Toward a Multidimensional Framework to Guide the Automated Generation of Text Types

Eduard Hovy

Information Sciences Institute
of the University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292-6695
U.S.A.
tel: +1-310-822-1511 x 731
fax: +1-310-823-6714
email: hovy@isi.edu

Julia Lavid

Departamento de Filología Inglesa
Facultad de Filología
Universidad Complutense de Madrid
280040 Madrid
Spain
tel: +34-1-394-5862
fax: +34-1-394-5396
email: lavid@dit.upm.es

1 Introduction

A central concern limiting the sophistication of text generation systems today is the ability to make appropriate choices given the bewildering number of options present during the planning and realisation processes. As illustrated in several systems [Hovy 88, Bateman & Paris 89, Paris 93], the same core communication may be realised in numerous different ways, depending (among other factors) on the nature and relation of the interlocutors, the context of the communication, the media employed, etc. The combinatoric number of possibilities of all such factors is extremely large. Since most of them are not well understood at this time, automated text generation may appear to be a hopeless endeavour.

Fortunately, the picture is not altogether bleak. Given that certain types of communicative situations consistently give rise to characteristic recognisable genres or text types, one can attempt to characterise each genre or text type in terms of the set of generator decisions or rules responsible for producing those characteristics, and then create prespecified, genre-specific, collections of features, formulated as decision rule criteria, for subsequent use (this point has been made before, in [Patten 88] and [Bateman & Paris 89]). With this aim in mind, two major questions arise:

1. Is there a regular categorisation of genres or text types?
2. How can one most easily determine the genre-determining features for given texts?

In this paper we address both questions. First we report on work developing a functionally motivated framework to provide a matrix for the description, comparison, and classification of a body of texts. This framework can act as the background for research on discourse phenomena, text planning, and realisation, and can enable groups working with different texts to relativise their results in terms of the matrix. The approach involves a systematic search for correlations between linguistic form and function in discourse, a discovery of the relation between meaning and wordings that accounts for the organization of linguistic features in each text type. This task cannot be fully performed without linking the functions of particular linguistic features to variation in the communicative situation, since, as users and receivers of language, people

[0]The first author was supported by ARPA Contract MDA-904-91-C-5224. The second author's portion is based on Deliverable R1.1.1a for DANDELION (ESPRIT Basic Research Project 6665).

produce texts whose communicative function has to be interpreted in terms of the concrete situation in which they were produced. The knowledge of the meaning potential associated with a generic situation is called *register.*

Register has been the subject of much research in Linguistics [Ferguson 83, Brown & Fraser 79, Hymes 74], especially in Systemic-Functional Linguistics [Halliday & Hasan 89, Ure 71, Gregory 88, Ghadessy 88, Cafferel 1991], etc. Within SFL, various perspectives have been taken: Halliday views register from the lexicogrammatical perspective, while [Martin 92] sees it operating at the semiotic level. With a phenomenon as complicated as register, it is inevitable that conflicting pictures exist; however, in this paper we do not devote too much time to any specific view, but rather take a slightly more general approach to make our points relevant to all. We view registers simply as stable configurations of features at all levels — semiotic, grammatical, lexical, phonological — linked together. In the first part of the paper, then, we outline several high-level and somewhat more general than usually provided register networks, drawn from a variety of sources and organized according to communicative metafunction.

With regard to the second half of the paper, we describe a semi-automatic method to determine genre-defining features for a given text, and show how the degree of genre-specificity can be measured quantitatively. This follows on register-oriented work in computational research on language generation, in particular that of [Patten 88, Bateman & Paris 89, Bateman & Paris 91]. Our work in some ways follows upon that of Bateman and Paris, who outline an ambitious 5-step method for the definition of register and the control of a generator program, using three variations of a sentence as illustration: 1. text analysis; 2. classification of features according to user; 3. classification of features with respect to register type; 4. creation of register networks; and 5. specification of generator control. We take a less ambitious and somewhat different approach to some of the same issues (steps 1, 3, and 4), and develop a semi-automated feature collection technique using as illustration 10 clauses from the instruction stage of a recipe. The contribution of this paper is twofold:

1. somewhat more high-level and comprehensive register networks, drawn from several sources and organized according to communicative metafunction (in contrast to steps 3 and 4);

2. a semi-automated abductive method for identifying grammatical features that are register-defining (in contrast to step 1).

2 The components of the communicative situation

According to Halliday, language performs three principal functions simultaneously: the *ideational function* (to understand the interlocutors' physical, mental, and emotional environment), the *interpersonal function* (to act on other people in it); and the *textual function* (to employ the media and situation at hand for optimal communication) [Halliday 85]. In a each instantiated communication, the speaker performs a series of linguistic choices from these three metafunctions of language: in Systemic terms, he or she selects features from language-based system networks assigned to the three different functions.

The communicative situation — topic, interlocutors, context, etc. — is closely correlated with and helps determine the configuration of meanings selected from these three functional components of language. Given this correlation, each particular communicative situation is partitioned into three regions corresponding to the linguistic ones: the experiential meanings of the text reflect the FIELD, the interpersonal meanings reflect the TENOR, and the textual

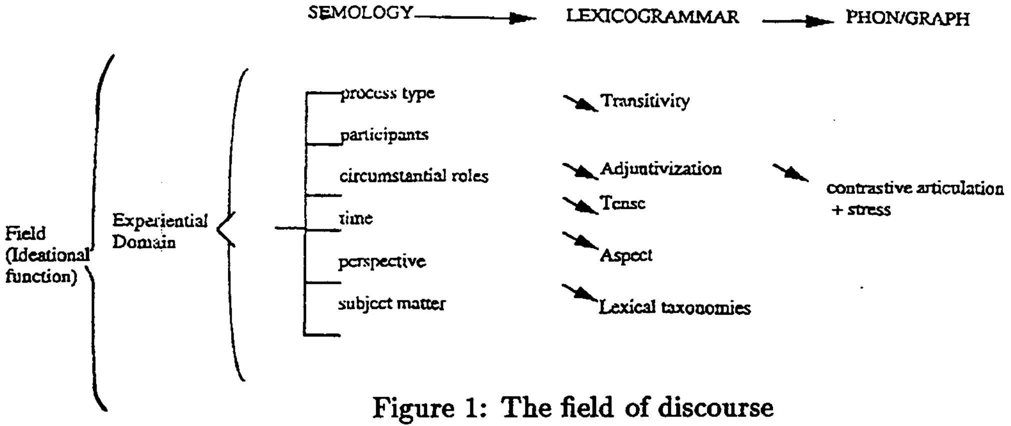

Figure 1: The field of discourse

meanings reflect the MODE of the discourse. We can say that field, tenor, and mode are the actual selections (from the ideational, the interpersonal, and the textual components of the language code respectively) taken in a particular event surrounding and including the language act.

In the remainder of this section we briefly describe the three aspects of communication. More details are provided in the longer version of this paper, available from the authors.

The field of discourse. According to Halliday, the field of discourse refers to "what is happening, to the nature of the social action that is taking place: what is it that the participants are engaged in, in which the language figures as an essential component" [Halliday & Hasan 89]. The field of discourse can also be called the text's experiential domain which includes the text's subject matter, that is, its ideational or propositional content. The network in Figure 1 illustrates these aspects.

The tenor of discourse. Where field predicts the range of meaning potentials in the experiential component of the language code, the tenor of discourse predicts the selection of options in the interpersonal component. According to Halliday and Hasan, "the tenor of discourse refers to who is taking part, to the nature of the participants, their statuses and roles: what kinds of role relationships obtain among the participants..., both the types of SPEECH ROLE that they are taking on in the dialogue and the whole cluster of socially significant relationships in which they are involved" ([Halliday & Hasan 89], p. 27). The tenor of discourse involves the selection of a number of options in the subsystems that configure the participants' speech roles. Among these speech roles we distinguish two principal types: one set of systems is concerned with the NEGOTIATION OF SPEECH ROLES, the other is concerned with the SPEECH MODALITIES. Figure 2 contains some of these options in a systemic network.

The mode of discourse. The mode of discourse has traditionally been seen as composed of selections from three simultaneous parameters: the LANGUAGE ROLE, the MEDIUM, and the CHANNEL OF DISCOURSE. The LANGUAGE ROLE is a continuum with the two ends of the scale being whether the language is constitutive or ancillary (the language in a face-to-face service encounter being ancillary since it accompanies an activity and is not the sole meaningful activity, and the language of a physics research paper being constitutive since the text creates the entire exchange). The MEDIUM OF DISCOURSE deals with the process of text creation, with the degree of sharing the process of text creation between the interlocutors. The CHANNEL OF DISCOURSE is the modality through which the language is received, including typically the options GRAPHIC and PHONIC. Early work on register (e.g., [Gregory & Carroll 78]) often glossed medium as being congruent with the option between speaking and writing, but we can now go further

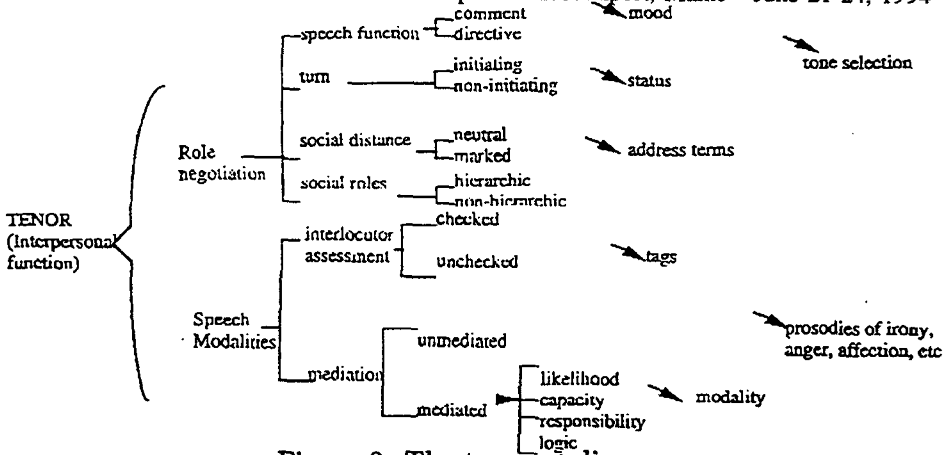

Figure 2: The tenor of discourse

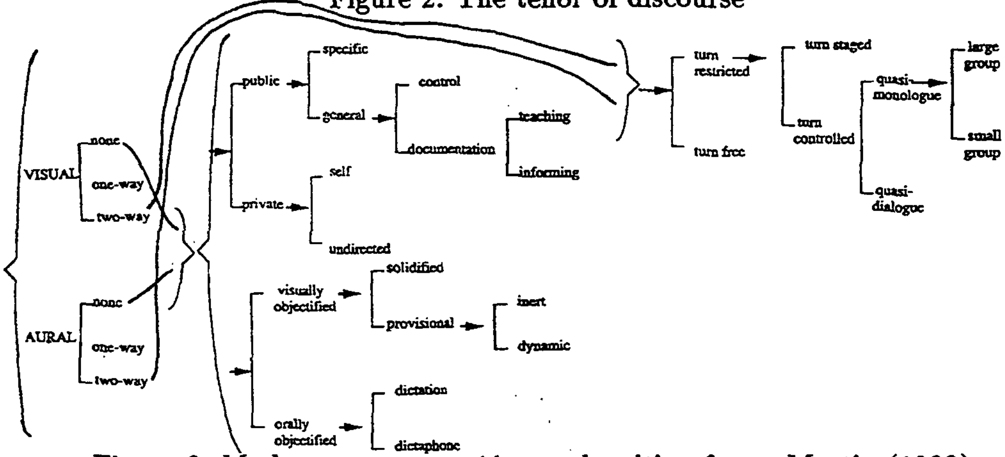

Figure 3: Mode systems: speaking and writing focus; Martin (1992)

and adopt more abstract characterizations as suggested by [Martin 92]. This is also necessary given the range of substantial empirical work (e.g., [Redeker 84, Biber 89] and others) showing that the spoken/written distinction *per se* is not a simple parameter. The lexicogrammatical consequences of the features shown in Figure 3 are discussed in [Martin 92].

3 Using the multidimensional analysis of texts for generation

As discussed in [Matthiessen 94], register can be interpreted (and therefore implemented in a sentence generator) in three ways:

- Probability variations of choices within systems: Each register imposes its idiosyncratic probability distribution upon the choice preferences within appropriate systems, so that while the grammar remains the same throughout, the generator's traversal of the grammar will vary according to registerial probabilities;
- Core system with extensions for variation: Each register adds some idiosyncratic systems at appropriate points of the grammar while leaving the remainder unchanged;
- Completely separated system networks: Each register has a distinct subgrammar, and no common core exists. This is the approach taken in [Patten 88, Bateman & Paris 91]. In this sense, register-specific language is treated like a sublanguage [Kittredge & Lehrberger 81].

We follow the first approach. In this section, we outline a method of semi-automatically determining probability distributions for each register, taking as example the instruction stage of a recipe:

> *Remove fruit and 2tbs of juice from the can, then discard the rest. Put all ingredients into a saucepan and slowly bring to the boil. When hot, pour into a food processor and process to a smooth sauce. For extra texture reserve 1–2 pieces of fruit, mash, then add this to the finished sauce.* (*SHE Magazine*, June 1993)

What are the lexicogrammatical features that express the features of field, tenor, and mode? For fully worked out systems, tracing them through the labyrinthine networks is tedious at best. For partially worked out systems, the connections between the higher level networks such as field and the lower level networks of the grammar often do not exist, and so another method is required for determining the registerially determinating features at the lower levels.

One such method, suggested in [Bateman & Paris 91], is to perform grammatical (and presumably lexical) analyses of sample texts by hand. While (as they nicely illustrate) this is possible for small samples, the problem of ensuring coverage and consistency for larger samples can quickly become daunting. For this reason, we propose a "bottom-up" abductive method, using the generator as a tool, that is considerably easier, since it is semi-automatic. The method involves the following steps:

1. For each sentence in the sample text type under consideration, create an input specification for the generator.
2. Run the generator on each input specification and check that the output sentences are correct. Collect the lexicogrammatical features for each sentence.
3. Classify the features for each sentence according to register type (field, tenor, or mode) and constituent type (clause complex, clause, noun phrase, lexical, etc.).
4. Count the number of times each feature appears in the whole test sample as a percentage of the number of times its constituent type appeared. For example, if the NP feature DETERMINED appears 9 times for 10 noun phrases in a sample, then we say the *involvement* of this feature is 90%. Graph or tabulate the distribution of feature involvement as *number of features* vs. *percentile.*
5. Through inspection of the resulting table, determine the register-determinate cutoff point — the point after which features appear too seldom to be indicative of the text type. This point will appear at the 'knee' at which the curve begins to rise rapidly for small increases of involvement.

We use the sentence generator Penman to generate the sentences in the sample text we selected, and collected the features it needed. The total number of features (including duplication) came to 543. Of these, 48 features appeared every time they could (i.e., were present every time a syntactic constituent of the appropriate type was generated: 10 at the clause complex level, 19 at the clause level, and 19 at the NP level). That is, 48 features had an involvement of 100%. We then graphed out the distribution of feature involvements. Notwithstanding the small sample size, we found a striking regularity: the involvement distribution was bimodal, with some features appearing very often (over 80%) and almost all the remainder appearing infrequently (under 30%, for the clause and NP levels, and under 60% for the clause complex level). That is, the middle range between 80% and 30% involvement contained significantly fewer features than either of the extremes. This we interpret as follows: when features appear often, they appear *very* often, and thus specify the genre characteristics. On the other hand, if features do not appear often, they appear seldom, only as needed to produce the particular clause(s) in which they appear. The degree to which features with high involvements appear can be thought of as the degree to which they co-specify the genre, and thus the "strength" of their propensity for selection during the text and sentence planning processes.

The following tables summarize (full information appears in the long version of this paper).

	Clause-complex level		**Clause level**		**NP level**	
% feature involvement	number of features	% of total features	number of features	% of total features	number of features	% of total features
100%	10	62.5%	19	15.4%	19	21.3%
>=80%	10	62.5%	34	27.6%	43	48.3%
mid-range	6	37.5%	24	19.5%	12	13.5%
<=30%	0	0%	65	52.8%	34	38.2%

A look at the genre-defining clause level features may prove instructive; as expected from looking at the text, features such as `IMPERATIVE`, `IMPERATIVE-INTERACTANT`, and `NONFINITIVE-VOICE` appear frequently:

```
19 at 100%
  START CLAUSES CLAUSE FULL MOOD-UNIT NONCONJUNCTED NO-WH-SUBJECT POSITIVE
  TRANSITIVITY-UNIT NONACCOMPANIMENT NONMATTER NONROLE NO-SPATIAL-EXTENT
  NO-TEMPORAL-EXTENT NO-TEMPORAL-LOCATION ACTIVE-PROCESS NOT-PHASE
  VOICE-LEXVERB LEXICAL-VERB-TERM-RESOLUTION

15 at 90%
  CLAUSE-SIMPLEX INDEPENDENT-CLAUSE INDEPENDENT-CLAUSE-SIMPLEX JUSSIVE
  NONINTERNAL-SUBJECT-MATTER IMPERATIVE IMPERATIVE-INTERACTANT MATERIAL
  IMPERATIVE-SUBJECT-IMPLICIT UNMARKED-POSITIVE DO-NEEDING-VERBS
  NONFINITIVE-VOICE NONCAUSE NONMANNER IMPERATIVE-UNTAGGED

 0 at 80%
```

4 Conclusion

The abductive method for text characterization presented here has several advantages, in our opinion. An important advantage is that it focuses human effort not on text analysis (which is difficult and prone to error and inconsistency) but rather on generator input creation (which can easily be checked). Also, the graphed distribution of feature involvements provides an immediate visual clue as to which features are indeed register-determinate and to what degree they are so. In turn, this allows the register-grammarian to express grammar decision rules (or system network options, in the case of SFL) in terms of probabilities with some empirical confidence. Another benefit is that the method assists with text type characterisation, by pointing out (through dramatically lower involvement values) when different text types or stages are mixed.

References

[Bateman & Paris 89] Bateman, J.A. and Paris, C.L. (1989) Phrasing a Text in Terms the User Can Understand. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence IJCAI-89*, Detroit, Michigan.

[Bateman & Paris 91] Bateman, J.A. and Paris, C.L. (1991) Constraining the development of lexicogrammatical resources during text generation: towards a computational instantiation of register theory. In Ventola, E. (ed), *Recent Systemic and Other Views on Language*, Amsterdam: Mouton.

[Biber 89] Biber, D. (1989) *A typology of English texts. Linguistics* 27 (3–43).

[Brown & Fraser 79] Brown, P. and Fraser, C. (1979) Speech as a marker of situation. In Scherer, K.A. and Giles, H. (eds), *Social markers in Speech* (33–62). Cambridge: Cambridge University Press.

[Cafferel 1991] Cafferel, A. (1991) Context Projected onto Semantics and Consequences for Grammatical Selection. Written version of paper presented at International Systemics Congress XVIII, International Christian University, Tokyo.

[Ferguson 83] Ferguson, C. A. (1983) Sports announcer talk: Syntactic aspects of register variation. In *Language in Society* 12 (153–172).

[Gregory & Carroll 78] Gregory, M. and Carrol, S. (1978) *Language and Situation: Language varieties and their social contexts.* London: Routledge and Kegan Paul.

[Gregory 88] Gregory, M. (1988) Generic Situation and Register: a Functional View of Communication. In Benson, J.D., Cummings, M.J. and Greaves, W. (eds) *Linguistics in a Systemic Perspective.* London: John Benjamins Publishing Company.

[Ghadessy 88] Ghadessy, M. (1988) *Registers of Written English: Situational Factors and Linguistic Features.* London: Frances Pinter.

[Halliday 85] Halliday, M.A.K. (1985) *An Introduction to Functional Grammar.* London: Edward Arnold.

[Halliday & Hasan 89] Halliday, M.A.K. and Hasan, R. (1989) *Language, Context and Text: A Social Semiotic Perspective.* Oxford: Oxford University Press.

[Hovy 88] Hovy, E.H. (1988). *Generating Natural Language under Pragmatic Constraints.* Hillsdale, N.J.: Lawrence Erlbaum Associates Publishers.

[Hymes 74] Hymes, D.H. (1974) *Foundations in Sociolinguistics.* Philadelphia: University of Pennsylvania Press.

[Kittredge & Lehrberger 81] Kittredge, R. & Lehrberger, J. (1981) *Sublanguage: Studies of Language in Restricted Semantic Domains.* Berlin: De Gruyter.

[Martin 92] Martin, J.R. (1992) *English Text: System and Structure.* Amsterdam: Benjamins.

[Matthiessen 94] Matthiessen, C.M.I.M. (1994) Register in the Round. In *Register Analysis: Theory and Practise*, Ghadessy, M. (ed), London: Pinter (221–292).

[Paris 93] Paris, C.L. 1993. *The Use of Explicit Models in Text Generation.* London: Francis Pinter.

[Patten 88] Patten, T. (1988) *Systemic Text Generation as Problem Solving.* Cambridge: Cambridge University Press.

[Redeker 84] Redeker, G. (1984) On differences between spoken and written language. In *Discourse Processes* 7 (43–55).

[Ure 71] Ure, J. (1971) Lexical density and register differentiation. In Perren, J.L. and Trim, J.L.M. (eds) *Applications of Linguistics: Selected Papers of the 2nd International Congress of Linguistics.* Cambridge: Cambridge University Press.