject position. In a list structure, all sentences have the same local-focus, as in the above example.

4.2 Digress

From any fact in a text, we allow a *digression* in order to add some background about that fact. In terms of the ideational potential, a digression involves taking the non-focal entity in a fact, and selecting one or more facts involving that entity. For instance, to digress from `designer(J-999,King)`, we can select one or more facts starting at entity `King`. If more than one digressing fact is selected, they are organised using a *list* strategy as outlined above. The digressing facts are in the normal case realised straight after the fact which introduced the digression.[6]

In a digression, the digressing entity becomes the local-focus, as can be seen in the following example:

> *This item is a brooch. It was designed by Jessie King. King was British. he worked in London. It is in the Art-Deco style.*

We also allow digressions on digressions, such as:

> *This item is a brooch. It was designed by Jessie King. King worked in London. London is in the South of England.*

However, digressions on digressions are disfavoured because they can quickly lead the text away from the object of interest.

4.3 Aggregate

Another strategy we allow to be applied to lists is called *aggregation*. Aggregation, a term quite common in computational linguistics, refers to the process of joining two separate clauses together into a single clause. In Systemics, we might call this *clause combining*, although as we will see below, that is only one possible form of aggregation.

Hua Cheng is responsible for the work on aggregation in ILEX. For more details, see Cheng (1998).

Two common types of aggregation include:

〈 **Clause Combining**: two or more facts are expressed as a clause-complex, either through parataxis or hypotaxis. We only deal with the paratactic case here. If both facts share some elements (Subject, Finite, etc.), then the shared elements can generally be dropped, e.g.,

King was Scottish and ~~he~~ lived in London.

It is made of silver and ~~it is made of~~ enamel.

- **NP Embedding**: facts which share an entity are expressed as modifiers within an NP which expresses the common entity, e.g.,

  This item was made by a British designer called Jessie M. King, who lived in London.

  This single clause, generated by ILEX, aggregates the `nationality` and `workplace` facts into the referring expression for King, producing text which is more succinct, and also more fluent, than a simple listing of facts.

4.4 Generalise

A Generalise strategy involves moving from any one fact to a fact which involves the generalised class which corresponds to that fact (see the discussion of generalisations in section 2).

In terms of the ideational potential, this involves a movement from a fact to a class entity following the dotted line which represents this relation.[7] This class entity then becomes focal, and one or more facts concerning the class of entities can be included.

Two common substrategies involve intermixing a generalise move with either an exemplify or exception relation:

- **Generalise/Exemplify**: *This item is in the Arts and Crafts style. Arts and Crafts style jewels usually have an elaborate design; for instance this jewel has floral motifs.*

- **Generalise/Exception**: *Arts and Crafts style jewels usually use oval-shaped stones but this jewel uses faceted stones.*

In both cases, the example or exception preferably relates back to the current focus, as in the above examples. However, if the database does not record whether the current item is an example or exception to the current example, then a recent item can be used, e.g.,

This item is in the Arts-and-Crafts Style. Arts and Crafts style jewels usually have an elaborate design (for instance the previous item has floral motifs).

We might even combine these two substrategies:

This item is in the Arts-and-Crafts Style. Arts-and-Crafts jewels tended to be hand-made. For example, this item was hand-made. However, the previous brooch was mass-produced.

4.5 Compare

Comparison is an alternative strategy to *list* in that it provides a top-level structuring to the text (while the other strategies here provide modifications to the basic list structure).

We provide comparisons to other items when there is nothing salient left to say about the current object. The program first locates a recently discussed object which is similar to the current one. It then provides a comparison using a schema: *Similarities ^ Differences*. A sample comparison generated by ILEX is shown below. Some work needs to be done to get the text more fluent:

> This pendant-cross resembles the previous item in that, like the previous item, it identifies the wearer as a Christian. However it differs from the previous item in that it was made in 1925, whereas the previous item was made in 1910.

The work on similarity and comparisons was implemented by Maria Milosavljevic (see Milosavljevic 1997; 1999).

4.6 Other Strategies

ILEX allows various other discourse-forming strategies to be used, all involving the linking two clauses together via a single rhetorical relation showing the relation between two facts. For instance, sometimes we know that one fact is more specific than another, so a `Specialise` relation is asserted between them in the ideational potential, and realised as an RST-Specialise relation in the rhetorical structure, e.g., *This item has an elaborate design in that it has floral motifs.* Similar relations are also possible for `Concession`, `Contrast`, etc.

5. Choosing which Discourse Strategies to Apply

So far we have just shown how the various discourse strategies which ILEX can use relate to linkages between facts in the ideational potential. The questions remains: for a good object description, which discourse strategies should we apply, and where in the text should we apply them.

There is no easy answer to these questions. Partially the answers are in the specification of the genre, which could state preferences for each discourse strategy, or perhaps contexts in which one discourse strategy is more valuable.

Partially also the answer is in the gallery's particular style. The type of exhibit label they desire will influence the degree to which each strategic resource is used. Are contrasts valued? or examples? Should lists be long or short? Should we stick to describing the object? Or are generalisations more important to the educational goals of the curators?

One of the requirements of the ILEX system was that the entire text should fit on the visible page (no scrolling of the web page was allowed). We are thus constrained to produce texts of fixed length.

Within that length, only so many facts can be realised. We thus have a case where the various discourse strategies are in competition for the limited screen real-estate. If we digress to talk about the designer, we can spend less time describing the jewel itself, or comparing it to similar jewels.

5.1 Selecting facts by Relevance

To decide which strategies are used on a page (and thus which facts appear on the page, and how they are organised), we use the notion of *relevance*: information included in the text must be in some way relevant to the goal of the discourse. Given that our goal is to describe an object, we can say that information can be included if it is relevant to describing the object.

We attempt to select content which is relevant both *ideationally* (how connected is the information to the object of focus?) and *interpersonally* (how important does the curator feel the information is? How interesting do we judge the information for the reader?). We will describe this approach below.

5.1.1 Ideational Relevance

We can define relevance in terms of the ideational potential: A fact is relevant to the entity of description if:

- it directly describes the entity, i.e., the fact is one which is linked directly to the entity;
- it provides background on an introduced participant, i.e., is linked to a relevant fact via a shared participant, e.g., when describing a brooch by naming its designer, facts providing background on the designer also become relevant.

〈 it provides background on the fact itself, i.e., the fact is linked to a relevant fact via a conceptual relation. The facts linked to a relevant fact via `Contrast`, `Generalisation`, `Instantiation`, etc., in our ideational potential all shine some more light on the relevant fact, and thus may be relevant to the description of the entity of focus.

Given the recursive application of these criteria, and the high degree of inter-connectedness of the ideational potential, it might be the case that *all* facts in the ideational potential are relevant on this basis. For this reason, we need to take into account *degrees of relevance* -- the further away from the focal entity we move, the less relevant the information becomes. A fact directly concerning the entity will have high relevance, while a fact which provides background to this fact will be of less relevance.

Let us assume that our content selection algorithm is given the task of selecting the *n* facts most relevant to some entity. A simple content selection model might simply select all facts at distance 1 from the entity (those involving the entity). If more facts are needed, the algorithm just selects facts at distance 2 (facts sharing a common entity, or related by a relation node to already selected facts). This process continues until enough facts are included.

ILEX improves on this simple process by taking into account the fact that some types of connections between facts preserve relevance better than others. It may be more desirable to include a fact connected via a `Generalisation` relation than one connected via a common entity. For this reason, we assign each relation type a level of relevance-maintenance, ranging from 1.0 (no lessening of relevance) to 0.0 (no preservation of relevance). If `Generalisation` is assigned a high level of relevance-maintenance, then facts connected by this relation will be included in the text before facts connected by weaker connections.

The ideational relevance of any fact is derived by multiplying together the relevance-maintenance levels of each link back to the entity being described. For instance, if a `Generalisation` relation has a level of 0.9 and a link through a common entity has a level of 0.5, then a generalisation on a fact linked through a common entity would have structural relevance of 0.45. If multiple paths exist, the lowest score is assumed.

This methodology is set out more clearly in O'Donnell (1997), where it is used for producing variable-length documents.

5.1.2 Interpersonal Relevance

Another factor we wish to capture in our content selection algorithm is that some facts are, in themselves, more worthy of including in the text. To calculate the intrinsic 'worth' of a fact, three factors are considered:

i. The (assumed) *interest* of the information to the user, e.g., jewellery made of plastic or paper is considered interesting, because such materials are unusual in jewellery.

ii. The *importance* of the information to the system's educational goals, e.g., the system considers it important to educate on stylistic development, so facts about styles are rated highly.

iii. The level of *assimilation* of the information -- the degree to which we can assume the reader already knows the information. This could be because the information is part of the reader's assumed world knowledge, or because we have previously told the user the information. Assimilation of a fact varies between 1.0 (fully assimilated) and 0.0 (totally unknown).

The three values interest, importance and (1 - assimilation) are multiplied together to calculate the interpersonal relevance of each fact. This represents how valuable the fact is to say, regardless of what object is currently being described.

We have no special theory about where interest and importance come from. User surveys could be done to produce values, or curators could produce estimates for various user types.

5.1.3 Combining Ideational Relevance and Interpersonal Value

To find the overall value of including the fact in the text, we need to combine the fact's interpersonal value with its ideational relevance. At present we just multiply the two together (each is between 0.0 and 1.0).

5.2 Using Relevance Ratings to Extract a Topic Tree

I have tried to show that each of the discourse strategies can be related to a particular way of linking facts within the ideational potential: a list strategy draws upon those facts linked to a single entity; a digression moves from one of these facts to a new entity and takes that entity as focal; a generalise move follows one of the links to a generalised entity, while exemplify, contrast, etc., depend on the relational link between one fact and another (comparison is one exception here; and aggregation will be treated below in the realisation section).

The content selection process ranks each fact by relevance. Because the path between facts plays a part in this (the structural relevance), we record, for each of the selected facts also the path of connection from the page focus to the fact (this is necessary because there may be alternative paths from the focus to a given fact).

We can represent the selected facts, and the connections between them, in terms of a *topic tree*, as in Figure 5. At the top of the tree, we have the page focus, the artefact being described: J-999. Three facts connected to that were selected as relevant. These facts form the basic list structure of the text.

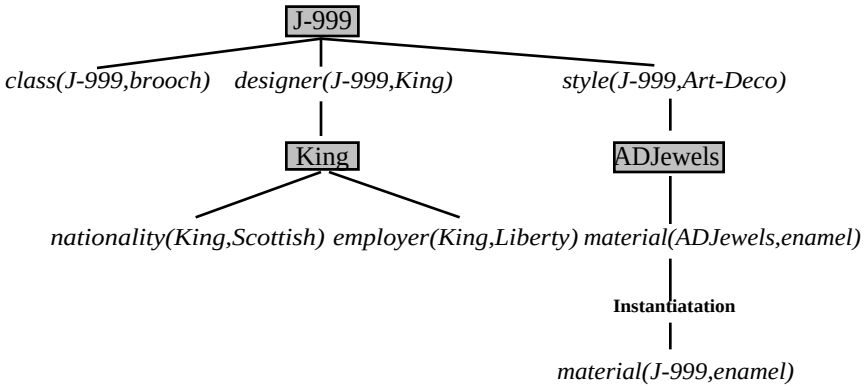

Figure 5: The Topic Tree.

The *designer* fact introduces a digression, whereby two facts concerning *King* are inserted. *King* becomes the local focus in the expression of these facts, which are also expressed as a list.

The *style* fact introduced a generalisation, changing the local focus to *ADJewels*. One fact about this class was considered relevant. Another fact, regarding the materials of *J-999*, was linked to this generalisation via an `Instantiation` relation. Note that this materials fact could have been included in the top-level list, except that we assign higher structural relevance to facts connected by relation links than those appearing in simple lists. By favouring relation structures over list structures, we produce more complex text, which feels more natural.

5.3 Applying Aggregation to the Topic Tree

Aggregation is an operation on the topic tree. We can choose to merge any facts in a list structure with either i) one of the other facts in that list structure, or ii) with the fact which introduced the list structure (if it is the result of a digression). In this instance, we chose to aggregate the *nationality* and *employer* facts..

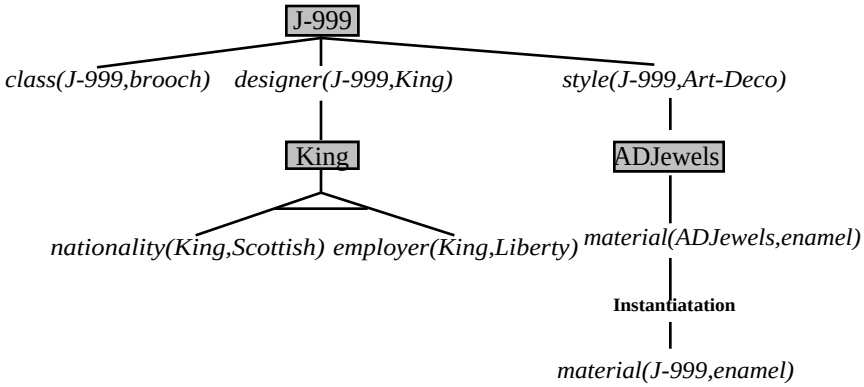

Figure 6: The Topic Tree after Aggregation.

6. Realising the Discourse Structure

The topic tree constructed in the last section is a long way from being a finished text. In this section, we very briefly discuss how the discourse structure is realised into text.

6.1 Building Rhetorical Structure

Once we have selected the facts we have to say, and structured them into a topic tree, we need to realise the facts as a rhetorical structure. The various discourse strategies each have defined mappings onto rhetorical structure. The List strategy maps onto an RST `Joint` structure. We model the digression as an `Elaboration` structure. The Generalise and Instantiate strategies map onto RST relations of similar names. The topic tree shown above would produce the rhetorical structure shown in Figure 7.

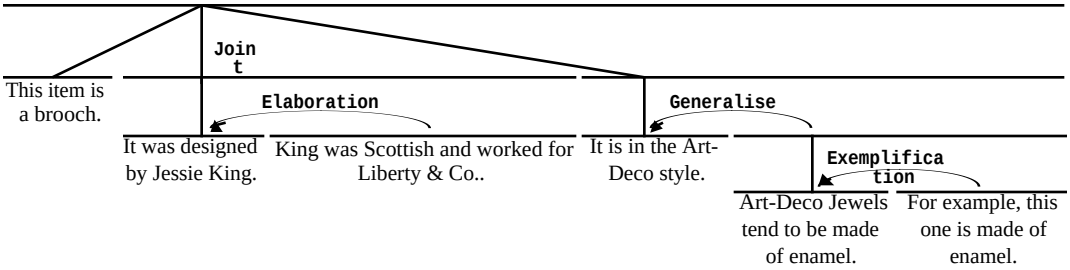

Figure 7: The Rhetorical Structure of the Text.

The RST structure is one step towards the realisation of the text. ILEX contains knowledge of how to express facts joined by various RST relations (as separate sentences, or as hypotactic or paratactic clause complexes), and inserts the appropriate conjunctive for the relation.

6.2 Sentence Realisation

The system then needs to realise each fact as a clause. ILEX contains a set of rules which map facts onto clause structure, allowing for variations in systems such as finite/non-finite clauses, voice, usuality, tense, etc. Sentence generation is done using the WAG sentence generator (O'Donnell 1996). The surface string (text) is then extracted from this structure.

6.3 Reference

Each reference to an entity needs to be expressed as an NP which adequately refers to the entity in context. ILEX can use a range of referring expressions to help create text which is quite natural, including definite reference: *this item, the previous item, the silver and enamel brooch;* indefinite reference: *a designer called Jessie King*; pronominal reference: *it, he,* etc. For more details see O'Donnell *et al* 1998.

7. Evolutionary Computing

One way in which Computational Linguistics is changing is in a movement away from modelling the process of language -- how is a text composed, towards modelling the product of language. Evolutionary computing is at the forefront of this change.

In the previous part of this paper, we outlined one *method* for constructing a text from basic content. The idea behind evolutionary methods is to randomly construct a first text, then to apply random mutations to the text, at each stage favouring the resulting texts which are evaluated as better texts. The process of random mutation and selection, when applied over thousands of generations, results in a text which is moving towards an optimal compositional structure (at least in regards to the evaluation metric used).

We start by constructing the text by arbitrarily selecting content to express from our knowledge source, and randomly sequencing the content. Rhetorical relations are asserted between these elements where possible. Clauses are realised with default thematic selections (e.g., active voice, any adjuncts after the verb).

This text is then evaluated using an evaluation metric. The metric is a function which rewards features of coherence we like, and penalises features we do not like. For instance, each instance of an illegal focal progression subtracts from the coherence score. The use of a strong rhetorical relation (e.g., exemplification), as opposed to a weak one (e.g., conjunction), is rewarded. Long distance between a satellite and its nucleus would also be penalised.

We then produce a number of offspring of this text, each produced by applying a mutation. A mutation might involve:

- Adding or deleting elements to a List structure:
- Adding or deleting a digression;
- Adding or deleting a generalisation:
- Change the sequence of clauses in the text;
- Change the rhetorical relation between two elements;
- etc.

Each of the resulting texts are then evaluated. This first round creates a *population* of texts. The process above is then repeated for each. At each round, a limited number of the population is selected for reproduction, those with a higher value having a higher

chance of being selected. In addition, a certain number of texts are deleted from the population at each round, being more likely for those with lower values.

The beauty of this approach is that it changes the focus of computational linguistics from one of *how do we compose a text?* to one of *how do we evaluate the quality of a text*? This focus is far more within the realm of the everyday linguist.

The process of mutation has its parallel in the human-writer process of *revision* – a change to the text to improve its coherence. However, a human will usually only apply a revision if it improves the text. The computer applies revisions without reference to improvement. However, texts which include poor revisions are less likely to reproduce, and thus favourable revisions will tend to win out. For an application of evolutionary computing to text composition in ILEX, see Mellish *et al.* (1998a).

8. Summary

We started out by asking "how do we compose a text?". Approaches which depend on a single theory of text structure, such as GSP, RST, method of development, etc., cannot answer this question by themselves. Instead, we posited that Enkvist's notion of text production using multiple discourse production strategies could be usefully applied to the task of text composition.

To test this hypothesis, we first explored a corpus of museum artefact descriptions to seek out the set of strategies which were usual in the genre. We then detailed a system for automatic text composition which makes use of multiple discourse strategies, and showed that the texts produced by the system are reasonably natural, yet can exhibit quite complex structures.

The following is a typical text produced by ILEX. It is not perfect, but quite good for machine-produced text:

> This jewel is a necklace and was made by Arthur and Georgie Gaskin. It uses green and white enamel, the colours of the suffragette movement. It is also in the Arts and Crafts style. It was made in 1910. It has an elaborate design; indeed Arts and Crafts style jewels usually have an elaborate design (for instance the pendant-necklace has floral motifs). This jewel was produced by a single craftsman (indeed Arts and Crafts style jewels were usually produced by a single craftsman).

Finally, we discussed one of the exciting new directions of change in the field of automatic text composition, the use of evolutionary computing, which shifts the emphasis from *how do we compose a text?* to *how do we evaluate what is a good text?*

The application of linguistic theories to computational problems can produce products which have commercial value. Various museums have shown interests in using the ILEX system, and the system is still being developed towards this end.

On the other hand, applying linguistic theories to computational problems also throws light on the linguistic theories themselves. There is a lot of description of linguistic analyses such as GSP, conjunction, reference, method of development, etc. However, a problem arises when we try to integrate these analyses into a single linguistic model. How does, for instance, method of development interact with the generic structure or a text, or with the Given/New structure.

In systems like ILEX, which need to explain most linguistic phenomena, these diverse approaches are integrated. This sometimes produces interesting observations of the linguistic theory, and show which of the theories are useful, and which are not.

9. Notes

I would like to thank Susana Murcia Bielsa for her comments on this paper.

1 The ILEX project was an EPSRC funded project, involving Chris Mellish, Jon Oberlander, Alistair Knott, Hua Cheng and the author. The implementation of the text composition process owes much to Alistair Knott. For descriptions of the ILEX project, see Knott *et al* (1997); Mellish *et al* (1998b); Oberlander *et al* (1997, 1998). Other researchers have contributed to the ILEX system, including Maria Milosavljevic and Janet Hitzeman.

2 Generic Structure Potentials (GSPs) can be used as a partial answer to text composition, and in fact has been used for automatic composition of artist's biographies in at least one work: Bateman and Teich 1995. However, such an approach will only work in highly regimented genres. While less regimented genres may demonstrate a degree of top-level schematic regularity, it is rarely possible to specify generic structure all the way down to the sentence level. GSPs by themselves do not explain how we compose a text.

3 This particular type of generalisation was recognised and described by Alistair Knott, see: Oberlander *et al.* (in press).

4 Other works by the author, such as the WAG sentence generation system (see O'Donnell 1994, 1995), assume a systemic semantics of this kind.

5 Our approach allows only facts concerning two entities, which is a limitation. However, the binary assumption simplifies our implementation. Complex sentences can however be formed via aggregation of simpler sentences.

6 Knott allows for a variation from this strategy if the insertion of the digressive material would disrupt the List structure too much. The digressing facts are placed after the list, as a new topic.

7 Alistair Knott, who introduced this type of strategy, calls this a *Predarg move*.

10. Works Cited

Bateman, John and Elke Teich. "Selective Information Presentation in an Integrated Publication System: an Application of Genre-driven Text Generation". **Information Processing and Management (Special Issue on Summarising Text)** 31(5) (1995): 753-768.

Cheng, Hua. 1998 "Embedding New Information into Referring Expressions". **Proceedings of COLING-ACL'98 (Student Session)**, Montreal, Canada, 10-15 Aug. 1998. 1478-1480.

Dane_, Frantisek. "Functional Sentence perspective and the Organisation of the Text." **Papers on Functional Sentence perspective.** Ed. Frantisek Dane_, The Hague: Mouton, 1974. 106-128.

Enkvist, Nils Erik. "Text Strategies: Single, Dual, Multiple." **Language Topics: Essays in Honour of Michael Halliday.** Volume II. Eds. Ross Steele and terry Threadgold. Amsterdam: Benjamins. 1987. 203-211.

Fries, Peter. "Themes, Methods of Development, and Texts." **On Subject and Theme: From the Perspective of Functions in Discourse**. Eds. R. Hasan and P. Fries. Amsterdam: John Benjamins, 1995. 317-359.

Goring, Elizabeth. Personal Interview. 18 Nov. 1995.

Halliday, MAK and Ruqaiya Hasan. **Cohesion in English**. London: Longman. 1976.

Hasan, Ruqaiya "Text in the Systemic Functional model." **Current Trends in Textlinguistics**. Ed. W. Dressler. Berlin: de Gruyter, 1978.

Hunterian Museum, University of Glasgow. "Coin Gallery." 24 February 1999. 19 January 2000. <http://www.gla.ac.uk/Museum/HuntMus/MoneyAndMedals/ComTok.html>

Knott, Alistair, Michael O'Donnell, Jon Oberlander and Chris Mellish. "Defeasible Rules in Content Selection and Text Structuring." **Proceedings of the 6th**

European Workshop on Natural Language Generation, March 24 - 26, 1997, Gerhard-Mercator University, Duisburg, Germany.

Mann, William C. and Sandra A. Thompson. "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization." **Text** 8 (1988): 243-281.

Martin, James R. "Conjunction: the Logic of English Text." **Micro and Macro Connexity of Discourse**. Eds. Janos S. Petöfi and Emel Sözer. Hamburg: Buske (Papers in Text Linguistics 45), 1983.

- - -. **English Text: System and Structure**. Amsterdam: Benjamins. 1992.

Matthiessen, C. and S. Thompson. "The Structure of Discourse and 'Subordination'." **Clause Combining in Grammar and Discourse**. Eds. J. Hainim and S. Thompson. John Benjamins Publishing Company. 1988. 275-329.

McKeown, K. R. **Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text**. Cambridge: Cambridge University Press, 1985.

Mellish, C., A. Knott, J. Oberlander and M. O'Donnell. [1998a] "Experiments using Stochastic Search for Text Planning." **Proceedings of the Ninth International Workshop on Natural Language Generation**, Niagara-on-the-Lake, Ontario, Canada, 1998.

- - -, M. O'Donnell, J. Oberlander and A. Knott. [1998b] "An Architecture for Opportunistic Text Generation." **Proceedings of the Ninth International Workshop on Natural Language Generation**, Niagara-on-the-Lake, Ontario, Canada, 1998.

Milosavljevic, Maria. "Augmenting the User's Knowledge via Comparison". **Proceedings of the 6th International Conference on User Modelling**. 2-5 June, 1997, Sardinia. 119-130.

- - -. **Maximising the Coherence of Descriptions via Comparison**. PhD Thesis, Macquarie University, Sydney, Australia 1999.

Oberlander, J., C. Mellish, M. O'Donnell and A. Knott. "Exploring a Gallery with Intelligent Labels." **Proceedings of the Fourth International Conference on Hypermedia and Interactivity in Museums**. Paris, September, 1997. 153-161.

- - -., M. O'Donnell, A. Knott and C. Mellish. "Conversation in the Museum: Experiments in Dynamic Hypermedia with the Intelligent Labelling Explorer." **New Review of Hypermedia and Multimedia**, 4 (1998): 11--32.

- - -, A. Knott, M. O'Donnell and C. Mellish. "Beyond Elaboration: Generating Descriptive Texts with Non-canonical Syntax." **Text Representation: Linguistic and Psycholinguistic Aspects**. Eds. T Sanders, J Schilperoord and W Spooren. Benjamins, in press.

O'Donnell, Michael. "Input Specification in the WAG Sentence Generation System." **Proceedings of the 8th International Workshop on Natural Language Generation**, Herstmonceux Castle, UK, 13-15 June, 1996.

- - -. "Variable-Length On-Line Document Generation". **Proceedings of the Flexible Hypertext Workshop of the Eighth ACM International Hypertext Conference**, Southhampton, UK, 1997.

- - -, M., H. Cheng and J. Hitzeman. "Integrating Referring and Informing in NP Planning." **Proceedings of the Coling-ACL '98 Workshop on the Computational Treatment of Nominals**, August 16, 1998, Universite de Montreal, Canada. 46-55.

Smithsonian Institution. "Smithsonian Gem & Mineral Collection." 30 March 1994. 19 January 2000. <http://galaxy.einet.net/images/gems/gems-icons.html>.