Jonathan J. Webster

THE 'PROLEX' PROJECT

Introduction

Socioculturally defined groups typically possess a distinctive way of communicating about their specific interests; business and management professionals are no exception. The PRO(fessional) LEX(is) Project at Singapore has as one of its primary objectives the description of those usage rules governing the appropriate use of a given profession's technical vocabulary and style of expression. In other words, we are endeavouring to discover the system of rules underlying the sociolinguistic competence of members of a socially-defined group; in this instance, as defined by occupation. In keeping with Adam Makkai's proposal (1980:125) for an Associative Lexicon for 20th-Century English, we will attempt "to do justice to what there is in human sociopsychological reality" by organizing our representation of professional lexis by frequency, collocational range, semantic nesting and socio-linguistically significant typological categories. Subsequent to text collection and computer-based data analysis, we propose to develop the PROLEX files into a software tool for the professional capable of being incorporated onto a microcomputer with applications for text processing and computer-assisted learning.

Prior to the collection of texts to be included in our analysis of professional lexis, we must first identify those variables by which we hope to define a truly representative corpus. Wölck (1976) suggests incorporating a brief, exploratory type study of the targeted community into the initial stages of one's project. Such a study should reveal sufficient 'anthropological data' to permit the selection of a text-based sample which adequately characterizes the dimensions of diversity within the professional community under study. In this case, our focus will be on the occupational lexis of the business—management professional as it would be called for in letters, reports, summaries, memoranda, notices, journal pieces, magazine articles, etc. Once we have determined the size and range of our text-based sample, we may then proceed in our analysis of texts using the Oxford Concordance Program, a general-purpose computer program capable of making frequency counts, constructing concordances, and testing for the collocational range of targeted lexical items.

Towards a typology of business English

In his preface to the book The Manager and His Words (Gillibrand and Maddock 1982), Ross Stainton, former Chairman of British Airways, writes: "Perhaps in the business world as much as in technical and semi-technical matters, jargon has crept in and normal usage has been distorted, so that the businessman visitor to Britain, or other English speaking countries, may well need some coaching in this aspect of language." The LONGMAN DICTIONARY OF BUSINESS ENGLISH is one recent attempt at system-

atically describing "the language of commerce and economics, textbooks, newspapers, journals, company reports, and commercial law." From Macmillan publishers comes the INTERNATIONAL DICTIONARY OF MANAGEMENT containing entries such as E_1.

E_1: short account ... (1) Account of someone dealing in short sales.(2) Total short sales of particular commodity, security or share on the open market.

As Makkai notes, "Dictionaries, by and large, have tended to ignore the associative groupings of lexemes as they form NATURAL SEMANTIC NESTS around concretely observable and abstract (non-observable entities, and their traditional reliance on alphabetization endeavored presenting a totality of the available lexis while ignoring frequency of usage, exact range of dialectal habitat, the speaker's sociological status, etc." (1980:127). Dictionaries of Business English likewise submit to this "psychologically quite unmotivated tyranny of the alphabet" (1980:127), leaving their users uninformed as to a term's typologically-based frequency count and collocational range. By 'typologically-based' we mean that our analysis of each term should be informed as to the 'type' of text within which use of the term has been recorded. Of course, this presupposes some prior classification of texts by whatever criteria can be shown to be sociolinguistically significant. To illustrate further what we mean by typological information, note the following excerpts from a "Glossary of Commercial Terms" compiled by Duncan Macintosh for his textbook English for Business.

E_2: after sight ... Used in conjunction with bills of exchange or "drafts" to indicate ...

E_3: without engagement ... A term sometimes used by merchants when quoting the price of certain articles liable to sudden fluctuations to indicate ...

Typological information tells us WHO uses the term, to WHOM they address the term, WHERE they use the term, WHEN they use the term and in WHAT area of business and management activity the users participate.

As for WHERE they use the terms, we would most likely set up a category for business letters within which would be included letters of enquiry, letters placing orders, letters of adjustment, letters requesting payment, letters promoting goodwill, sales letters, etc. (cf. Gartside 1980). This category of business letters should itself then be classed under the broader heading of 'business correspondence' and documents along with memoranda, notices, reports, invoices, telexes, etc.

If what we wish to describe, however, is not only the language used to conduct business, but also the language used to talk about business, then we should also include samples selected from The Wall Street Journal, Harvard Business Review, Business Week, as well as any other text — whether book, journal, magazine or newspaper — the selection of which will afford a more complete and accurate representation of available publications. Initially, text samples will be classified according to topic — accounting,

advertising, banking, commerce, etc. - and also by type of pub-
lication - intended regional coverage, expected readership
sophistication, etc.

The Oxford Concordance Program (OCP) permits the user to
render a computer version of the text corpus fully referenced
to indicate those features by which each sample, or portion
therefrom, is typologically classified. We commit for processing
only those samples identified as belonging to that reference
category targeted for analysis. OCP further allows the user
to generate, sort, and print the reference in an index or con-
cordance.

To illustrate, we propose the following information categories
for the 'letter of enquiry':

T   Type of correspondence (e.g. Letter of enquiry)
F   Function (e.g. Requesting credit)
D   Date
S   Salutation
P   Paragraph
C   Closure

According to Makkai, "Lexicography urgently needs a syntactic-
ally sophisticated, computer-processed frequency count of English
lexis with systematic indications given of the social and geo-
graphical source of the form" (1980:128,129). We propose to
differentiate between the multiple syntactic functions of the same
lexical items. Lexical items are coded as to their syntactic
function using letters declared as diacritics which are attached
as parts of the words, but are not considered at the sorting
phase until the words can be distinguished in no other way.

Thus, if the word cash were to function syntactically as noun,
as in the sentence "They are low on cash", we would attach the
appropriate diacritic cash1 whereas, if the same word were to
function as a verb as in the phrase "to cash a check" we would
enter it as cash2. Coded in this way, we may now obtain a more
accurate frequency count per lexeme.

Short collocates with account in "short account", with bills
in "short bills", with shipment in "short shipment": like those
set phrases Makkai calls 'lexemic idioms', we propose to treat
these institutionalized word combinations as single lexemes.

We aim to discover the collocational range of those lexemes
constituting the professional lexis under review, using the OCP DO
CONCORDANCE command which instructs the computer to produce a list
of words together with their references and contexts (OCP Users'
Manual, 5.3.1.3, 71). In addition, by means of the PICK COLLOCATES
command, we obtain a concordance by which we are fully informed
as to the actual contexts in which the collocates appear, the typo-
logcial distribution of source texts (as indicated by the
reference category labels cited alongside the contexts), and also
the frequency count per collocation.

Preferring a TRUE ('thesaurus rendered updated ecologically') account of semantic nestings to the traditional HABIT ('heterogeneous alphabetical inventory') dictionaries, Makkai proposes modelling his Associative Lexicon after Roget's THESAURUS "with a number of major modifications". Addressing himself to the question of how we can improve on Roget's semantic nestings, Makkai (1980:131, 132) states,

> I believe that the answer lies in the controls to be exercised over semantic nestings by frequency listings and by collocational range. The adjective round, for example, should be marked C(ommon), + UK (all regions), + Australia (all regions), + New Zealand (all regions), + Canada (all regions), - C(ollocational), R(ange): fruits, objects, balls, human body characteristics - followed by some examples such as round apple, round ball, round figure, etc.

We propose to structure the PROLEX entries similarly by semantic nestings informed firstly as to the typological distribution of lexemes, and secondly by frequency and collocational range.

Interfacing PROLEX entries with the discourse frame data-base

The PROLEX entries, once encoded, will be interfaced with the DF.DB, a data-base consisting of discourse frames modelled after Kenneth Pike's notion of grammatical 'tagmeme', "a constituent of a construction seen from the point of view of its four general features: slot, class, role and cohesion" (1979:33).

In order to facilitate implementation of a data-base containing all the information about grammatical constructions as represented in Pike and Pike (1979) by generative formulas and grammatical tagmemes, we have devised a mode of representation called 'discourse frames' which, following the suggestion of Winston and Horn (1981:292, 293), are represented as nested association lists. Winston and Horn represent a frame as an association list nested to a depth of four levels. If we picture an association list as a list of pairs, then the four levels would appear as follows:

    [FRAME.NAME [SLOT   [FACET   [VALUE]]]]

A frame can have any number of slots; slots can have any number of facets; facets can have any number of values. Borrowing the feature names from Pike and Pike, we have devised the following model for a discourse frame:

    [FRAME.NAME   [SLOT   [CONDITION   [.....]   ...]
                            [ROLE   [.....]   ...]
                            [CLASS   [.....]   ...]
                            [COHESION   [.....]   ...]]]

Given the names of the frame, slot, facet, and value as input, the function F.PUT constructs the discourse frames. By way of example, if we were setting up the discourse frame for LETTER. ENQUIRY for the first time, we would key into the computer:

    F.PUT "LETTER.ENQUIRY "PRE.MAR "CONDITION "OBLIGATORY

The result would be a new variable name, LETTER.ENQUIRY, with a value of [LETTER.ENQUIRY [PRE.MAR [CONDITION [OBLIGATORY]]]]. At the same time we will have stored the same structure on the property list of LETTER.ENQUIRY under the property FRAME. For an existing frame, we have the same function, F.PUT, to store additional SLOTs, FACETs, or VALUEs. Even once the frame is seemingly complete new information, whether it be a new slot, or new facet for an already existing slot, or even a new value for an already existing facet, we still need only apply the same function, F.PUT. On the other hand, if we wish to remove a particular slot or facet or value, we need only key in the appropriate function and appropriate input.

Similar to the generative formulas of Pike and Pike (1979) after which they were modelled, the discourse frames may be used to generate successively higher levels of discourse units from word to phrase to clause to sentence to paragraph to letter. Discourse frames include the range of possible variants for any given construction.

Incorporated into frames such as ADJ(ective), C(ount) N(oun), R(oo)T, T(i)M(e).M(ar)K(e)R, etc., are control slots which signal that they are endpoints, at which point the PROLEX files must be retrieved and a search made for the appropriate lexical filler.

The PROLEX System File combines both the PROLEX entries and the DF.DB to serve as a useful aid to the professional in generating business correspondence. Thus, if a user wished to write a letter of enquiry, he might first call up a menu indicating types of correspondence, and then proceed to interact with the machine, supplying the necessary information as requested by the computer to produce the desired product.

The PROLEX user should also be able to access individual lexical entries according to the manner in which they have been encoded, i.e. by typological distribution, semantic nesting, frequency, or collocational range. Furthermore, as Makkai points out, since "people are accustomed to looking up words in dictionaries in the HABITual way" (1980:127), users should also be able to look up entries alphabetically.

Conclusion

What I have sought to outline in this paper are the methodological means by which we expect to obtain a detailed description of professional lexis, specifically, Business English. Incorporating suggestions put forward by Makkai (1980:128-130), we have set for ourselves the following research objectives:

(1) to classify text samples according to those typological categories determined to be sociolinguistically significant;

(2) to provide a syntactically sophisticated, computer-processed frequency-count of professional lexis with systematic indications given of the social and geographical source of the form;

(3) to discover the collocational range of those lexemes con-
stituting the professional lexis under review;

(4) to obtain a TRUE ('thesaurus rendered updated eco-
logically') account of semantic nestings.

References

Gartside, L. (1981) Model Business Letters. London: Macdonald and
Evans
Gillibrand, P. and Maddock, V. (1982) The Manager and His Words.
Oxford: Pergamon Press
Hockey, S. and Marriott, I. (1980) Oxford Concordance Program:
Users' Manual, Version 1.0. Oxford: University Computing
Service
Macintosh, D. (1980) English for Business. Hong Kong: Book
Marketing
Makkai, A. (1980) "Theoretical and practical aspects of an Associa-
tive Lexicon for 20th-Century English" in Theory and Method
in Lexicography: Western and Non-Western Perspectives ed. by
L. Zgusta. Columbia, S. Carolina: Hornbeam Press
Naterop, B.J. et al. (1980) Business Letters for All. Singapore:
Oxford U.P.
Pike, K.L. and Pike, E.G. (1982) Grammatical Analysis. Arlington:
Summer Institute of Linguistics/University of Texas
Webster, J.J. (1983) "The PROLEX Project: procedures and applica-
tions" in Asian Culture I ed. by T. Tan and P. Woo. Singapore:
Society of Asian Studies
Webster, J.J. (1983) "Discourse frame theory: a computerized learn-
ing strategy for developing writing skills" in Proceedings of
the Malaysian National Computer Conference '83 ed. by D.S.S.
Lee. Kuala Lumpur, Malaysia: Malaysian Computer Society
Winston, P.H. and Horn, B.K.P. (1981) LISP. Reading, Massachusetts:
Addison-Wesley
Wölck, W. (1976) "Community profiles" Mimeo, State University of
New York at Buffalo