Automated Multi-Purpose Text Processing

John Wilkinson and **Chrysanne DiMarco**
Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
E-MAIL: jgwilkin@palestrina.uwaterloo.ca
E-MAIL: cdimarco@logos.uwaterloo.ca
VOICE: (519) 888-4443; FAX: (519) 885-1208

Abstract

Multiple versions of a document may need to be produced for different purposes and so may require quite different stylistic structures. This paper describes how we are building stylistic control into both natural language parsers (Pundit) and generators (Penman) to handle variations in hospital patient educational materials. Stylistic knowledge is formally represented in terms of a multi-level stylistic grammar which defines a systematic correspondance between low-level syntactic structures and high-level stylistic effects.

1 The importance of style in natural language processing

The importance of dealing with stylistic aspects of language in computational systems is undeniable. People communicate a great deal of information through stylistic nuances, and a knowledge of how these subtleties influence meaning is part of a full understanding of language. Systems that could analyze the effects of style on communication would provide information about the implicit meaning that is contained in a text. And generation systems that could control style would produce text that intentionally conveys a specific communicative effect. Both stylistic analysis and generation could be used in applications, such as text critiquing, second-language instruction, and machine translation, for which understanding the effects of how something is said is as important as understanding what is said. Ultimately, computational stylistics should be a part of any system that attempts to deal with 'real-world' language.

But very few natural language understanding systems have attempted to deal with issues of style,[1] and those that do have generally taken a simplistic and heuristic approach. Stylistic analysis has not yet developed the systematic and rigorous methods of syntactic analysis and semantic interpretation. Part of the reason is obvious: understanding style is hard. Stylistic effects are difficult to articulate and even more difficult to define.

2 Previous and current research

In the past several years, we have worked to address the problems of *syntactic style*, understanding how particular syntactic structures can convey corresponding stylistic effects. We are currently involved in a collaborative research project with Dietmar Rösner's group at the Forschungsinstitut für anwendungsorientierte Wissensverarbeitung (Ulm, Baden-Württemberg) on "Multi-purpose text generation from knowledge bases." The two participating subprojects are both concerned with the generation of multiple versions of natural language text from a single knowledge base. In their TechDoc project, Rösner's group is addressing issues of multilingual generation in the production of technical manuals. In our HealthDoc project, we are developing systems for the automated generation of customized hospital patient education materials. Both groups are studying problems in text planning, sentence planning, and lexical choice that are common to the generation of multiple versions of documents, whether multilingual or multi-purpose. Our research has focused on the stylistic complexities of real-world natural language analysis and generation in applications where the subtleties of language can significantly influence the effectiveness of the text produced. Our approach is to develop and use theo-

[1] By *style*, we do not mean literary style, but rather the style of texts such as high-quality magazines and newspapers, technical manuals, and business correspondence.

ries of text composition—in particular, theories of rhetorical structure and lexical semantics and fine-grained meaning. This work brings together several elements of earlier research by our group: unilingual lexical choice (Miezitis 1988); representing and preserving stylistic nuances in translation (DiMarco 1990; DiMarco and Mah 1994); and, more generally, analyzing and generating stylistic nuances in text (Green 1992; DiMarco and Hirst 1993 Makuta-Giluk and DiMarco 1993; Hoyt 1993; Hoyt and DiMarco 1994; Hirst 1995; Green and DiMarco (to appear)).

In this paper, we take the problem of formally representing stylistic knowledge as our starting point. It is our belief that stylistic knowledge must first be formalized, rendered in a well-defined representation, before a computational analysis of style can be attempted. And it is our further contention that a formal representation will facilitate a very transparent implementation. We show how a formal representation of syntactic style can be used as the basis for a general-purpose stylistic analyzer that could be used as a component of a natural language processing system. Then, we discuss how this work is being extended to the development of natural language generation systems that can account for pragmatic nuances of text.

3 A theory of syntactic style[2]

A formal representation of stylistic knowledge should ideally be based on an underlying linguistic theory: there should be a vocabulary of concepts, a clear definition of how the concepts are related, and a systematic way for building new concepts out of existing ones. In our earlier work, we presented a computational theory of syntactic style that is a multilevel representation of stylistic grammar rules. This section summarizes the details of this work as presented in (DiMarco and Hirst 1993). Green (1992) develops the linguistic underpinnings of the theory; Hoyt (1993) presents the representation of the complete theory in a syntactic stylistic grammar.

3.1 Fundamental concepts

In designing a computational theory of style, we constructed a vocabulary of stylistic concepts at three levels of abstraction:

- *Primitive elements* are stylistically significant syntactic properties of sentence components.
- *Abstract elements* are general stylistic properties of groups of sentences.
- *Stylistic goals* are the writer's intentions for high-level pragmatic properties of text.

At all levels, the guiding principle of the theory is that style is *goal-directed*, that is, linguistic choices are made to achieve specific stylistic goals, such as clarity or abstraction. Therefore, we tie low-level syntactic choices to high-level stylistic goals. The fundamental concepts that are used to integrate the multiple levels of the theory are stylistic *concord* and *discord*, which we define as follows:

Concord: A stylistic construction that conforms to the norm for a given genre.

Discord: A stylistic construction that deviates from the norm.[3]

3.2 Primitive elements of style

At the lowest level of the theory, there are two views of sentence structure, *connective* and *hierarchic*:[4]

Connective ordering: The result of cohesive bonds drawing together components in a linear ordering.

Hierarchic ordering: The result of bonds of subordination and superordination drawing together components in a nested ordering.

The connective and hierarchic orderings are used in the definition of primitive stylistic elements to provide a precise syntactic basis to the theory, yet also allow a mapping to the abstract elements.

We use the terms *conjunct* and *antijunct* with superscripts to indicate the degree of connectivity or disconnectivity. Syntactic components are classified as either *conjunct*^{5} or *conjunct*^{6} (excessively connective), *conjunct*^{3} or *conjunct*^{4} (strongly connective), *conjunct*^{2} (moderately connective), *conjunct*^{1} (mildly connective), and *conjunct*^{0} (neutral). Similarly, the terms *antijunct*^{0} through *antijunct*^{4} are used to indicate increasingly disconnective effects; *conjunct*^{0} and *antijunct*^{0} are the same.

There is a complementary vocabulary of primitive elements for the hierarchic view. The stylistic effects of syntactic components are correlated with the degree of subordination or superordination;

[2]The material in this section appeared previously in (Hoyt and DiMarco 1994).

[3]Discord, in our view, is not necessarily 'bad'. Indeed, it is the strategic use of discord, deviation from the norm, that can give expressiveness to writing.

[4]These two complementary kinds of analysis are implicit in the work of most stylists and rhetoricians.

the classifications are analogous to the connective: *subjunct*^{4} through *subjunct*^{0} (decreasingly subordinate) and *superjunct*^{0} through *superjunct*^{4} (increasingly superordinate); *subjunct*^{0} and *superjunct*^{0} are the same.

We adapted the work of Halliday and Hasan (1976) on cohesive relations to assign classifications to the connective elements. Halliday and Hasan consider substitution, including ellipsis, to be the most strictly cohesive relation, followed by reference, and then conjunction. We adopted this ranking, and so we classify intrasentential substitution and ellipsis as strongly connective (*conjunct*^{3}), reference as moderately connective (*conjunct*^{2}), and conjunction as mildly connective (*conjunct*^{1}). We also classify interpolation, parenthetical constructions, as disconnective (*antijunct*^{2}).

In assigning a hierarchic classification to a syntactic component, we adapted Halliday's (1985) work on *subordination*, specifically, embedding and hypotaxis, and the definition of the term *superordination* by Quirk *et al.* (1985). We classify embeddings as strongly subordinate, *subjunct*^{3}, and hypotactic structures as only mildly subordinate, *subjunct*^{1}.

3.3 Abstract elements of style

The primitive elements of style are combined into patterns of *abstract elements* that describe general stylistic properties related to syntactic parallelism, structure nesting, and linear ordering. The abstract elements are defined as follows:

Homopoise: A sentence with interclausal coordination of syntactically similar components.

Heteropoise: A sentence in which one or more parenthetical components are syntactically 'detached' and dissimilar from the other components at the same level in the parse tree.[5]

Monoschematic: A sentence with a single main clause with simple phrasal subordination and no accompanying subordinate or coordinate clauses.

Centroschematic: A sentence with a central, dominant clause with one or more of the following optional features: complex phrasal subordination, initial dependent clauses, terminal dependent clauses.

Polyschematic: A sentence with more than one central, dominant clause and at least one dependent clause.

Resolution: A shift in stylistic effect that occurs at the end of a sentence and is a move from a relative discord to a stylistic concord.

Dissolution: A shift in stylistic effect that occurs at the end of a sentence and is a move from a relative concord to a stylistic discord.

The remaining abstract elements describe *concordant* or *discordant* stylistic effects in particular positions. The basic elements are *initial concord*, *medial concord*, and *final concord*, with a similar range of *discord* elements.

3.4 Stylistic goals

As we have noted, the abstract elements are defined in terms of the lower-level primitive elements. The abstract elements are in turn used as the basis for the definition of higher-level *stylistic goals*. Stylistic goals can be organized along orthogonal dimensions. For example, a writer might try to be clear, or obscure, or make no effort either way. *Clarity* and *obscurity* are thus opposite ends of a stylistic dimension. Likewise, the goals of *concreteness* and *abstraction* form a dimension, and so do *staticness* and *dynamism*.

We adapted descriptions of stylistic goals from textbooks of style and rewrote these descriptions in terms of our abstract elements. Clarity, for example, is characterized by *simple* monoschematic sentences, *centred* centroschematic sentences, and *parallel* homopoisal sentences. Concreteness is associated with sentences that *highlight* a particular component: these are our heteropoises and discords. And *staticness* is characteristic of *'fixed-form'* sentences in which there is little stylistic variation, that is, monoschematic or homopoisal sentences.

4 ASSET—A stylistic analyzer

The theory of style has been applied in a straightforward manner to computerized stylistic analysis. Hoyt (1993) implemented such an analyzer in Prolog, called ASSET, for certain simple parts of the stylistic grammar; it has now been extended to cover the entire grammar. Briefly, the process of stylistic analysis involves accepting an input sentence, parsing it to obtain the syntactic structure, marking each part of the sentence with the appropriate primitive elements, deriving the abstract elements for the sentence from these, and finally, determining the stylistic goals that the sentence achieves. It should be

[5] A heteropoise can be *initial*, *medial*, or *final*, depending on the position of the parenthesis in the sentence.

noted that the theory of style does not imply this particular sequence of events; for instance, stylistic analysis could theoretically happen in tandem with parsing, rather then after it. However, for the sake of modularity, the two are separated in this implementation.

4.1 The Pundit parser

The first step in stylistic analysis is obtaining a parse tree for the input sentence. Hoyt chose the Pundit parser (PUNDIT REF) for this task. Pundit is a large natural language parser based on Sager's (1981) Linguistic String grammar. This grammar is represented as a series of context-free (Backus-Naur Form, or BNF) rules along with a set of restrictions that are responsible for enforcing context-sensitive properties of language, such as agreement. Pundit covers a fairly broad range of syntactic constructions, including many that are stylistically interesting.

Pundit is capable of performing semantic analysis on sentences in addition to syntactic parsing, but this ability is not needed in ASSET. By modifying the Pundit code slightly, it is possible to capture just a syntactic parse of a sentence for use in stylistic analysis.

4.2 The transformation module

ASSET is designed to be independent of the parser used. Therefore, the stylistic analysis itself is performed without regard for the source of input. However, because the actual implementation is based on the Pundit parser, it requires a 'transformation module' to transform the output of the parser into the form of input required by ASSET. This is actually the largest part of the implementation because there is no simple correspondence between the output of Pundit and the terms of the stylistic grammar. During transformation, each grammatical category produced by Pundit must be translated into a construction 'known' by ASSET, eventually producing an equivalent parse tree that conforms to the grammar of style.

4.3 The annotation module

Once the parse tree has been transformed into a form to which the stylistic grammar can be directly applied, ASSET's 'annotation module' labels each node of the tree with stylistic information. Here, Hoyt introduced a new level to the grammar of style: the transition elements. Transition elements were introduced primarily for ease of implementation, and are not considered to be a change in the theory of the grammar. They have the same names as the abstract elements discussed above, but are applied to nodes in the parse tree, not to the sentence as a whole. For example, the transition element 'centro' is assigned to phrases that may be part of a centroschematic sentence. The process of annotation thus involves the traversal of the parse tree from leaf to root, with each node being assigned primitive and/or transition elements as applicable, based on the syntactic class of the node and the analysis of its children. After this annotation, the abstract elements for the sentence are determined from the transition elements at the highest level, and the stylistic goals in turn from these abstract elements (Figure 1).

4.4 The Significance of ASSET

ASSET is important to the theory of syntactic style in two respects: First, it demonstrates that this approach is theoretically and computationally feasible. The straightforward implementation of the theory shows that it is not impractical to deal with subtelties such as syntactic style in natural language processing systems. Second, ASSET provides a means to test the theory of style. By using the analyzer on sentences from various sources, weaknesses and strengths of the stylistic grammar can be discovered. Resulting changes in the theory can be easily incorporated into ASSET for further verification.

5 An approach to generating stylistically customized text

We are now extending our work on stylistic analysis to the generation of text that conforms to given pragmatic constraints. Our particular application is the automated generation of hospital patient educational materials, customized to the patient's age, reading level, culture, and medical condition. We are particularly interested in how the linguistic style of the texts interacts with persuasive effects; for example, we would expect that anti-smoking material aimed at adolescent males would differ from the kind intended for pregnant women, and, within each group, other personality and demographic characteristics will influence the nature of an effective persuasive message.

Instead of following the usual generation paradigm of trying to generate a complete customized document from scratch, we are starting from a 'master' document that contains all the material relevant for a given type of medical therapy, for all variations in patient condition. We will convert the master document into the sentence-planning representation language (SPL) used by Penman (CITE), the generation system we are working with. The SPL relevant to a particular patient's set of medical variables will be selected out of the master document—what will remain will be semantically complete, but possibly incoherent, lacking in cohesion, and stylistically awkward. From this point on, separate text planning, sentence planning, and lexical choice 'repair engines' will work on the SPL in turn, producing, through a 'waterfall' process, a gradually modified SPL that will generate increasingly better text. This is a novel approach to the problems of generation, and we anticipate that we may be able to finesse many of the currently intractable problems, while being able to produce high-quality texts.

We propose to develop systems for knowledge-based document generation that would provide support to a user in setting up the document, in maintaining the coherence of the text, and in conveying the appropriate stylistic effect for a given situation. More generally, there is a great need for computer support for technical documentation and writing, which would be supported by tools such as those we will generate. There would be enormous value to any organization in report-generation systems simple and reliable enough to be embedded in user interfaces; even partial solutions would be helpful. The world market for documentation management and technical writing is large and continuing to grow.

References

DiMarco, Chrysanne (1990). *Computational stylistics for natural language translation.* Doctoral dissertation, Department of Computer Science, University of Toronto [published as technical report CSRI-239].

DiMarco, Chrysanne and Hirst, Graeme (1993). "A computational theory of goal-directed style in syntax." *Computational Linguistics*, **19**(3), September 1993, 451–459.

DiMarco, Chrysanne and Mah, Keith (1994). "A model of comparative stylistics for machine translation." *Machine translation*, **9**(1), 1994, 21–59.

Green, Stephen J. (1992). *A functional theory of style for natural language generation.* Master's thesis, Department of Computer Science, University of Waterloo, 1992. [University of Waterloo Faculty of Mathematics Technical Report CS-92-48].

Green, Stephen J. and DiMarco, Chrysanne (to appear). "Stylistic decision making in natural language generation." Originally appeared in: *Proceedings, Fourth European Workshop on Natural Language Generation*, Pisa, 1993, 155–158. Extended version (22 pages) accepted Summer 1994 to appear in a book of selected papers from the workshop to be published by Springer-Verlag.

Halliday, M.A.K. (1985). *Introduction to functional grammar.* Edward Arnold, London, 1985.

Halliday, M.A.K. and Hasan, Ruqaiya (1976). *Cohesion in English.* Longman Group Limited, London, 1976.

Hirst, Graeme (1995). "Near-synonymy and the structure of lexical knowledge." *AAAI Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity*, Stanford University, March 1995. [to appear]

Hoyt, Patricia A. (1993). *A goal-directed functionally-based stylistic analyzer.* Master's thesis, Department of Computer Science, University of Waterloo, 1993. [University of Waterloo Faculty of Mathematics Technical Report CS-93-48].

Hoyt, Pat and DiMarco, Chrysanne (1994). "A goal-directed multi-level stylistic analyzer." *Proceedings, 10th Canadian Conference on Artificial Intelligence*, Banff, May 1994, 23–30.

Makuta-Giluk, Marzena H. and DiMarco, Chrysanne (1993). "A computational formalism for syntactic aspects of rhetoric." *Proceedings, The First Conference of the Pacific Association for Computational Linguistics*, Vancouver, April 1993, 63–72.

Miezitis, Mara (1988). *Generating lexical options by matching in a knowledge base*, MSc thesis, Department of Computer Science, University of Toronto, October 1988. [Published as technical report CSRI-217.]

The Penman natural language generation group (1988). *The Penman documentation.* Information Sciences Institute/USC, 1988.

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey, and Svartvik, Jan (1985). *A comprehensive grammar of the English language.* Longman Group Limited, 1985.

Sager, N. (1981). *Natural Language Information Processing: A Computer Grammar of English and Its Applications.* Addison-Wesley.

Unisys Corporation (1989). *PUNDIT User's Guide.* Unisys Corporation.

```
Style Goals:  clarity
              concreteness


Abstract Elements:  heteropoise
                    centroschematic
                    initial-concord
                    medial-concord
                    final-concord
                    initial-discord
                    resolution


The Stylistic Parse Tree:
-------------------------
complete([centro,poly,hetero,concord, final-concord,medial-concord,
          initial-concord,initial-discord])
   dependent-clause([concord,initial-discord])

     non-finite-clause([initial-discord,concord])

       noun-phrase([no-subject],[no-subject], [no-subject-concord])
         np(wh-word)

        verb-phrase([mono,concord])
          verb([conjunct0],[subjunct0],[no-style])
            lexical-verb(WHISLTING)
          complements([concord,mono])
            complement-1([no-complement],[no-complement], [concord,mono])
              lex-complement(none)

          adverbial([conjunct1],[subjunct0],[centro])
            adjunct(SOFTLY)

   major([centro,concord,mono])

      noun-phrase([centro,concord,mono])
        nominal-group([mono,concord,centro])
          premodification([conjunct0],[subjunct0], [centro,concord,mono])
            premod(none)
          noun([conjunct3],[subjunct0],[centro,concord,no-style])
            pronoun(HE)
          postmodification([conjunct0],[subjunct0],[centro,concord,mono])
            postmod(none)

      verb-phrase([mono,concord,centro])
        verb([conjunct0],[subjunct0],[no-style])
          lexical-verb(WALKED)
        complements([centro,concord,mono])
          complement-1([mono,concord,centro])
            prepositional-phrase([centro,concord,mono])
              preposition([no-conn-style],[no-hier-style],[no-style])
                lex-preposition(ALONG)
              nominal-group([mono,concord,centro])
                premodification([conjunct1],[subjunct2],[centro,concord,mono])
                  adjectival([conjunct1],[subjunct2])
                    definite-article(THE)
                noun([conjunct0],[subjunct0],[no-style])
                  lexical-noun(HALLWAY)
                postmodification([conjunct0],[subjunct0],[centro,concord,mono])
                  postmod(none)
```

Figure 1: Sample ASSET parse for the sentence “Whistling softly, he walked along the hallway.”