DEMONSTRATION OF GENESYS:
A VERY LARGE, SEMANTICALLY BASED SYSTEMIC FUNCTIONAL GENERATOR

Robin P. Fawcett and Gordon H. Tucker

Computational Linguistics Unit
University of Wales College of Cardiff
e-mail: fawcett@uk.ac.cardiff, tuckerg@uk.ac.cardiff

This paper provides background material to the demonstration to be given at COLING '90 of the GENESYS component of the COMMUNAL system. A presenter of such a demonstration should say (1) what the system is good for; (2) why it is good for it; and (3) what makes it different from alternative systems. The system to be demonstrated is just a **part** (though the single most important part) of a much more complex system, and some of the answers to the questions must be related to the overall system. So I shall describe that first, and then the generator itself.

1. The COMMUNAL Project

The acronym **COMMUNAL** stands for **COnvivial Man-Machine Understanding through NAtural Language**. The long-term goal of the project is to contribute to the development of systems that will enable computationally naive people to interact naturally ('convivially') with the Intelligent Knowledge Based Systems (IKBSs) that we should expect a decade ahead. The reason why it promises well is that it is a holistic, integrated approach, built around a strong, rich theory of natural language in social interaction (Halliday's **systemic functional grammar (SFG)**. The concept of **choice between meanings**, which lies at the heart of the model, fits particularly naturally with the concept of **planning**, and this is probably why so many successful models of generation have drawn, either explicitly or implicitly, on SFG. Some of the differences from a well-known sister project (Penman) will be mentioned below.

The project is based at the University of Wales College of Cardiff and the University of Leeds, with financial support coming from RSRE Malvern and industry. Phase 1 ran from 1987-89, and the main funding has now been secured for Phase 2, which runs 1990-93 (with further negotiations in progress to enhance this).

In a young field such as natural language processing (NLP) it is important to develop alternative approaches. The COMMUNAL Project is based on a number of principles, some of which are different from the mainstream assumptions, with some being explicitly innovative. These principles include the following:

(1) The way in which 'knowledge' (or, preferably, 'belief') is organized should be influenced by the way in which language is modelled, rather than vice versa;

(2) Priority in the system is given to the **generator**, with the **parser** and **semantic interpreter** being derived from the generator;

(3) The organization of the grammar in the generator gives priority to **paradigmatic** (i.e. **systemic**) relations rather than to **syntagmatic** (i.e. **structural**) relations, and to the level of **semantics** (or **functions**, in Halliday's broad sense of the term), rather than to **form** (syntax, etc): i.e. it is a **systemic functional grammar**;

(4) The meanings of **lexical items** (i.e. vocabulary) are modelled in the same way as the meanings of syntax, i.e. in system networks; there is therefore no separate 'lexicon' (though a word list is available to the parser).

(5) To develop a full model of generation, it is necessary to model generation in the framework of **social interaction**, and not only monologue.

(6) Attention is given to the paradigmatic systems and syntagmatic structures of **discourse** (as well as sentences).

(7) Strong emphasis is placed on the semantic generation of **intonation** (as well as **punctuation**), thus generating an output ready for speech synthesis.

For a fuller overview of the principles underlying the project, see Fawcett 1986 and 1988; for the particular version of SFG around which the project is built see Fawcett 1980, 1987 and to appear b.)

The project has three sub-teams, each of whom works closely with the others. In Phase 1 these worked on: (1) **language generation** (at Cardiff: Fawcett, Tucker and Wright, assisted part-time by Tench and Young); (2) **language parsing and understanding** (at Leeds: Atwell (part time, the Leeds team leader), Souter and O'Donoghue); and (3) **beliefs, inferencing and planning** (at Cardiff: Wright and, part time, Atkinson). Thus, while the project has a firm linguistic base in NLP, it already encompasses, in a small measure, other central aspects of artificial intelligence (AI), and these will be greatly extended in Phase 2.

Note that COMMUNAL is an **interactive** system, rather than one that only generates **monologue**. In outline, the major components of the overall model are:

1. The **parser**.
2. The **semantic interpreter**.
3. The **belief system**, which includes general and specific beliefs about ('knowledge of') situations and things in some domain (currently, personnel management in a large organization); specific beliefs about the content of the preceding discourse, about various aspects of the current social situation, about the addressee(s) and their beliefs of all types, their attitudes, and their goals and plans.
4. The **planner**, which makes general plans, drawing on knowledge of:
   (a) **genres** (scripts, schemas, etc), introducing where appropriate sub-units such as **transactions** (see below) and more detailed plans, using
   (b) the **local discourse grammar**, which is modelled as a 'systemic flowchart' (i.e. a flowchart containing many small system networks at the choice points, and which generates **exchanges** and their structure),
5. the **lexicogrammar**, i.e. the **sentence generator** (see below).

There were four main achievements in Phase 1. The first was the size and scope of the **generator**. Its two main sub-components are the **system networks**, which are described more fully in the next section. The second achievement was to build a large **probabilistic parser**. The Realistic Annealing Parser (the RAP) parses any output from GENESYS on the basis of co-occurrence probabilities. In Phase 2 we shall also build a **non-probabilistic parser**, with the intention of incorporating the advantages of **both** in the final system. The third major achievement relates to the **semantic interpreter**. It has been a long-standing goal in NLP to build a large scale system that uses the **same grammar** to either **generate** or **interpret** a sentence. (Many current systems use a different grammar for each process.) We are developing at Leeds an interpreter which, roughly speaking, runs the realization rules in reverse, drawing where appropriate on the system networks. This component (which is called REVELATION because it reveals the 'meaning' from the 'wording') depends on the generator being built in such a way that this is possible; this will be refined and reported on fully in Phase 2. The final achievement is that small but principled mini-components were built for an exemplar situation to demonstrate the passage of information through the whole system; these will be considerably developed during Phase 2.

2. The sentence generator: GENESYS

The generator is called GENESYS, because it GENErates SYStemically, i.e. by using a systemic functional grammar. It is a **lexicogrammar** (to use the term of Halliday, the chief architect of the theory), because it generates from one unified semantic component both **grammar** (in the sense of syntax and grammatical items) and **lexical items** - as well as **punctuation** or **intonation**, as required. Its two main subcomponents are:

(a) The **system networks.** These are complex networks of choices between a wide variety of types of **semantic features**. Some denote **situations** and are realized in the clause. Types of meaning covered include **theme** and **information structure** as well as **transitivity, mood, negativity, modality, affective meaning** and **logical relationships.** There are equivalent system networks for the semantics of **things** and **qualities.**

(b) The **realization rules,** which turn the **selection expressions** (or 'bundles') of semantic features that are the output from passes through the system networks into quite relatively richly labelled **syntactic structures** - and into the entities at their terminal nodes, i.e. **items** (grammatical and lexical) and markers of **punctuation** or **intonation.** (For those unfamiliar with systemic grammars, it would help to see a demonstration; see also Tucker 1989.)

Here are a few facts to give you a perspective on GENESYS at the end of Phase 1. McDonald, Vaughan and Pustejovsky (1987:179), referring to a well-known NLP project at the University of S. California which also uses SFG, say: 'Nigel, Penman's grammar .... is the largest systemic grammar and possibly the largest machine grammar of any kind.' Although GENESYS was developed independently, it has in two years grown to be even larger than Nigel (using the criterion employed by Mann (the Director of Penman till 1989) to characterise Nigel's size; see further below). (Nigel reached its present size some years ago; it has not grown in the interim because the team have been working on other components of Penman). The first point about GENESYS is therefore its size and scope.

Areas of grammar covered so far include the following (using theory-neutral terminology possible, sometimes supplemented by specifically systemic functional terms): complex structures realized in any of eight different auxiliary verbal elements, most of which can co-occur in the same clause, realizing choices in MOOD, MODALITY, TENSE, ASPECT and VOICE, etc (including, as well as the usual forms that are standard in all treatments, "used to", "would" in its 'habit' sense, "be going to", "be about to"); an almost complete range of TRANSITIVITY types (defined in terms of configurations of participant roles, including covert roles); three types of 'passive' construction; many types of verb COMPLEMENTATION (including some 'extraposition'), finite and non-finite, handled as embedded clauses and regarded as situations that 'fill' the Phenomenon; realizations of INFORMATION STRUCTURE in both punctuation and intonation; in MOOD both polar and Wh-questions (including multiple Wh-items in the same clause); five types of potentially co-occurring TIME ADJUNCTS (including two types of frequency, and repetition); adjuncts of PLACE AND MANNER; Adjuncts expressing a LOGICAL RELATIONSHIP (e.g. clauses with "because ...", "if", purposive "to ..."); marked THEMATIZATION of all non-Subject PARTICIPANT ROLES (Complements) and of all types of CIRCUMSTANTIAL ROLES (realized in Adjuncts); the handling of MANNER ADVERBS in the same network as ADJECTIVES, with provision for COMPARATIVE and SUPERLATIVE meanings and forms, both regular and all types of irregular forms; a representative range of TIME ADVERBS; a full range of irregular verbs and noun plurals; complex NOMINAL GROUPS with provision for mass and count nouns, for NUMBER, for appropriately sequenced multiple pre-head MODIFIERS (which may themselves be filled by structures such as "fairly rich"), for three types of determiner selection, as in "five of the biggest of those apples", (i.e. (a) selection by QUANTIFICATION, including weak "one" and "a(n)", (b) selection by SUPERLATIVISATION, and (c) selection by one of three types of DEIXIS, realized in "the", demonstratives and possessives), and for post-head QUALIFIERS filled by prepositional groups (with clauses as qualifiers (i.e. 'relative clauses') currently being added); a full range of PRONOUNS (personal, possessive, demonstrative and indefinite) and PROPER NOUNS (names of individuals, with their own quite complex internal grammar, names of social groups, names of places); complex genitive constructions, e.g. "the new doctor's car's door"; PREPOSITIONAL GROUPS, with a representative range of PREPOSITIONS; a wide range of TEMPERING items for use with adjectives and adverbs (with embedded groups and clauses as in "bigger than him/it used to be" currently being covered); 'special' grammars of dates, addresses and human proper names; and much else.

Examples of sentences that can be generated include:

(1) I like being here.
(2) Ivy might have been going to be being looked at by them, but she ran out of the room.
(3) The Director doesn't like it that the new manager was about to leave because we didn't give him a bigger office.
(4) Some of the most unpleasant of that rather angry man's better friends are amazingly rich.
(5) | until next month/T/12 | he will be living at eleven Romilly Crescent/T/2 | Canton/NT/1 | (i.e. with intonation marked for input to a speech synthesizer, where "1" = fall and "2" = rise)

Note that so far the system only provides principled **motivation** for choosing between the semantic options for only a small subset of the semantics. The research strategy, as stated above, is first to model the full richness of the semantics and its realizations. These areas will be greatly extended in Phase 2.

An important theoretical difference between GENESYS and Nigel is that the system networks in GENESYS are structured specifically to express **semantic** choices, while Nigel's are at the level of **lexicogrammatical** form (though reflecting 'meaning', as always in a systemic grammar).

Let us attempt the difficult task of a comparison. If, following Mann, we use the number of systems to characterise the grammar's size, we find that Nigel has a little under 400 systems (i.e. 'choices between features') realized in grammar (Bateman, personal communication 1990), while GENESYS has about 600 (the greater number possibly reflecting the explicit semanticity of GENESYS). GENESYS also has hundreds of **systems that generate vocabulary** that are integrated into the lexicogrammar (as Halliday has always suggested should be done). In Nigel there is a separate lexicon (which was unfortunately necessary to meet a sponsor's research collaboration requirements with a more traditional approach), whereas in GENESYS the networks generate not only syntax and morphology but also lexis, intonation and punctuation. At the end of Phase 1 GENESYS had a total of around **1,100 systems** of semantic features. In addition GENESYS has some **1,400 realization rules**, involving about **4,000 operations, and these appear to be more complex than those in Nigel**. This results in part from the semanticity of the networks, and in part from the complexity of the phenomena covered (See Tucker 1989 for an overview of GENESYS at work.)

The above informal comparisons with the Nigel system illustrate the difficulty of comparing even closely related grammars. The fact is that there are **no agreed criteria for measuring grammars**, and even less are there criteria for evaluating **holistic models of language** that have (in addition to the usual areas of syntax and morphology), vocabulary, intonation and punctuation - and, in the case of GENESYS, **semantic networks that generate all of these**. Despite the lack of agreed criteria it is clear that GENESYS is already one of the largest of such models in existence.

As Phase 2 gets under way, we are already increasing the model's coverage. By its end we expect to more than double the number of grammatically realized systems, and so to be able to handle something approaching unrestricted syntax. (We have plans to implement certain novel possible solutions to some well-known syntactic problems, but even if all are successful with these there will of course still be many problems that remain at the end of Phase 2.) The **vocabulary in GENESYS** should grow to **3-4,000 word-senses** (or possibly more if an ambitious planned procedure is successful), and it will have a near-complete coverage of both **punctuation** and **intonation**, the latter being a complex matter where there is little previous work to build on; for the generation of **semantically motivated intonation** see Fawcett 1990b. GENESYS needs a special speech synthesizer to turn this into a phonetic output; we hope that there will be a sister project at University College London to work with us on this. Finally, GENESYS introduces **probabilities** into the operation of the system networks; see Fawcett (to appear a).

GENESYS was developed using Poplog Prolog on a SUN 3/50, using a special tool for writing and testing grammars developed by Wright (called DEFREL, because it defines relations).

3. Other future developments

In a complete model of generation we need, as well as a sentence generator, rich components for **belief, inferencing** and **planning** (which includes **discourse planning**), as described in section 1.

At present these components are much less well developed in COMMUNAL than in some other projects, as is to be expected in a project that is explicitly 'language-led'. But we do already have small but principled components that enable the system to accept sample utterances from the parser and interpreter; to add appropriately to its beliefs; to draw inferences from a new belief; and to make an appropriate discourse plan as input to GENESYS. (Comparisons with Penman are even harder here, because Penman generates monologue, not dialogue.) Phase 2 will develop these components further, exploring several new ideas. The parsers and semantic interpreter will also be developed further.

There are many possible **spin-off applications**.

One long-standing goal to which we hope to make a significant contribution is what is known as **text to speech**. This is the process of mechanically turning written text into natural-sounding speech. This must include **semantically motivated intonation**, which is something that has been handled inadequately so far in work on text to speech, but which GENESYS has already made good progress in attending to in a principled way. Other possible spin-off applications include interactive tutors for **automated language learning** (potentially achievable as an application because of the possibility of using quite limited domains), the development of **metrics of text complexity**, and possible applications in the **games industry**. At the appropriate point we shall seek industrial interest in our original sponsors and in others for these (and no doubt other) possible applications.

This project is still young (barely two years old, in terms of actual research). Many of the reports are still confidential to the partners, but papers referring to the project are now beginning to appear (see the references). As Phase 2 gets under way, COMMUNAL is already becoming increasingly well known to the relevant research communities, both academic and industrial.

If you would like to visit the project, please contact us.

The research reported here was supported by grants from RSRE Malvern under contract no. ER1/9/4/2181/23, by the University Research Council of International Computers Ltd, and by Longman.

References

Note that, for reasons of space, these have mainly been limited to papers about COMMUNAL that are in the public domain. Offprints/copies of the papers listed here by Fawcett and Tucker are available from the address below.

Berry, M.M., Butler, C.S., and Fawcett, R.P., (eds.) to appear. *Meaning and Choice in Language: Studies for Michael Halliday. Volume 2 Grammatical Structure: a Functional Interpretation. Newark, N.J.: Ablex.*

Fawcett, R.P., 1980. *Cognitive linguistics and social interaction: towards an integrated model of a systemic functional grammar and the other components of an interacting mind.* Heidelberg: Julius Groos and Exeter University.

Fawcett, R.P., 1986. 'Meaningful relationships'. Interview with Tony Durham, published in 'Over the horizon' column in *Computing: the Magazine* (6.2.86), pp. 6-7.

Fawcett, R.P., 1987. 'The semantics of clause and verb for relational processes in English'. In Halliday, M.A.K., and Fawcett, R.P. (eds.) 1987. *New developments in systemic linguistics, Vol 1: Theory and description*. London: Frances Pinter.

Fawcett, R.P., 1988. 'Language generation as choice in social interaction'. In Zock, M., and Sabah, G., (eds) 1988b. *Advances in natural language generation Vol 2.* London: Pinter.

Fawcett, R.P., 1990a. 'The COMMUNAL Project: two years old and going well'. In *Network No. 13.*

Fawcett, R.P., 1990b. 'The computer generation of speech with semantically and discoursally motivated intonation'. In *Procs of 5th International Workshop on Natural Language Generation*, Pittsburgh.

Fawcett, R.P., to appear a. 'A systemic functional approach to selectional restrictions, roles and semantic preferences'. Accepted for *Machine Translation.*

Fawcett, R.P., to appear b. 'A systemic functional approach to complementation'. In Berry, Butler and Fawcett, to appear.

McDonald, D.D., Vaughan, M.M., and Pustejovsky, J.D., 1987. 'Factors contributing to efficiency in natural language generation'. In Kempen, Gerard, (ed) 1987. *Natural language generation.* Dordrecht: Martinus Nijhoff.

Tucker, G.H., 1989. 'Natural language generation with a systemic functional grammar'. In *Laboratorio degli studi linguistici 1989/1.* Camerino: Italy: Universita degli Studi di Camerino (pp.7-27).

Tucker, G.H., to appear. 'Cultural classification and system networks: a systemic functional approach to lexical semantics'. In Berry, Butler and Fawcett, to appear.

Contact address: Dr R.P. Fawcett, Director, Computational Linguistics Unit, Aberconway Building, University of Wales College of Cardiff, Cardiff CF1 3EU, UK. e-mail: fawcett@uk.ac.cardiff and tuckerg@uk.ac.cardiff

25.5.90