

Prodigy: Probabilistic Deep Generation

Case for Support

Motivation

*Computational methods for generating language are lagging behind computational methods for analysing language in several ways, most obviously in that they have barely been used commercially. The main reasons for this are that systems for generating language take inordinate amounts of time to build, yet once built cannot be reused, and tend to be severely lacking in language variation which is easily perceived as lacking in quality. The current situation in language generation research is reminiscent of language analysis research in the late 1980s, when symbolic and statistical methods briefly formed entirely separate research paradigms. Language analysis soon moved towards a paradigm merger, realising that symbolic methods lacked the efficiency and robustness that probabilistic methods could provide, which in turn would benefit from the accuracy and subtlety of symbolic methods. A similar development is currently underway in the field of machine translation where — after several years of purely statistical methods dominating the field — researchers are now beginning to bring linguistic knowledge back in. The experience from these research fields suggests that higher quality can be achieved when the symbolic and statistical paradigms join forces. Recent research shows that this is likely to be true for language generation too. The aim of the **Prodigy** project is to develop, for the first time, a comprehensive, linguistically informed, probabilistic methodology for generating language that substantially improves development time, reusability and variation in language generation systems, and thereby enhances their commercial viability.*

Part I: Project team

Host institution

The University of Brighton has a track record of high-quality research in natural language technology (NLT) stretching back over fourteen years, initially in the Information Technology Research Institute (ITRI) and, since a reorganisation in 2005, in the Natural Language Technology Group (NLTG) in the School of Computing, Mathematical and Information Sciences (CMIS). The group's research has been recognised for many years as internationally leading¹, and was the principal contributor to the University's 2001 RAE grade 4 in Computer Science.

Funded by UK research councils (EPSRC, ESRC, MRC), the European Commission (Network of Excellence, IST, Esprit, INCO Copernicus, INTAS, eContent, LRE) and industry, the group's work has primarily centered on Natural Language Generation (NLG) and lexical representation, with associated interests in human computer interfaces and natural-language system architectures. Members of NLTG have participated in a range of externally funded research projects with NLG at their core, including RAGS, PILLS, NECA, CLIME, HALO, EPOCH and COGENT². This work has resulted in a range of significant research outputs, including WYSIWYM (an NLG-based technique for document authoring, question formulation and complex data entry), the RAGS generic architecture for NLG systems, and the word-sketch approach to computational lexicography (used in a recent edition of the Macmillan English Dictionary).

Brighton is one of the leading universities for research in the post-1992 sector, and CMIS has a strong research culture in computer science and related areas. CMIS staff have led a range of research projects funded by the UK research councils (AHRC, EPSRC, ESRC), the Department of Health, and the European Commission (Network of excellence, EU Leonardo, EU India), and the school provides a solid foundation of experience in administrative support for research.

Principal investigator's track record and plans for future research

Anja Belz joined ITRI in 2001 as a research fellow, and briefly worked on two EC-funded NLG-related projects — PILLS (Patient Information Leaflet Localisation System) and NECA (Net Environment for Embodied Conversational Agents) — before becoming co-author and named research fellow (later promoted to senior research fellow) on COGENT (Controlled Generation of Text, 2003-2006, EPSRC grant GR/S24480/01). At the beginning of 2006, she was appointed to the position of senior lecturer in CMIS, where her responsibilities include being module leader for the level 3 module Topics in Language Technology.

Belz's research has always centered around two main strands: (i) formal models for natural language and (ii) computational learning, with a particular focus on statistical modeling. Following a PhD³ and first post-doctoral position⁴ in natural language understanding (NLU), Belz has over the last few years increasingly worked in language generation,

¹Schneider and Rodd, 2001: International Review of UK Research in Computer Science. Report commissioned by EPSRC, BCS and IEEE.

²See <http://www.itri.brighton.ac.uk/projectsindex.html>.

³*Learning Finite-state Models for Natural Language Processing*, with Gerald Gazdar, at University of Sussex.

⁴Learning Computational Grammars, EU funded, under TMR scheme, large international research and training network, coordinated by John Nerbonne at University of Groningen.

and has now established herself internationally as a researcher in NLG. In 2004, she co-chaired (with ITRI colleagues) the Third International Natural Language Generation Conference (INLG'04). Last year, she co-chaired (with Sebastian Varges, Stanford University) the successful new Workshop on Using Corpora for NLG, and was invited to speak on evaluation in NLG at the European Language Resources Association's 10th anniversary workshop on NLT evaluation. Currently, she is co-chairing (with Robert Dale, Macquarie University) the Special Session on Sharing Data and Comparative Evaluation at INLG'06. She has reviewed submissions for funding bodies and journals, and is serving on four programme committees this year.

Over the past three years, Belz's research has focused in particular on comparative evaluation and probabilistic methods for NLG, and these two topics will be the cornerstones of her future research:

Comparative Evaluation for NLG. There is very little existing work on evaluation of alternative techniques applied to the same task (or comparative evaluation) in NLG, and at this moment we have little idea of which generation techniques or general approaches work better than others.

In collaboration with Ehud Reiter (University of Aberdeen), Belz has been investigating methods for comparatively evaluating NLG, and has carried out experiments to assess automatic evaluation metrics in terms of their ability to predict human judgments of NLG system outputs [8]. This research was only the second time in NLG that directly comparable results were produced for different techniques.

Belz and Reiter are continuing this line of investigation in the GENEVAL initiative [20, 7], the aim of which is to establish — in dialogue with the research community — a methodology for comparative evaluation in NLG. This will involve (i) identifying and creating data resources to be shared, (ii) developing human and automatic evaluation techniques that are reliable and trusted, and (iii) organising evaluation campaigns and associated workshops as forums for dialogue with the community and testbeds for different approaches to evaluation. New evaluation metrics and new insights into comparative evaluation in GENEVAL will directly benefit the benchmarking and general evaluation efforts to be undertaken in the **Prodigy** project.

Probabilistic NLG. Over the last decade, there has been a lot of interest in statistical techniques among researchers in NLG, a field that was largely unaffected by the statistical revolution in NLU that started in the 1980s. The potential advantages of statistical NLG methods are improved robustness, reusability and system development time, and an adaptable handle on decision-making. However, there has been a distinct lack of take-up of statistical techniques in NLG applications, for two main reasons: existing statistical NLG techniques (i) are inherently expensive, requiring the set of alternatives to be generated in full before the statistical model is applied to select the most likely; and (ii) have simply not been shown to produce outputs of high enough quality.

In the EPSRC-funded COGENT project, Belz developed the probabilistic *p*CRU language generation framework, as a more efficient, linguistically more informed, and statistically principled alternative to existing probabilistic NLG techniques. Tests have shown [5, 8] the *p*CRU approach to result in substantial improvements in development time and reusability, while the generated texts were judged better by human evaluators than human-written texts [8]. *p*CRU was the proof of concept that probabilistic NLG can be efficient and high-quality as well as robust and reusable, but left many issues unexplored. The **Prodigy** project aims to fully explore whether the combination of the probabilistic and the linguistic can be as beneficial for NLG as it has been for NLU.

Affiliated project members

Ehud Reiter, University of Aberdeen, UK: Ehud Reiter is a Senior Lecturer in Computing Science at the University of Aberdeen. He is a leading researcher in Natural Language Generation, with over 60 publications in NLG, including the first book ever written about building NLG systems; he also was area chair for NLG at EACL'06. His research includes the EPSRC-funded SUMTIME and BabyTalk projects mentioned in this proposal. He has previously collaborated with the investigator on research on evaluation of NLG systems [8, 20]. His role in the project is to advise on domain-independent input representation strategies (see Proposed Research, Section 2.3) and evaluation methods (Section 2.5).

Khalid Choukri, ELRA/ELDA, France: Khalid Choukri is executive director of the European Language Resources Agency (ELRA) and chairman and CEO of its operational body, the Evaluations and Language resources Distribution Agency (ELDA). The primary purpose of ELRA/ELDA is to create and distribute language resources and to evaluate human language technologies. ELRA/ELDA will support research efforts in Prodigy through advice on evaluation strategies, corpus creation and resource distribution (see also letter of support regarding proposed level of contribution).

Part II: Proposed research

1 Introduction and research context

Natural language generation (NLG) is the branch of language processing that maps non-language representations of information to language that expresses the information. NLG is a subtask in some applications, e.g. machine translation where it enables computational systems to compose a translation in one language from an analysis of a text in another language, and in human-computer interaction where it enables the system to formulate replies, instructions, explanations etc. NLG is also a task in its own right, for example when creating summaries, reports and descriptions from databases and other sources of nonverbal information. This stand-alone, data-to-text type of NLG is the focus of the **Prodigy** project.

The number of potential applications of data-to-text NLG technology is vast: most companies and organisations regularly turn some form of nonverbal information (e.g. account information, performance results) into letters, reports, manuals, etc. NLG technology can help economise the text-production process, but it can also make information available in verbal form that would otherwise be inaccessible or more time-consuming to process. NLG researchers have looked at a wide range of applications⁵:

1. *Economising text production*: serial letters for business and health providers; descriptions of museum exhibits, stock market movements, route directions; instructional texts such as cooking recipes, technical manuals, patient information leaflets; prognostic reports such as weather forecasts and pollen forecasts;
2. *Making nonverbal information more available*: (a) increasing accessibility to non-experts — explanations of proofs and complex phenomena, presenting information in expert systems; (b) speeding up the rate at which information can be processed — diagnostic reports such as medical reports, fault reports, error messages, air pollution reports; (c) providing a modality that might otherwise be wasted — e.g. where visual modalities are already busy, additional, nonverbal information can be converted into verbal form and provided over an audio channel.

Yet, despite its potential usefulness, NLG is not actually *used*: at present, there are virtually no commercial NLG systems. This lack of commercial take-up has four main reasons, all to some extent interrelated: the typical NLG system is (i) expensive and time-consuming to build, (ii) not reusable, (iii) restricted in its language, and (iv) not robust. NLG systems tend to be handcrafted as rule-based, deterministic decision-makers with hardwired conditions for rule application. Such systems are usable only in one specific application context (every application requires different contents and a different style), with the result that on the whole, no NLG system or any of its components can be reused. Handcrafting NLG systems is moreover an expensive and time-consuming process (building single applied systems tends to take up entire 3-year research projects), and language needs to be severely restricted in order to make the system-building task manageable. Finally, NLG systems tend to be reliant on fully specified inputs, lacking the robustness to deal with new, partially or improperly specified inputs.

In other fields of natural language processing (NLP) such as parsing, probabilistic methods have provided the answer to similar issues. In the early 1990s, NLU moved towards a merger between the initially separate statistical and symbolic paradigms, realising that symbolic NLP lacked the efficiency and robustness that probabilistic NLP could provide, which in turn would benefit from the accuracy and subtlety of symbolic NLP [13, p. 98]. The next generation of high-accuracy parsers, for instance, were based on approaches resulting from this paradigm merger.

Statistical techniques have begun to appear in NLG. A host of techniques have been applied to relatively small sub-problems of NLG, mostly in the last stage of generation known as surface realisation (e.g. NP type determination [19], and lexical choice [2]), but also in earlier, ‘deeper’ stages of generation (e.g. content selection [3], attribute selection in content planning [18], text planning [16]). Perhaps the best known statistical approach to NLG is generate-and-select NLG where all alternative realisations are generated in full and a statistical selection mechanism is then applied to select the most likely [15]. Efficiency is an issue with generate-and-select, because the number of alternative realisations can be vast, e.g. during surface realisation alone, trillions of alternatives may have to be generated [15]. Furthermore, statistical methods have not been shown to generate texts of high enough quality for real-world use. This is largely because n-gram models — used in much of statistical NLG — while doing useful work in speech recognition and tagging, are not sufficient to model the longer-range dependencies and structural regularities of tasks such as language parsing and generation.

In order to become viable for real-world applications (and ultimately, for commercial deployment), NLG needs to enhance its technologies in terms of **reusability**, **robustness**, and **development time**, at the same time ensuring **computational efficiency** and **high-quality outputs**. Probabilistic methods *are* key to achieving this, but in order to produce high quality outputs, probabilistic NLG needs to be *linguistically informed*, associating probabilities with linguistic objects, not the arbitrary same-length strings of words that n-grams are based on; in order to be feasible in terms of computational expense, probabilities need to *inform choice during the generation process*, not selection after it; and in order to improve significantly on development time, use of adaptable, probabilistic methods needs to be extended backwards into *deep generation*, and not be restricted to surface realisation with its relatively straightforward tasks of lexical and syntactic choice.

⁵Implemented systems exemplifying all of these and more can be found on John Bateman and Michael Zock’s NLG website: <http://www.fb10.uni-bremen.de/anglistik/langpro/NLG-table/>.

Taking the principal investigator’s previous EPSRC-funded research on probabilistic NLG (see Section 2.1) as a starting point, the **Prodigy** project will fully explore whether the combination of the probabilistic and the linguistic can be as beneficial for NLG as it has been for NLU. The aim is to achieve advances in output quality, efficiency, reusability and robustness by focusing on developing technology in two key areas:

- *domain-independent linguistic representation* — representation in deep and surface generation needs to be linguistically grounded in order to achieve high-quality language, and domain-independent in order to enhance reusability;
- *probabilistic decision-making* — decision-making based on probabilistic models enhances the robustness and reusability of systems, and probabilistic models can be the basis for highly efficient decision-making.

The principal outcomes of the **Prodigy** project will be:

- advances in our understanding of reusable technology for NLG;
- empirical results regarding requirements for domain-independent input representations for NLG;
- a methodology for probabilistic language generation incorporating deep as well as surface generation;
- an implemented suite of tools enabling researchers to build probabilistic generators; and
- new data resources of paired data-to-text NLG inputs and outputs.

2 Methodology

The first subtask in data-to-text NLG is to determine on the basis of the input data what information is going to be represented in the textual output. This task — often called content determination (CD) — involves different kinds of data processing and summarisation for different types of input data, and is necessarily domain-specific. While it is commonly assumed [21] that CD is part and parcel of NLG, Evans et al. [12] have argued that the non-linguistic part of CD should be considered a task separate from NLG proper. As one of our aims in **Prodigy** is to enhance reusability, we draw a similar distinction between (i) domain-specific CD and (ii) linguistic generation that can, in principle, be encoded in terms of representations and operations that are reusable from one domain to the next. We define the data-to-text NLG task as the mapping from abstract linguistic content representations to realisations.

2.1 Starting point: the *p*CRU approach

*p*CRU [6] is a probabilistic language generation framework that was developed (in the EPSRC project COGENT) with the aim of providing the formal underpinnings for creating NLG systems that are driven by comprehensive probabilistic models of the entire generation space (including deep generation). NLG systems tend to be composed of generation rules that apply transformations to representations (performing different tasks in different modules). The basic idea in *p*CRU is that as long as the generation rules are all of the form $relation(arg_1, \dots, arg_n) \rightarrow relation_1(arg_1, \dots, arg_p) \dots relation_m(arg_1, \dots, arg_q)$, $m \geq 1, n, p, q \geq 0$, then the set of all generation rules can be seen as defining a context-free language and a single probabilistic model can be estimated from raw or annotated text to guide generation processes.

*p*CRU uses straightforward context-free technology in combination with underspecification techniques (context-free representational underspecification, CRU, [4]), to encode a **base generator** as (i) a set G of expansion rules (of the form above) composed of n -ary relations $relation(arg_1, \dots, arg_n)$ where the arg_i are constants or variables over constants; and (ii) argument and relation type hierarchies. During generation, inputs are expanded under unifying variable substitution until no further expansion is possible. In non-probabilistic mode, the output is the set of fully expanded (fully specified) forms that can be derived from the input. The *p*CRU (*p*robabilistic CRU) **decision-maker** is created by estimating a probability distribution over the base generator from an unannotated corpus of example texts, in two steps. First, the corpus is converted to a **multi-treebank**. For each sentence, all (left-most) derivation trees licensed by G are determined and added to the corpus. In the second step, frequency counts are determined for each individual generation rule from the multi-treebank. The counts are converted into a probability distribution over G , using smoothing and standard maximum likelihood estimation. This distribution is used in one of several ways to drive generation processes, maximising the likelihood either of individual expansions or of entire generation processes.

A *p*CRU generator of weather forecasts in the SUMTIME domain [22], performing text planning, aggregation and realisation, has been implemented and tested extensively [5, 8]. Results showed improved efficiency compared to an equivalent generate-and-select method, and expert judges deemed text quality better than forecasts produced by expert forecasters. The tests also demonstrated robustness and significantly reduced development time (by an estimated 80% compared to a handcrafted system with the same black-box functionality), and that the generator was adaptable to different forecasting styles without any manual overhead.

*p*CRU research in the weather domain provided the proof of concept that probabilistic NLG can enhance system reusability and robustness and still be efficient and high quality. However, reusability was limited to within the same domain and while the simple probabilistic technology worked very well in the SUMTIME domain, it has not been shown to be powerful enough for a range of different, potentially more varied domains. The probabilistic model and application

techniques used so far need to be validated for other domains and, if necessary, extended or replaced. Furthermore, automatically adaptable decision-making is only one side of reusable NLG, what we need to look at next is domain-independent linguistic representation.

2.2 Data resources

We will create and test generators in a range of domains with varying style, complexity and text length, which is important for demonstrating reusability and scalability. NLG corpora are hard to come by, as NLG projects do not tend to distribute reusable data. We will use the three currently available English data resources that can be used for data-to-text NLG, but to ensure enough variation, we will also create two additional corpora:

1. SUMTIME-Wind corpus of paired wind data and forecasts [22]: about 3,500 pairs, text size 1-2 sentences;
2. ILEX corpora of museum database entries and descriptions of exhibits [14]; text size several sentences;
3. BabyTalk corpus of neonatal data and medical reports [17] (about 200 nurses' reports); text size several paragraphs;
4. Prodigy Recipes corpus: new corpus of about 1,000 short cake recipes (lists of instructions of 1–3 sentences) and input representations;
5. Prodigy Census corpus: new corpus of about 500 census statistics from the 2001 UK Census paired with several alternative verbal paraphrases of the statistics written by subjects; text length 1-3 sentences.

The main difficulty in creating corpora for data-to-text NLG is that while inputs and outputs are easily obtained separately, paired inputs/outputs are not a 'naturally occurring' data resource. For Prodigy Census, we will create the outputs manually, and for Prodigy Recipes, we will create the inputs semi-automatically. Doing this ourselves has the advantage that we can pair single inputs with several outputs (at least for evaluation data) which is crucial for evaluation by automatic metric which requires several reference texts instead of a single 'gold standard' reference. Writing paraphrases does not require specific expertise, and we will recruit paid subjects to do this part of the work.

For the Prodigy Recipes corpus, we will 'harvest' recipes from the vast number of available cookery websites (many of which allow free non-commercial use and some of which are not copyrighted) including textual variants of standard recipes. We will develop a semi-automatic method for creating input representations from automatic analyses of recipe texts and lists of ingredients (using existing MRS and SDRT parsers, see also Section 2.3).

For the Prodigy Census corpus, we will use sets of statistics from the 2001 UK Census which the Office of National Statistics makes available "free to all at the point of use"⁶ through the Census Access Project. We will group statistics into three levels of complexity: (i) single-point, e.g.: (Cambridge) People stating religion as: Christian1 63.8 → *63.8% of people living in Cambridge stated their religion as Christian*; (ii) single-dimensional, e.g.: (Cambridge) People stating religion as: Christian1 63.8 Buddhist1 0.6 Hindu1 0.9 Jewish1 0.6 Muslim1 2.1 Sikh1 0.2 Other religions1 0.5 No religion1 23.8 → *Among people living in Cambridge, 63.8% stated their religion as Christian, 0.6% as Buddhist, 0.9% as Hindu, 0.6% as Jewish, 2.1% as Muslim, and 0.2% as Sikh. 0.5% stated other religions, while 23.8% stated they had no religion, and 7.5% did not state a religion*; and (iii) multi-dimensional, where e.g. statistics for Cambridge are compared to East of England and England and Wales.

Compared to the huge corpora characteristic of NLU, these are small data resources. However, application domains in NLG are vastly more restricted than in NLU (which tends to develop and evaluate applications for *unrestricted* language). E.g. the SUMTIME corpus has only 90 words (not counting wind directions) and a handful of syntactic structures, despite being the inputs and outputs collected from a real-world text production task.

2.3 Towards domain-independent representations for NLG

We will create an abstract linguistic content representation formalism for data-to-text NLG that is sharable between all five Prodigy application domains, beginning by generalising over the domain-specific input representations. Approaching domain-independent representation for NLG in this way will produce valuable insights into its feasibility without the more ambitious aim of a fully formalised universal scheme. We will carry out a survey of existing data-to-text systems, comparing the strategies for representing content that they employ. For example, data-to-text systems often use a level at which 'messages' are represented, as in the following train timetable domain example from Reiter and Dale [21]:

```
[ message-id: msg02
  relation: DEPARTURE
  arguments: [ departing-entity: CALEDONIAN-EXPRESS
              departing-location: ABERDEEN
              departing-time: 1000 ] ]
```

We will also look at using elements of existing semantic and discourse representation formalisms, in particular Minimal Recursion Semantics (MRS, [10]) and Segmented Discourse Representation Theory (SDRT, [1]) for which parsers exist. At this stage, we envisage a syntactically flat representation language based around events, entities, attributes of events and entities, and rhetorical relations. To give a rough idea, the following is a simple example of a representation

⁶<http://www.statistics.gov.uk/census2001/op2.asp>.

(along with example realisations) that would work for both the train timetable information example above and the SUM-TIME domain used in previous *p*CRU research (events are labelled E_i , entities X_i , proper names are in double quotes, and assumed is a structured set of relation and argument type definitions):

```
E1:DEPART(X1), X1:''Caledonian Express'', TIME(E1,1000), LOCATION(E1,''Aberdeen''), E2:ARRIVE(X1),  
TIME(E2,2000), LOCATION(E2,''London'')
```

The Caledonian Express leaves Aberdeen at 10am. The Caledonian Express arrives in London at 8pm.

The Caledonian Express departs Aberdeen at 10:00, and arrives in London at 20:00.

```
E1:VEER(X1,X2), X1:S, X2:SW, TIME(E1,_,1800), MANNER(E1,slow), E2:VEER(X2,X3), X3:NW, SEQUENCE(E1,E2)  
S-ly slowly becoming SW-ly. Later NW-ly.
```

Southerly gradually veering south-west by 18:00, then becoming north-westerly.

We will also investigate ways of making the relations and rules that encode the mapping from inputs to realisations (the generation space) more generic. We will incorporate a semantic level of representation based on a simplified version of MRS [10]. This will enable us to reuse and adapt (parts of) the English Resource Grammar [9] for surface realisation, and to focus on generic representation for the deeper parts of generation. Here we envisage two intermediate levels of representation corresponding roughly to text plans and sentence plans [21], the former based on SDRT [1], the latter on underspecified MRS.

Shared input representations and overlapping generation spaces should in principle make it possible to encode a single domain-independent base generator for all Prodigy domains, and then to train the decision-maker on domain-specific corpora to automatically adapt the generator to a given domain. For example, the train departure and weather forecasting domains above have distinct styles. When the decision-maker is trained for the train timetable information domain, it might generate the following for the above input from the weather forecasting domain: *The southerly wind veers south-west by 18:00. The south-westerly wind then becomes north-westerly.* Conversely, the decision-maker trained for weather forecasting might generate the following for the above train timetable input: *Caledonian Express leaving Aberdeen at 10:00, arriving London at 20:00.*

2.4 Probabilistic decision-making

The quality of generated texts depends on representation strategies for encoding generation spaces, but also on the suitability and adaptability of the probabilistic model that controls navigation through the generation space. After producing baseline results with existing *p*CRU techniques, we will investigate alternative techniques for estimating and using probabilistic models: (i) alternative types of corpora — e.g. corpora that have been annotated automatically (using existing MRS [10] and SDRT [1] parsers) instead of ‘raw’ corpora; training from corpora exemplifying a particular style, level of formality, affect, etc., in addition to training from what is known in NLG as a target corpus (a collection of examples of the exact texts the system is supposed to generate); (ii) alternative ways of estimating a probability distribution, including different smoothing techniques and derivation-tree disambiguation; and (iii) alternative techniques for using the probabilistic model during generation, e.g. maximising the likelihood of the intended meaning given the realisation and context.

Implementing and reimplementing generators using existing and newly developed techniques will give us a clear idea of the suitability of probabilistic context-free expansion rules with variable and constant arguments. Using context-free models has a number of important advantages, including trainability from raw corpora and low computational cost in training and generation, but it may prove too restrictive a formalism type. We will therefore also investigate various alternatives, including probabilistic graphical models such as Bayesian belief networks. Exact training algorithms for these tend to be computationally intractable in the general case, so we will look at model subclasses with polynomial training algorithms and at polynomial approximate algorithms.

2.5 Evaluation

There is currently no common NLG evaluation standard. The approach to evaluation involving both automatic and human-based evaluation outlined below conforms to our current ideas for the GENEVAL initiative [20, 7], but we will update our approach in line with any emerging standards.

Automatic evaluation is a viable way of measuring progress during the development of core technology, provided that scores are calculated against reference texts by several authors. However, our main evaluation effort will be in the form of three rounds of combined human and automatic evaluations. Human evaluation is more expensive and time-consuming to organise, but — given a sufficient number of subjects — gives a more reliable estimation of quality, as well as one that is more trusted by the NLG community.

For the automatic evaluations, we will hold out a portion of each corpus (about 5% of input/output pairs) for use as a validation data set. We will use the remainder in 10-fold cross-validation (possibly more, depending on variation) evaluations where the data is split 10 ways and the probabilistic model is estimated (trained) on 9 parts and tested on 1 part, such that each of the ten parts is used once as the test set. The validation set will not be used except in final evaluations. The main automatic metric will be NIST [11] which has been shown [8] to correlate highly with human judgments (Pearson correlation coefficient > 0.8) when comparing different NLG systems. We may also use a range of other metrics including any we will develop in the GENEVAL initiative.

In the human evaluations we will use fewer test pairs (around two dozen, depending on number of systems to be compared), and domain experts (at least for the SUMTIME and BabyTalk domains) as well as non-experts. We will use a similar set-up to the one in previous SUMTIME evaluations [8], i.e. randomly selected test data, and a repeated Latin squares experimental design, where subjects are shown randomised baseline texts and topline texts composed by humans, in addition to outputs from the systems to be evaluated. Subjects will score systems on a 10-point scale, assessing language quality and appropriateness of content.

We will also perform task-based evaluation: for the Prodigy Recipes generators, by recruiting students to use the recipes and then answer questions about ease of understanding, total cooking time, level of expertise etc.; and for the Prodigy Census generators, by asking subjects to fill in the percentages in the original tables (faster if texts are clearer).

3 Research programme

1. Creating data resources: Corpora need to be in the format of paired input content representations and corresponding human-written output texts. All existing corpora to be used already have (domain-specific) content representations, but (apart from the SUMTIME corpus) may require some preparation⁷. The main work will go into creating the two new resources. Ample time and budget is allowed for all resource creation tasks (see Work Packages appendix).

2. Building baseline systems: We will implement simple domain-specific *p*CRU generators for all domains except SUMTIME for which we already have one (which took under a month to build). This will verify previous results for *p*CRU, and provide us with baseline systems to compare newly developed techniques against. We will use automatic evaluation to assess generation space definitions during system building.

3. Evaluation I: In the first round of combined human and automatic evaluations we will assess the performance of the baseline systems against an absolute baseline of randomised generation, existing handcrafted systems, and human-written texts using methodology as described above (Section 2.5). This will give us a clear idea of the general suitability of context-free methods over a range of domains.

4. Prodigy input representation formalism: We will begin by generalising over the domain-specific representations, moving on to a single content representation formalism that is generic with respect to all Prodigy domains.

5. Reimplementing baseline systems: We will reimplement the five baseline generators, using the new Prodigy input formalism, also maximising overlap between generation space definitions. During system building we will again use automatic evaluation which will help assess the quality of different generation space encodings.

6. Evaluation II: As evaluation round I, but for the reimplemented generators.

7. Prodigy probabilistic control: We will investigate alternative methods for probabilistic control, in particular different ways of obtaining and exploiting probabilities. Continuing to use the domain-independent input representation formalism, we will also investigate alternatives to context-free models and techniques. We will continuously evaluate progress against baseline technology by automatic methods. The aim is to create a range of tools based on best-performing techniques to support creation of probabilistic NLG systems.

8. Final implementation of generators: Using the Prodigy input formalism and final probabilistic methodology, we will create final versions of generators for the five domains.

9. Evaluation III: We will thoroughly assess the success of the final Prodigy technology by human evaluations of text quality as well as task-based evaluation of the usefulness of the recipe texts and census paraphrases.

Most of the research will be carried out by the research fellow, but the principal investigator (PI) will take an active role in developing representation strategies and techniques for probabilistic control. The PI will also carry out most of the work on creating the Prodigy Recipes and Census corpora (recruiting paid subjects for creating outputs for the latter). Research effort is divided in this way, because a corpus similar to Prodigy Census will also be used in GENEVAL (see Track Record section above), and because it is a self-contained work package requiring a different set of skills from the work on domain-independent representation and probabilistic control.

Among the affiliated project members, Ehud Reiter will advise on the Prodigy input formalism and evaluation techniques, and Khalid Choukri will advise on evaluation methods, corpus creation and resource distribution.

Details and deliverables for all work packages implementing Items 1 to 9 above, as well as software packaging, can be found in the Work Packages appendix to this proposal.

4 Prodigy project in context

Beneficiaries: The Prodigy project will make a theoretical and technological contribution to advancing the field of NLG in terms of system reusability, robustness and development times, through development of generic approaches to linguistic representation for NLG and use of automatically adaptable probabilistic techniques.

⁷We have two back-up corpora should the effort involved prove infeasible: Dale's StockReporter data and Reiter et al.'s corpus of pollen forecasts.

The research community will also benefit in tangible terms through the resources and software tools that we are planning to release under appropriate research licencing conditions. The former will help produce further comparable evaluation results which are much needed in the field, and the latter will make it possible for researchers with no expertise in statistical methods to use probabilistic NLG techniques.

There is currently virtually no commercial use of NLG technology. Methodology to be developed in **Prodigy** will help overcome some of the issues standing in the way of commercial deployment.

Dissemination: We will report our research through conventional academic means such as conference presentations, journal publications and the establishment of a website providing more technical information and specific deliverables.

A crucial component in our work on standardising representations will be a continuous dialogue with the NLG community, and to this end we will establish a panel of advisors made up of the two affiliated project members (see Project Team, p. 2), and several other representative members of the NLG community.

Releases of resources will be publicised on the relevant mailing lists (SIGGEN, Corpora, etc.). We may use some of the data in shared task evaluations in the GENEVAL work.

We will generally seek to achieve high visibility. Print and broadcast media have in the past been very interested in applied NLG projects such as SUMTIME and BabyTalk at Aberdeen⁸.

References

- [1] N. Asher and A. Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.
- [2] S. Bangalore and O. Rambow. Corpus-based lexical choice in natural language generation. In *Proceedings of ACL'00*, pages 464–471, 2000.
- [3] R. Barzilay and M. Lapata. Collective content selection for concept-to-text generation. In *Proceedings of HLT/EMNLP'05*, Vancouver, 2005.
- [4] A. Belz. Context-free representational underspecification for NLG. Technical Report ITRI-04-08, Information Technology Research Institute, University of Brighton, 2004.
- [5] A. Belz. Statistical generation: Three methods compared and evaluated. In *Proc. of ENLG'05*, pages 15–23, 2005.
- [6] A. Belz. pCRU: Probabilistic generation using representational underspecification. Technical Report NLTG-06-01, NLTG, CMIS, University of Brighton, 2006.
- [7] A. Belz and A. Kilgarriff. Shared-task evaluations in HLT: Lessons for NLG. In *Proc. INLG'06*, to appear.
- [8] A. Belz and E. Reiter. Comparing automatic and human evaluation of NLG systems. In *Proc. EACL'06*, pages 313–320, 2006.
- [9] A. Copestake and D. Flickinger. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of LREC'00*, 2000.
- [10] A. Copestake, D. Flickinger, and I. Sag. Minimal recursion semantics: An introduction. Draft, available online at <http://www-csli.stanford.edu/aac/papers/newmrs.ps>, 1999.
- [11] G. Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the ARPA Workshop on Human Language Technology*, 2002.
- [12] R. Evans, P. Piwek, and L. Cahill. What is NLG. In *Proceedings of INLG'02*, pages 144–151, 2002.
- [13] G. Gazdar. Paradigm merger in NLP. In Robin Milner and Ian Wand, editors, *Computing Tomorrow: Future Research Directions in Computer Science*, pages 88–109. Cambridge University Press, 1996.
- [14] A. Isard, J. Oberlander, I. Androustopoulos, and C. Matheson. Speaking the users' languages. *IEEE Intelligent Systems Magazine: Special Issue on Advances in Natural Language Processing*, 18(1):40–45, 2003.
- [15] I. Langkilde. Forest-based statistical sentence generation. In *Proceedings of ANLP-NAACL'00*, pages 170–177, 2000.
- [16] M. Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL'03*, 2003.
- [17] A.S. Law, Y. Freer, J.R.W. Hunter, R.H. Logie, N. McIntosh, and J. Quinn. A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. *Journal of Clinical Monitoring and Computing*, 19:183–194, 2005.
- [18] A. Oh and A. Rudnicky. Stochastic language generation for spoken dialogue systems. In *Proceedings of the ANLP-NAACL'00 Workshop on Conversational Systems*, pages 27–32, 2000.
- [19] M. Poesio, R. Henschel, J. Hitzeman, and R. Kibble. Statistical NP generation: A first report. In R. Kibble and K. van Deemter, editors, *Proceedings of ESSLLI Workshop on NP Generation*, pages 30–42, 1999.
- [20] E. Reiter and A. Belz. GENEVAL: A proposal for shared-task evaluation in NLG. In *Proceedings of INLG'06*, to appear.
- [21] E. Reiter and R. Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997.
- [22] S. Sripada, E. Reiter, J. Hunter, and J. Yu. Exploiting a parallel TEXT-DATA corpus. In *Proceedings of Corpus Linguistics 2003*, pages 734–743, 2003.

⁸E.g. New Scientist reported SUMTIME work on weather forecast generation in issue 2518, 28 September 2005, page 27; BBC Scotland reported BabyTalk research, see <http://news.bbc.co.uk/1/hi/scotland/4669466.stm>.