Stochastic text generation

Jon Oberlander and Chris Brew

Phil. Trans. R. Soc. Lond. A 2000 **358**, 1373-1387
doi: 10.1098/rsta.2000.0592

THE ROYAL SOCIETY

Stochastic text generation

By Jon Oberlander and Chris Brew

Human Communication Research Centre, Division of Informatics,
University of Edinburgh, Edinburgh EH8 9LW, UK

Natural language generation systems must achieve fluency goals, as well as fidelity goals. Fluency helps make systems more usable by, for instance, producing language that is easier for people to process, or which engenders a positive evaluation of the system. Using very simple examples, we have explored one way to achieve specific fluency goals. These goals are stated as norms on 'macroscopic' properties of the text as a whole, rather than on individual words or sentences. Such properties are hard to accommodate within a conventional architecture. One solution is a two-component architecture, which permits independent variation of the components, either or both of which can be stochastic.

Keywords: natural language generation; statistical methods; maximum-entropy modelling

1. Introduction: generation and understanding

Natural language generation (NLG) research aims at systems that produce coherent natural language text from an underlying representation of knowledge. Systems must produce language—single sentences or more complex discourses—which (i) faithfully represents the relevant knowledge, and also (ii) do this in a natural sounding way. These have been termed the fidelity and fluency goals, respectively (Ward 1993). The fluency goal leads to important differences between research in NLG and that in natural language understanding (NLU).

An NLU system has to recover meaning representations from input strings of text or speech. Whether or not a given string sounds natural, elegant, or forceful is immaterial. What matters is that an NLU system should be able to extract some meaning, and that the meaning should correspond as closely as possible to that intended by the string's speaker or writer.

At one level, NLG can be characterized as the inverse of this process. The system has to recover strings of text or speech from input meaning representations. It has therefore been argued that it should be possible to develop representations and processes that are reversible, and can thus be used for both NLU and NLG (Shieber 1988, 1993). However, the fluency goal introduces problems specific to NLG that are of relatively little significance in NLU. McDonald (1993) put it in the following way.

> Existing comprehension systems as a rule extract considerably less information from a text than a generator must appreciate in generating one. Examples include the reasons why a given word or syntactic construction is used rather than an alternative, what constitutes the style and rhetoric appropriate to a given genre and situation, or why information is clustered in one pattern of sentences rather than another.

In other words, there are many sets of surface strings that are equivalent, as far as an NLU system is concerned, but distinct as far as an NLG system is concerned. NLU systems could, in principle, distinguish elements in the equivalence sets. However, if they do, they will be reasoning under uncertainty, and this price does not seem worth paying, since identifying the style of a paragraph of text is of much less immediate use than identifying its propositional content.

It might be accepted that NLG systems must aim at naturalistic output, but argued that it is perfectly satisfactory if their fluency falls far short of human standards. For instance, telephone users trying to book airline tickets might prefer to have an efficient, 'unnatural' dialogue with something that is obviously a machine, rather than have to endure a polite, helpful and more 'natural' dialogue generated by a machine that was trying to pass the Turing test. If this is so, then the additional complications introduced by the fluency goal can be largely sidestepped by practical NLG systems.

However, the problems of fluency cannot be so easily ignored. In § 2, we discuss two kinds of fluency-related goal, and argue that achieving these goals is advantageous from a usability engineering perspective. In § 3, the focus is then on two particular textual properties related to these fluency goals: the distribution of sentence lengths, and vocabulary diversity. These properties apply to a text as a whole, rather than to individual words or sentences. As a result, § 4 presents a general system architecture that allows such properties to be independently varied.

2. Two kinds of desirable generation behaviour

People have expectations about the ways in which other people will talk. They also have personalities, which influence the ways in which they prefer to talk, and be talked to. In this section, we will argue that facts like these have significant implications for the design of practical generation systems.

(a) *Syntax and the maximization of expectedness*

Psycholinguists have noted that people often appear to use more words in their utterances than are strictly necessary. On the one hand, according to Grice's maxims of quantity and quality (Grice 1989), a speaker will attempt to optimize their utterance, making it as brief as possible, while still accurately distinguishing the intended meaning from any other candidates. Departures from the optimally efficient utterance will lead to their hearer deploying inferential effort, to calculate what further information the speaker meant to convey; the considerate speaker will, therefore, select the content and form of their utterance, so as to avoid suggesting such false or misleading inferences (Joshi 1982). For example, consider a situation containing three animals: one small white cat and two dogs, one large and black, and the other small and white. It is usually assumed that an optimal description of the first dog is either '*the large dog*' or '*the black dog*', whereas '*the large black dog*' will be suboptimal, since it contains two adjectives where one will do; it suffers from a degree of redundancy (Dale 1992; Reiter 1990).

On the other hand, there is substantial psycholinguistic evidence that the behaviour of human speakers involves the production of non-minimal utterances, and their hearers expect this behaviour. Thus, '*the large black dog*' may after all be the way the majority of speakers choose to describe the situation above (cf. Levelt (1989) for a survey). Conversely, hearers do not expect speakers to produce optimal, minimized utterances; in fact, such an unexpected utterance would provoke its hearer to search for reasons for its speaker's failure to use an expected utterance.

The expectation of non-minimality has implications for several areas of NLG. Let us consider two by way of example: the selection of referring expressions; and the aggregation together of sentences containing common elements.

First, Dale & Reiter (1995) discussed the former, focusing on the case of definite noun phrases (NPs). They propose that both people and NLG systems should strive to produce the most expected utterance, if they are to avoid unwanted implicatures. Elsewhere, we have argued that this notion of 'expected utterance' is not without its problems (Oberlander 1998). However, for current purposes, we can observe that one way of ensuring that an NLG system generates more expected utterances (and fewer unexpected ones) is to have it prefer to generate texts containing NPs with a distribution of lengths similar to that found in a relevant corpus of language use.

Secondly, the process of aggregation is often required in NLG systems that map a given set of propositions into a set of independent clauses. As Meteer (1992) has noted, such an approach avoids certain difficulties by ensuring a perfect match between what the system chooses to say, and the means available for saying it. However, if Sue met John yesterday, and also met Jane, such a simple system could express this as '*Sue met John. Sue met Jane.*', but not as '*Sue met John and Jane*'. Some of these systems therefore exploit aggregation to allow the derivation of the latter, under specific circumstances. In the current context, aggregation can be seen as a method for restoring naturalness to utterances: it converts a set of minimal, unexpected sentences into a set of less minimal, more expected sentences. In gross terms, it leads to fewer, longer sentences; more precisely, it leads to fewer sentences, more varied in length. By analogy with the NP case, it can, therefore, be seen that a further way of ensuring that an NLG system generates more expected utterances (and fewer unexpected ones) is to have it prefer to generate texts containing sentences with a distribution of lengths similar to that found in a relevant corpus of language use.

The difficulty for existing approaches to NLG is that sentence length is an emergent property of many low-level decisions. Once it is decided that sentence lengths should be distributed in a certain fashion, we have stipulated a global textual target, whose attainment does not follow from any individual lower-level decision.

(*b*) *Personality and the maximization of user satisfaction*

Researchers in personality psychology have investigated the extent to which people's visual appearance and non-verbal behaviour can create impressions in other people in considerable detail. Much of this work is based on the 'big-five' theory, which sees the most significant personality dimensions as extroversion (or dominance versus submissiveness), affection (warmth versus coldness), conscientiousness (competence versus incompetence), neuroticism (anxiousness versus relaxation), and openness to experience (liberalism versus conservativism) (Pervin & John 1996). In this paradigm, it has been found, for instance, that relative facial maturity creates impressions of competence and dominance (Berry 1991), and, more generally, that non-verbal information can allow external judges to give subjects scores on some of the dimensions that correlate well with the subjects' self-assessments, and those of their friends (Borkenau & Liebler 1992).

It has often been assumed that non-verbal information has a greater effect on impression formation than verbal information; partly as a result, verbal correlates of personality have been investigated somewhat less. However, it has recently been demonstrated that for perceived personality features such as competence and dominance, verbal behaviour has at least as strong an influence as non-verbal behaviour (Berry *et al*. 1997). In fact, a substantial amount of work has been carried out on language variables supposed to relate to gender (Lakoff 1975; Newcombe & Arnkoff 1979) or power; these include the use of tag questions, hedge expressions, and indirect speech acts. Although the results on some of those language variables have been mixed, it has consistently been found that certain simple measures of a speaker's vocabulary diversity correlate well with their perceived dominance and competence (Bradac 1990; Bradac *et al*. 1988). In particular, a speaker's type-to-token ratio (TTR) is directly related to their perceived competence (so long as the ratio is calculated in a way that controls for the length of their discourse).

The reason why personal style is an issue for NLG is that Moon & Nass (1996) have shown that a computer user prefers to work with a computer whose natural language messages have been designed to project personality parameters similar to the user's own. In particular, it was found that dominant-type users prefer computers using dominant-type language (here, the absence of hedge expressions); submissive users prefer computers using language like their own. As Reeves & Nass (1996) have emphasized, preference affects both subjective satisfaction, and the user's estimates of the computer's speed, efficiency and design.

The lesson for NLG then follows: it may be well worth the trouble of controlling output language so as to project a personality that matches the user's. Unfortunately, as with the maximization of expectedness, the projection of personality via TTR control merely establishes a target, without specifying any method for attaining it.

3. Fluency goals in text generation

Expectedness and personality issues are no doubt related to each other, and to other facets of fluency in natural language. However, for current purposes, there are two common factors. First, meeting the goals of achieving expectedness and projecting a personality are worthwhile engineering objectives, since they help avoid false implicatures—and, hence, reader effort—and they should improve user satisfaction. Secondly, they involve properties of the text as a whole. In this section, we investigate in further detail the concrete examples of sentence length and vocabulary diversity. Obviously, these do not reflect the full complexity of either expectedness or personality; however, they serve to effectively illuminate the broader issues.

(*a*) *Sentence length*

The first example is very simple: we stipulate that NLG systems should be able to control sentence length, producing sentences that are neither too short nor too long. This is a condition on the distribution of sentence lengths.

Sentence length is not always a useful criterion for discriminating between the work of different human authors. Mosteller & Wallace (1984) report a study by

Table 1. *Summary statistics for sentence length*

author	μ	σ^2	σ^2 (binomial)
Shakespeare	13.17	326.83	186.645
Twain	16.50	214.83	288.62
Lambs'	32.09	753.93	1061.53
Shakespeare (trigram)	13.23	286.56	188.32
Twain (trigram)	16.09	195.90	274.82
Lambs' (trigram)	32.07	901.337	1060.41
Shakespeare (bigram)	13.16	265.37	186.227
Twain (bigram)	16.32	209.47	295.49
Lambs' (bigram)	31.80	906.82	1043.14
Shakespeare (unigram)	12.85	174.68	178.09
Twain (unigram)	16.48	272.43	288.36
Lambs' (unigram)	30.50	955.98	960.50

Mosteller and Williams that conclusively demonstrates that it is not able to discriminate between the writings of Hamilton and of Madison in the *Federalist* papers. Nonetheless, some authors do differ from one another. To see how, we take Mark Twain's *Tom Sawyer*, Shakespeare's *Henry V* and Charles and Mary Lamb's *Tales from Shakespeare*, tokenize them, and measure sentence length.

The results are shown in the first three lines of table 1. The second column gives the mean sentence length, the third the empirical variance, and the fourth the variance that would be expected if the text had been produced by a binomial process (from which it would follow that the distribution of sentence lengths was geometric). Note that the means are very different from each other, and that Shakespeare's sentence length has higher variance than that of the corresponding binomial process, whereas the other authors use a narrower range of sentence lengths.

The next step, for current purposes, is to understand more about the properties of sentence length. It is known that text generated by sampling from simple n-gram models of text preserves some subjective impression of authorial style (Dewdney 1990). Sampling from a kth-order Markov model also preserves the expected frequency of all n-grams of order k or lower, but it is not obvious that this will be sufficient to capture sentence length. Table 1 lists means and variances for sentence length in 100 000-word samples of trigram, bigram and unigram approximations to the text produced by Twain, Shakespeare and the Lambs.† As the order of the Markov approximation decreases, the approximation to the sentence-length distribution becomes less exact. As a comparison, we quote the theoretically expected variance for a binomial distribution with the empirical mean. Both the bigram models and the trigram models capture regularities in the author's use of sentence length. Table 2 shows examples of the output.

† These are representative runs of the trigram, bigram and unigram simulations. The simulations were replicated several times with similar results. For the bigram and trigram simulations the implementation of the `shaney` generator from Kernighan & Pike (1999) was used. For the unigram distribution we used our own implementation.

Table 2. *Samples from models of Shakespeare's Henry V.*

source	text
Shakespeare	, armed in their stings , Make boot upon the summer ' s velvet buds , Which pillage they with merry march bring home To the tent – royal of their emperor ; Who , busied in his majesty , surveys The singing masons building roofs of gold , The civil citizens kneading up the honey , The poor mechanic porters crowding in Their heavy burdens at his narrow gate , The sad - ey ' d justice , with his surly hum , Delivering o ' er to executors pale The lazy yawning drone .
Trigram	HENRY . We are in God ' s peace ! I have an excellent armour ; but in loving me you should love the lovely bully . What men have defeated the law ; Who when they were as cold as any ' s ambition ! He was thinking of civil wars when he was a merry message . KING HENRY . Thou dost thy office fairly . Turn head and stop pursuit ; for we hear Your greeting is from him , you men of mould . Abate thy rage , abate thy manly rage ; Abate thy rage ,
Bigram	retreat . By my hand into the clouds , as cold as any military man in divers functions , Setting endeavour in continual motion ; To - morrow is Saint Crispian . He trots the air , a subject for a sovereign ' s ambition ! He was thinking of civil wars when he speaks , The King hath heard that men of mould . Abate thy rage . Use lenity , sweet chuck . NYM . They know your worthiness . My liege , as you shall read that my Nell is dead i ' faith , my cousin Suffolk
Unigram	great , , of . and nothing Who than ; , ; they gentleman ecus . that Till Britaine of Where Salisbury even about unprovided that sum Gainst . behind serve a it offend perdurable ; friends sort spirit whereof them English me mouth not Would thy put of peers civil ' pasture our READABLE the d , ? madame if that Isabel DAUPHIN need widow KING a shall ' like . wonderful he The Southampton ? the Consideration terre Hugh an snatchers is ' keep repose IS Exeunt ry , mothers inward was words are BOY another I , Europe

The trigram and bigram models sometimes have no choice but to produce verbatim copies of parts of the text on which they are based, as can be seen in the multiple occurrences of '*was thinking of civil wars when he*'. Nonetheless, some flexibility remains, and this fact will be exploited in due course.

Against this background, the key question is: how could a system achieve (or at least approach) a stipulated sentence-length distribution in generated text? Notice that the distribution will typically be that of a given target text, but nothing hinges on this.

Clearly, it is possible to impose Twain's sentence-length distribution on Shakespeare's text, by deleting all Shakespeare's sentence boundary markers, and running the result through a program that stochastically adds punctuation in the proportions used by Twain. But this brute-force approach is inappropriate, because nothing prevents sentence boundaries from being added in places where Shakespeare's text cannot support them.

A more realistic technique can be framed in the following terms. For the sake of argument, we will assume an architecture in which the output of a conventional NLG system is reviewed by a separate component responsible for sentence length. Call the first component the author and the second the reviewer. Various issues then arise as the author and the reviewer attempt to collaborate to produce a mutually acceptable version.

Firstly, sentence length is typically an emergent property of a large number of authorial decisions, few of which are based solely on stylistic considerations. The reviewer can indicate that a particular sentence is the wrong length, but it falls to the author to implement any change. The author's repertoire may not include a version of the sentence that changes the length while preserving propositional content and still meeting other authorial goals. In this case, the author may have an invidious choice to make. Even when an appropriate alternative version is in the author's repertoire, it may be a challenging task to find the parts of the authorial decision-making process that it would be most appropriate to modify. Without a principled means of doing this, the author is going to struggle to meet the reviewer's objections.

Secondly, the distribution of sentence length is itself an emergent property of a large number of decisions about the lengths of individual sentences. The reviewer may criticize the author's sentence-length profile without attributing blame to particular individual sentences. It now falls to the author to select and modify sentences—to aggregate them, in traditional NLG terms—so as to adjust the sentence-length distribution. In general, the difficulty that the author faces is that of reducing a target for a macroscopic property of the text to a prescription for change at the level of individual authorial decisions.

To take a physical analogy, the sentence-length distribution is like the temperature of a gas, while the length of an individual sentence is like the speed of a molecule within that gas. Just as knowledge of the temperature of a gas imposes little constraint on the speed of a particular molecule, so knowledge of the sentence-length distribution does not, on its own, determine the length of any individual sentence. The constraint applies to the ensemble of decisions made, not to any individual decision.

(*b*) *Vocabulary diversity*

The second example is slightly more complex than sentence length: we stipulate that NLG systems should be able to meet targets on vocabulary diversity. There are several ways of presenting this, as follows.

(1) As noted earlier, in the clinical, forensic and personality literature, the vocabulary diversity is often estimated using TTR. To avoid a dependency on the size of the text sample, TTR is measured not on the whole document, but on a series of fixed size bins.

(2) An allied measure (Yule 1944) is Yule's K, which for words w with frequency $|w|$, has the form:

$$K = 10\,000\frac{\sum |w|^2 - \sum |w|}{\left(\sum |w|\right)^2}.$$

Putting aside the constant factor of 10 000, this is the probability that two words drawn at random from the text will be identical. This will decrease as TTR increases.

(3) Another indicator of vocabulary diversity is the distribution of the distance between successive occurrences of the same word. One version of this calculates inter-token distance separately for each type in the vocabulary, while another sums over all types to produce a single figure. This too will decrease as TTR increases. Note that the distribution of sentence length is just the distribution of the distance between successive sentence boundary markers. Sentence length is, therefore, a special case of vocabulary diversity.

The measures listed above are sensitive to the frequency profile of words within the vocabulary, but it would make no difference if each English word were systematically replaced by a corresponding number, French word or Chinese character. So long as tokens can be checked for equality, the measures can be obtained. Given parallel word lists drawn from Twain and Shakespeare, it is possible to impose Twain's vocabulary choice on Shakespeare by replacing the nth most frequent word in Shakespeare's vocabulary with the nth most frequent word in Twain's. If punctuation is passed through unchanged, we will also have Shakespeare's sentence-length distribution. But the result would be gibberish, failing for the same reason as the brute-force attempt to impose sentence-length distribution: inadequate account has been taken of context.

4. An architecture for stochastic text generation

In the following sections we will display a general methodology for producing NLG systems that achieve (or at least approach) goals for macroscopic properties of text. We will do this by introducing an unconventional NLG architecture (Knight & Hatzivassiloglou 1995; Langkilde & Knight 1998), which we modify to meet our needs. Langkilde & Knight's (1998) Nitrogen uses a probabilistic model to select among analyses proposed by a non-deterministic generator. The system has the following architecture.

(i) A symbolic generator capable of generating alternative answers.

(ii) A word lattice produced by the symbolic generator.

(iii) A statistical extractor capable of unpacking and evaluating alternative paths through the word lattice.

It is worth noting the goals that Nitrogen and its successors have been set, since these differ significantly from the mainstream goals of NLG. The key goal is to produce output irrespective of the poverty of the input to the generation process. There is no guarantee that the output will be correct, although the aim is to make it as fluent as possible. The generator is non-deterministic because its input comes from machine translation and is too impoverished to completely determine the output. (It may, for example, lack information about number and case.) The language model can be arbitrarily sophisticated, but, to date, the reported experiments use simple n-gram models.

The computational problem faced by the language model is well studied, because it arises when the input to a parser is the output of a speech recognizer. The only difference is the source of the uncertainty, which needs to be resolved. The speech recognizer's language model is attempting to reconstruct an utterance from a lattice of perceptual data, whereas Nitrogen's language model is attempting to find an appropriate text from a word lattice produced by the underlying generator.

Nitrogen's non-deterministic generator is purely possibilistic, and, in current terms, plays the role of the author, while a language model plays the role of the reviewer. Significantly, however, instead of passing a single text to the reviewer, the author passes a lattice representing a very large space of possibilities. Instead of sending back requests for revision, the reviewer can simply choose the best of the available alternatives, secure in the knowledge that all of these are at least marginally acceptable to the author. The preferences of the reviewer still predominate. Strong authorial preferences will still be overridden, even on the basis of very small differences in the reviewer's language model. But this property is not essential, since nothing hinges on the fact that the author is purely possibilistic. For our purposes, what matters is Nitrogen's clear separation between the roles of the author and the reviewer.

(*a*) *An architecture for imposing a sentence-length distribution*

Consider a variation of the Nitrogen framework in which both author and reviewer are modelled by stochastic processes. The author is a trigram model built from text by Shakespeare, while the reviewer is a statistical model of the sentence-length distribution occurring in the same text. For the latter component, one could use: a binomial; Katz's K mixture (Katz 1996); a negative binomial (Church & Gale 1995); or any other convenient distribution. The lattice produced by the author contains, as before, only paths that are at least minimally acceptable to the author, but these paths are now annotated with weights derived from the trigram model. Quasi-Shakespearean text with a quasi-Shakespearean sentence-length distribution can be generated by allowing the reviewer to choose a high-scoring path through the lattice of alternatives provided by the author. Every path will be drawn from Shakespeare's trigram model, but the choice is up to the reviewer.

We can vary author and reviewer independently. So we could impose Twain's sentence-length preferences on Shakespeare's trigram source, or vice versa. This technique is gentler than the naive approaches presented earlier, because the output will contain only trigrams represented in the original source. Conversely, unless the trigram lattice contains a sufficient choice of paths, it may not be possible to match the reviewer's sentence-length norms as closely as would be possible with the more brutal methods.

Standard techniques for finding paths through weighted lattices—as detailed, for example, in Jelinek's (1997) textbook—are applicable to the reviewer's path-decoding task. But the reviewer must decide the logically prior question of how to combine the weights provided by the author with the preferences arising from the sentence-length model. An obvious answer is to use a linear combination between the weights of the two models,

$$P(s_1 \to s_2) = \lambda P_{\text{author}}(s_1 \to s_2) + (1-\lambda) P_{\text{reviewer}}(s_1 \to s_2),$$

where λ is an adjustable parameter. Extreme values will allow either of the two component distributions to be used to the exclusion of the other.

i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	...
window	onto	a	text	.	This	is	the	window	which	is	...
1	2	3	4	5	6	7	8	1	9	7	...

Figure 1. Labelling items in a bin.

(*b*) *An architecture for imposing vocabulary diversity*

The general architecture developed for imposing sentence length applies to the more complex task of imposing a pattern of vocabulary diversity. The original formulation of the author as a trigram source stays in place, but the shift of focus from sentence length to vocabulary diversity entails that the simple distribution used by the reviewer must be replaced by something more elaborate. Because the sentence-length distribution is a special case of the distribution of inter-token distance, the more elaborate techniques can also be used for the simpler problem. This is appropriate, because the focus on sentence length to the exclusion of other factors was an idealization. In the real world, multiple authorial goals compete to be satisfied, and the techniques developed in this section are designed to deal with this situation.

We begin by making precise the aim of this section. The aim is to simulate the distribution of inter-token distance produced by a particular author. This can be done by first defining a set of features that are sufficient to capture the salient features of distribution, and then incorporating these features into maximum-entropy models. Standard maximum-entropy techniques for carrying out feature selection and weight estimation (Berger *et al*. 1996; Della Pietra *et al*. 1997; Rosenfeld, this issue) are assumed. The merit of these techniques is exactly that they are standard, requiring from the user only the definition of an appropriate space of possible features.

The TTR is calculated over 25-word bins. The TTR will be less than 1 only when the bin contains repeated words. Starting from the left of the bin, label words as they appear, giving each word type a distinct label. This process is illustrated in figure 1. The representation in the third line is sufficient to calculate all the measures of vocabulary diversity, but does not mention individual words, so will abstract away from the vocabulary choices of a particular author. A set of features can be defined over that representation. As is usual with maximum-entropy models, the intention is not to preselect appropriate features, but to define a large set of features, sufficient to cover the intended regularities, and then to delegate feature selection and model building to standard algorithms.

Suitable features are defined by the following predicate templates.

(i) The label I appears at position p.

(ii) Positions p and q carry the same label.

(iii) Positions p and q are both labelled with I.

(iv) The label I is repeated, with an inter-token distance of d.

The motivation for these features is to find attributes that can be measured for one author but applied to the output of another. Both the features and the representation underlying them are open to revision. In particular, it might be better not to relabel punctuation symbols or common closed-class words, which play a role different from those of content words. Since every author uses these symbols, it will still be possible to apply the results of training on one author to the evaluation of another.

A maximum-entropy model of Shakespeare's vocabulary diversity can be created by grouping the text into 25-word bins, carrying out the relabelling process, and counting relevant events as they occur within the bins. Similar models for Twain, the Lambs, or groups of writers can be created. Given the usual precautions against over-fitting, the maximum-entropy algorithms will build each model based on a limited set of important features. This model can then play the part of the reviewer in the architecture that was outlined in the previous section. It is therefore possible to impose one author's vocabulary profile on another's text, in a principled fashion.

(*c*) *Further macroscopic properties*

Maximum-entropy modelling is a powerful general technique. It has the crucial advantage of being able to handle situations in which the features used are not independent. Because TTR and the other measures of vocabulary diversity are properties of the configuration of elements within the bin, and cannot be simply ascribed to the effect of any individual choice by a binomial or multinomial process, the maximum-entropy approach is warranted by our application. But the key point is that in this respect the TTR measure is representative of a large class of macroscopic properties of text for which we may wish NLG systems to respect specified norms.

First consider those properties relevant to the achievement of 'expectedness' goals. To make a text easy to process and free of misleading structures, it is certainly not sufficient simply to meet a target for sentence length. In the earlier discussion of expectedness, at least one other target was introduced—the distribution of noun phrase lengths—and it is easy to see how a treatment of this feature might follow that for sentence lengths. However, there are many other factors that are associated with readers' expectations, and, very often, they flow from the type of interaction between author (or speaker) and reader (or hearer). Biber (1986) proposes a number of linguistic factors that are associated with significant differences between one genre of language and another. His 'abstract versus situated content' dimension, for instance, opposes language containing more nominalizations, prepositions, agentless passives or '*it*'-clefts with that containing more place and time adverbs, relative pronoun deletion or subordinator-'*that*' deletion. Indeed, such genres create expectations in readers, and often these expectations concern a particular group style. There has been computational work on the achievement of stylistic goals in text (cf. Danlos 1987; Hovy 1988), and this has isolated a number of syntactic patterns that contribute to stylistic goals (DiMarco & Hirst 1993). The patterns are described in terms of balance, dominance and position, and the goals in terms of clarity versus obscurity, concreteness versus abstraction, and staticness versus dynamism. DiMarco & Hirst (1993) supply a stylistic classification of primitive elements (such as adjectivals and adverbials), which, given a set of grammar rules, have effects contributing to the higher-level patterns and goals. This work therefore represents a plausible source of features for training a maximum-entropy model.

Secondly, turning to personality issues, it is again clear that TTRs represent only the tip of a linguistic iceberg. Unlike style, there has been little computational work on creating favourable personality impressions, and it is true that psychologists have devoted more attention to non-verbal factors influencing impression formation. However, enough has been discovered about relevant linguistic factors to begin to explore their incorporation in a maximum-entropy model. Taking just the work of Berry *et al*. (1997), further factors include the frequencies of adjectives denoting negative and positive emotions; propositional attitude verbs; self-referents; negations; and present versus past tense. In addition, it is generally accepted that overall word count is associated with perceived dominance, as is the avoidance of tag questions and hedge expressions.

5. Conclusion

We have argued that NLG systems need to be able to achieve fluency goals, as well as fidelity goals. Undoubtedly, it is important that they faithfully represent relevant knowledge. However, from the point of view of usability engineering, it is also important that they produce language that is easy for people to process, and which engenders a positive evaluation of the system itself. Using very simple examples, we have explored one way of achieving fluency goals. Because the goals are stated as norms on macroscopic properties of the text, the system architecture must allow such norms to be stated, and that a clear mechanism must be provided whereby macroscopic properties can emerge from ensembles of microscopic decisions. We have suggested a revision of the Nitrogen architecture as a model of what is needed in such systems. Because our examples have been highly idealized, there is an open question about the feasibility of our approach as a means for generating useful text.

We have used a Markov trigram generator as a stand-in for the author, and emphasized the fact that a second supervisory component takes the responsibility for selecting between its outputs, and for deciding to what extent its expressed preferences will be respected. This is not to insist that the Markov generator is appropriate technology. However, it does appear desirable to continue to use a separate supervisory component, together with a non-deterministic generator, albeit one that is more elaborate than the Markov generator. Like any conventional NLG system, this non-deterministic generator must have an authorial repertoire wide enough to cover the ideas that it needs to express and the situations in which it has to express them. But, crucially, it does not have to fully understand the conditions under which the use of particular options and combinations of options are appropriate, since that task is delegated to the reviewer.

One dimension of variation worth exploring in future work is the complexity of the communications medium shared by author and reviewer. The proposed use of a weighted word lattice will probably be too limiting. On the one hand, were we to move to an unconstrained blackboard architecture, the benefits of the existing modularization of the system would be at risk. On the other hand, in unpublished work, Langkilde has already proposed an extended Nitrogen architecture in which the two components share access to a probabilistically annotated parse forest. This allows the authorial component to provide the reviewer with more information about the structure of the space of possibilities that it has considered.

Finally, by incorporating stylistic features already proposed within the NLG literature, stochastic systems offer a novel approach to the task of generating genuinely fluent language.

The research reported here was carried out at the Human Communication Research Centre (HCRC) at the University of Edinburgh. The first author was, until recently, an Engineering and Physical Sciences Research Council (EPSRC) Advanced Fellow, whose work on the particular topic reported here was made possible by a Macquarie University Visiting Research Scholarship. Therefore, we gratefully acknowledge the support of the Economic and Social Research Council for the HCRC's research programme, as well as the generosity of both the EPSRC and Macquarie University, Sydney. Special thanks to Gerald Gazdar and Karen Spärck Jones for their extremely helpful comments, and to Sebastian Varges for insightful discussions.

References

Berger, A. L., Della Pietra, S. & Della Pietra, V. 1996 A maximum entropy approach to natural language processing. *Comp. Ling.* **22**, 39–71.

Berry, D. S. 1991 Attractive faces are not all created equal: joint effects of facial babyishness and attractiveness on social perception. *Personality Social Psych. Bull.* **17**, 523–531.

Berry, D. S., Pennebaker, J. W., Mueller, J. S. & Hiller, W. S. 1997 Linguistic bases of social perception. *Personality Social Psych. Bull.* **23**, 526–537.

Biber, D. 1986 Spoken and written textual dimensions in English: resolving the contradictory findings. *Language* **62**, 384–411.

Borkenau, P. & Liebler, A. 1992 Trait inferences: sources of validity at zero acquaintance. *J. Personality Social Psych.* **62**, 645–657.

Bradac, J. J. 1990 Language attitudes and impression formation. In *Handbook of language and social psychology* (ed. H. Giles & W. P. Robinson), pp. 387–412. Wiley.

Bradac, J. J., Mulac, A. & House, A. 1988 Lexical diversity and magnitude of convergent versus divergent style shifting: perceptual and evaluative consequences. *Lang. Commun.* **8**, 213–228.

Church, K. W. & Gale, W. A. 1995 Poisson mixtures. *Natural Lang. Engng* **1**, 163–190.

Dale, R. 1992 *Generating referring expressions: building descriptions in a domain of objects and processes*. Cambridge, MA: MIT Press.

Dale, R. & Reiter, E. 1995 Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Sci.* **19**, 233–263.

Danlos, L. 1987 *The linguistic basis of text generation*. Cambridge University Press.

Della Pietra, S., Della Pietra, V. & Lafferty, J. 1997 Inducing features of random fields. *IEEE Trans. Pattern Analysis Machine Intell.* **19**, 1–13.

Dewdney, A. K. 1990 *The magic machine*. San Francisco, CA: Freeman.

DiMarco, C. & Hirst, G. 1993 A computational theory of goal-directed syntax. *Comp. Ling.* **19**, 451–499.

Grice, H. P. 1989 *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Hovy, E. 1988 *Generating natural language under pragmatic constraints*. Hillsdale, NJ: Lawrence Erlbaum.

Jelinek, F. 1997 *Statistical methods for speech recognition*. Cambridge, MA: MIT Press.

Joshi, A. K. 1982 Mutual beliefs in question answering systems. In *Mutual knowledge* (ed. N. V. Smith), pp. 181–197. Academic.

Katz, S. M. 1996 Distribution of content words and phrases in text and language modelling. *Natural Lang. Engng* **2**, 15–59.

Kernighan, B. W. & Pike, R. 1999 *The practice of programming*. Reading, MA: Addison-Wesley.

Knight, K. & Hatzivassiloglou, V. 1995. Two-level, many-paths generation. In *Proc. 33rd A. Mtg of the Association for Computational Linguistics, Cambridge, MA, June 1995*, pp. 252–260.

Lakoff, R. 1975 *Language and women's place*. New York: Harper Row.

Langkilde, I. & Knight, K. 1998 Generation that exploits corpus-based statistical knowledge. In *Proc. 36th A. Mtg of the Association for Computational Linguistics and 17th Int. Conf. on Computational Linguistics, Montreal, Canada, August 1998*, pp. 704–710.

Levelt, W. 1989 *Speaking: from intention to articulation.* Cambridge, MA: MIT Press.

McDonald, D. D. 1993 Issues in the choice of a source for natural language generation. *Comp. Ling.* **19**, 191–197.

Meteer, M. 1992 *Expressibility and the problem of efficient text planning.* London: Pinter.

Moon, Y. & Nass, C. 1996 How 'real' are computer personalities? Psychological responses to personality types in human–computer interaction. *Commun. Res.* **23**, 651–674.

Mosteller, F. & Wallace, D. L. 1984 *Applied Bayesian and classical inference: the case of the Federalist Papers.* Springer.

Newcombe, N. & Arnkoff, D. B. 1979 Effects of speech style and sex of speaker on person perception. *J. Personality Social Psych.* **37**, 1293–1303.

Oberlander, J. 1998 Do the right thing... but expect the unexpected. *Comp. Ling.* **24**, 501–507.

Pervin, L. A. & John, O. P. 1996 *Personality: theory and research*, 7th edn. Wiley.

Reeves, B. & Nass, C. 1996 The media equation: how people treat computers, television, and new media like real people and places. Stanford, CA: CSLI.

Reiter, E. 1990 The computational complexity of avoiding unwanted computational implicatures. In *Proc. 28th A. Mtg of the Association for Computational Linguistics, Pittsburgh, PA, June 1990*, pp. 97–104.

Shieber, S. M. 1988 A uniform architecture for parsing and generation. In *Proc. 12th Int. Conf. on Computational Linguistics, Hungary, August 1988*, pp. 614–619.

Shieber, S. M. 1993 The problem of logical-form equivalence. *Comp. Ling.* **19**, 179–190.

Ward, N. 1993 *A connectionist language generator.* Norwood, NJ: Ablex.

Yule, G. U. 1944 *The statistical study of literary vocabulary.* Cambridge University Press.

Discussion

N. Nicolov (*University of Sussex, Brighton, UK*). I like the separation of 'author' and 'reviewer' functions in your model. But do you not run into problems of requiring an exponentially large volume of text to be generated by the author for review? Could the reviewing function perhaps be integrated into the generator, so as to reduce the volume of generated text?

J. Oberlander. Running the generator iteratively to generate all possible texts is certainly not feasible. This is why we have proposed word lattices as an alternative, though this still means a lot of work. However, I would resist any proposal to reintegrate the reviewer with the generator because we like the idea of a language-independent reviewer. We want to explore what the separation of these two functions will bring us.

R. Rosenfeld (*Carnegie Mellon University, Pittsburgh, PA, USA*). I too liked your separation of author and reviewer functions, and would like to see how far you can push it. Your generator is a trigram model, have you considered how you can produce meaningful text and edit it further?

J. Oberlander. One possibility would be to use a template-based generator and smooth its output. Perhaps we could use the Nitrogen model and add our stochastic evaluator onto the end of that to ensure that the macroscopic text properties we want are preserved.

K. R. McKeown (*Columbia University, New York, USA*). My question is similar to Roni Rosenfeld's. Your work is the first to introduce a stochastic element into natural language generation. But it looks like a swing from the purely symbolic to the purely statistical. Langkilde's work, which you cite, tries to push grammar to a minimum, but at the cost of ignoring the large amount of good work on producing large-scale grammars for generation. How can you connect more with symbolic work in order to get a better balance?

J. Oberlander. We agree that the work on large-scale grammars, which serve fidelity goals, should not be ignored, and while we wish to push the stochastic approach as far as it will go, the framework here is designed to allow us to build directly on the existing symbolic work, including, particularly, the work to date on stylistic features of discourse. The current approach was inspired, in part, by an observation of John Bateman's concerning a comment made by one of the museum curators about a particular piece of jewellery. The curator's discourse was very fluent, and had apparently included extra content purely to satisfy fluency goals. So, we are here focusing on the cases where fluency can be just as important as fidelity. But, ultimately, getting even these cases right will still involve the exploitation of linguistic features uncovered in traditional symbolic approaches.

K. I. B. Spärck Jones (*University of Cambridge, UK*). Surely this distinction between content and form is too simplistic? After all, even reordering changes content by changing emphasis.

J. Oberlander. Indeed, the distinction is not absolute, and nor is the division of responsibility between author and reviewer. The form of the final text is not simply the work of the reviewer. The set of generated texts available for review depends on the sophistication of the lattice from which they are produced. Equally, the content of the final text is not simply the work of the author. So, to strategically influence the outcome, the clever writer must generate multiple versions which will survive the cuts the reviewer makes. This is somewhat akin to the journalistic style of iterative deepening, whereby a news story is written by revisiting the same topic in progressively more detail so that it may later be cut to virtually any length by an editor.

F. Pereira (*AT&T Laboratories, Florham Park, NJ, USA*). You said that you do not want to push the critique into the generator. It is important to distinguish what is computed (an evaluation function over generated texts) from how it is computed (by separate modules doing text generation and evaluation). The virtue of your architecture is that it keeps this distinction very clear, but it need not be this way. An integration at the implementation level might yield computational efficiencies without sacrificing the functional model.

J. Oberlander. The short answer is 'yes'! A longer answer, on reflection, is that the two modules can indeed be combined in more than one way. The obvious combination keeps the distinction intact, but more efficient combinations may require much more complex composition. We would like to try the former first.