The COGENT Project is looking at generic (wide-coverage and reusable) NLG, in particular the issue of nondeterminism (multiple outputs for the same input) and its two main sources, wide syntactic/lexical coverage and underdetermined semantics. Nondeterminism can be controlled in two ways: (i) increasing the degree of specificity of the inputs to generation, and (ii) reducing the set of realisations consistent with a given input. Most surface NLG that aims to be generic requires either highly specified inputs (e.g. FUF/SURGE, KPML) with no informed choice among alternative realisations, or permits underdetermined inputs and uses selection mechanisms for alternative realisations (e.g. statistical NLG). In both cases, the degree of specificity of the inputs is fixed, which may imply having to make random or default choices, or having to consider vast numbers of alternatives (e.g. trillions in Nitrogen according to Langkilde, 2000).

Context-Free Representational Underspecification

CRU is a framework for defining underspecifiable representation formalisms. A CRU-grammar $G$ is formally a context-free grammar with atomic features, and defines both a fully specified language, $L(G)$, and the different ways in which it may be underspecified. CRU-grammars are intended for generating semantic expressions, rather than strings of words. The set of underspecified expressions, denoted $U(G)$, is the set of all partially derived forms, where a partially derived form is a sentential form which contains at least one nonterminal and from which at least one string in $L(G)$ can be derived. The idea is that every string in $L(G)$ can be underspecified by any of its partially derived forms.

CRU-grammars can encode generation spaces as in the diagram above. The grammar on the right encodes a toy generation space based on one of the examples in Reiter & Dale, 1997. It generates 24 structured semantic representations, e.g.:

1. $\text{DEF(x, caleonian_express(x), leaves london at AND(10am, 6pm))}$
2. $\text{DEPARTURE --> EXISTS(H0, X, H1, H2) TRAIN(H1, X) DEPART(H2) LOC_TIME(H2)}$
3. $\text{AM(H) --> am(H)}$
4. $\text{DEPART(city, H0, H1, H2) --> and(H0, H1, H2)}$
5. $\text{AND(H0, H1, H2) --> and(H0, H1, H2)}$
6. $\text{LOC_TIME(H) --> London(H) at(H) AND(H, H1, H2)}$
7. $\text{5(H1) AM(H1) 1(H2) PM(H2)}$
8. $\text{AM(H) --> am(H)}$
9. $\text{PM(H) --> pm(H)}$
10. $\text{DEPART(city, H0, H1, H2) --> and(H0, H1, H2)}$

For this fully specified expression (“the Caledonian Express leaves London at 10am and 6pm”), the grammar licenses 80 different underspecified expressions (achieved in CFG framework by a standard syntactically flat encoding), including:

- $\text{EXISTS(X, TRAIN(X), DEPART london at AND(10am, 6pm))}$
- $\text{DEPART(city, H0, H1, H2) --> and(H0, H1, H2)}$
- $\text{AM(H) --> am(H)}$
- $\text{PM(H) --> pm(H)}$
- $\text{DEPART(city, H0, H1, H2) --> and(H0, H1, H2)}$
- $\text{LOC_TIME(H) --> London(H) at(H) AND(H, H1, H2)}$
- $\text{5(H1) AM(H1) 1(H2) PM(H2)}$

The grammar exemplifies two of the most commonly used underspecification mechanisms: metaconstants and structural underspecification. The AND-rules are an example of the latter, as they underspecify argument order: \text{AND(10am, 6pm)} expands to \text{AND(10am, 6pm)} and \text{AND(6pm, 10am)}. This kind of logical underspecification can help address the logical form equivalence problem (Shieber, 1993). The remaining rules are examples of underspecification by metaconstant, where some nonrecursive expression is expanded to a prespecified set of expressions. Taking a more linguistic view, the phenomena being modelled include lexical variation (synonyms, paraphrases), referring expressions and propositional logical equivalence.

What’s new in terms of underspecification?
- Enumeration of fully specified forms is tractable.
- CRU-grammars define a hierarchy of underspecified forms.
- There is no restriction on partial underspecification.
- There is a straightforward probabilistic extension (PCFGs).
- CRU allows an arbitrary number of different degrees of underspecification (rather than just two).

Work in progress …
- Full formal specification of CRU.
- Toy CRU-grammar for patient information leaflets.
- Flexible CRU interface to the LKB/ERG realiser.

Acknowledgements:
- Supported under EPSRC Grant GR/S24480/01.