THE PENMAN LANGUAGE GENERATION PROJECT

Information Sciences Institute
of the University of Southern California

William C. Mann
Eduard H. Hovy

The Penman Project at USC/ISI has been conducting research in natural language processing since 1978. It presently consists of six technical staff, organized into three principal efforts. The first two are funded by DARPA: - Natural language generation (Penman)
- Text structure development (text planning)
- Natural language understanding (parsing)

The natural language sentence generation program Penman provides computational technology for generating English sentences and paragraphs, starting with input specifications of a non-linguistic kind. The research goals underlying Penman are threefold: to provide a useful and theoretically motivated resource for other research and development groups and the computational community at large, to provide a framework in which to conduct investigations into the nature of language, and eventually to provide a text generation system that can be used routinely by system developers. Penman is being used by computer scientists (as the output medium of their programs, among others projects in human-computer communication, expert system explanation, and interface design) and by linguists (as a reference and research tool).

Nigel, the English grammar, is the heart of the single-sentence component of Penman. Nigel is a network of over 600 nodes, each node representing a single minimal grammatical alternation. Nigel is based on the theory of systemic linguistics (a theory of language and communication developed by Halliday and others). Penman contains a noun group planner that is still under development. It also contains a number of additional information resources, such as a lexicon of words and a very general taxonomic model of the world, which is used to categorize the entities of any domain for which it is to generate language. Finally, over the last two years, we have been investigating the planning and generation of multisentential paragraphs. In order to plan coherent order of clauses, we used relations from Rhetorical Structure Theory of Mann and Thompson, operationalized in the form of plans. Using these plans, the text structure planner operates in top-down hierarchic expansion fashion, patterned after the system NOAH.

Penman is currently being used by three domains at USC/ISI:
- An integrated multimedia interface system (II), in which paragraphs of English text, planned and generated by Penman, are combined with maps, menus and other display methods, so as to be suitable for command and control use. The II Project is being led by Dr. Yigal Arens.
- The Program Enhancement Advisor (PEA) is an experimental expert system that interactively advises programmers on how their Lisp programs might be improved. It contains an explanation facility that uses Penman's grammar to generate text that explains how PEA works. PEA is being developed as a Ph.D. project by Johanna Moore.
- The Digital Circuit Diagnosis system (DCD) is an experimental expert system that diagnoses faults in digital hardware. Like PEA, it contains an explanation facility that uses Penman's grammar to generate output. Text is generated that explains the definitions of entities within

DCD and the reasoning that lead to the diagnosis. DCD is being developed by Dr. Cecile Paris.

Recent achievements include the development of an input notation that is very flexible and easy to use, as well as the completion of a set of documentation about Penman. We have started distributing the system at a nominal cost, and are constantly searching for new users.

Penman is one of the most comprehensive language generation programs in the world today; it can generate, in some way, almost any information that can be represented. Our goals are to extend Penman and to provide ways of controlling it in order to generate multiple versions of the same input, as well as to complete the paragraph structure and noun group planners.