Tribble, C., (1989) The use of text structuring vocabulary in native and non-native speaker writing , *MUESLI News* Jun-89: 17-20

THE USE OF TEXT STRUCTURING VOCABULARY IN NATIVE AND NON-NATIVE SPEAKER WRITING

A number of MUESLI News readers attended the IATEFL SIGs meeting, held at Avery Hill in January. The article below is Chris Tribble's summary of the talk he gave there and there's a plea on the back page for more student texts to enable him to continue his research.

Background

This paper arose from a specific problem associated with teaching writing to overseas post-graduate students on an in-sessional course in Queen Mary College part of a Service English programme run by Bell Language Institute. The majority of the students I was teaching were working towards further degrees in a wide variety of disciplines and were all volunteers on an EAP Writing Course. Their problems related to difficulties that students seem to have in using text structuring language appropriately. 1 initially thought that the students' difficulties might lie in their command of the sorts of procedural lexis that have been discussed by Widdowson and, later, McCarthy, and embarked upon a study that intended to investigate the ways in which non-native speakers managed the exploitation of this type of vocabulary. An examination of the data l have available to me at the moment revealed that this was a false assumption. There was very little evidence of "the sorts of difficulty they present to learners". While l shall provide examples of this in the third section of this presentation, there were other pointers to something that had not struck me at all before and which was of immediate and practical relevance to my students. It is on this that l shall focus primarily. This combination of needs and opportunities has lead me to start a process of investigation. The process is neither complete nor nearing completion. It is, however, at a point when it seemed useful to try to report what l am trying to do.

Available analytic tools

The specific means of analysis which are currently available to me are a PC based word processor; the Oxford Concordancing Program (OCP) and a text-oriented programmingl language, SNOBOL4+. Before describing the textual problems l have been trying to deal with and the aspects of text that have, so far, been amenable to analysis, l would like to outline the potential of each of these instruments. By doing this the limits on the project itself will be easier to appreciate. WordPerfect 4.2 is a commercially available word processor that is widely accepted for business applications but which also has considerable potential as a tool for text study. Most modern WP packages have word search facilities in which a word or series of words can be defined by typing them as a "search string". Once specified, a preset series of key strokes will cause the cursor to travel to points in the text where the search string occurs. This is useful but common.

The advantage of WordPerfect is that not only is it possible to record a sequence of keyboard instructions as a form of program a "macron but also, it is possible to work with two documents simultaneously available to the operator. It is thus possible for example to call in another macro and start a second search and so on. This sort of facility can be used in relation to texts of unlimited length (other than by disk capacity) and provides a very powerful tool.

When one takes into account other facilities that exist in the program, such as line sorting, spell-checking, text marking, data-base and arithmetic facilities it becomes clear that the word processor is not just a glorified typewriter but something that can provide the means for a very wide range of research applications. Microsoft WORD has many of these facilities but l have not worked with another package where they are so well integrated although l understand that Samna is a package worth investigation. WordPerfect 5 which has just come on to the market has even more bells and whistles and should be seriously considered by anyone looking for a WP package for a PC. The Oxford Concordancing Program has been available for some time as a mainframe/mini computer

tool an d has recently been made available for the PC. It is able to construct text concordances, word indexes and word lists of enormous sophistication. Using OCP it is now possible for someone working at home to build concordances on the basis of affixes, subsections of words, words, phrases, collocations or combinations of these and to sort them by left or right contexts, as lemmas of verbs and so forth as well as exploiting tag codes to show the text provenance of each entry.

SNOBOL4

SNOBOL4 is a programming language originally developed by Bell Laboratories in the United States. It is now being used for a wide range of text analysing and processing activities and is attractive because it is specifically oriented to the sorts of string that are of interest to those of us who work in the humanities or linguistic sciences words. Chris Butler and Susan Hockey , through their writing, and Mike Stubbs, by inspiration and example, have been my guides into what has been, for me, a completely new area. Unlike a word processor, SNOBOL4 is able to deal with an, effectively, unlimited number of search strings in any particular operation. The first purposes to which I put it were involved in the construction of concordances or as a means of checking and selecting particular contexts within concordances that had already been prepared by SNOBOL4. The eventual purpose l have in mind is the construction of dictionaries I lexicons that can then be used as a means of identifying and tagging specific word categories.

Choice of analytic tool

Because the text elements that l was most interested in were primarily words and words that were relatively unambiguous it was obvious that OCP was going to be the most useful tool for the present purpose. It enabled me to construct selective tables of the words that existed in my target files and then make multiple searches for a wide range of strings across a text of unlimited length. The results of these searches could be ordered in any way that l wished and could be held as printout or files that could then be worked on with a word processor or SNOBOL4 for further analysis and exploitation.

Available sources

The first source of native speaker text used in this study was drawn from Kurzweiled articles (Kurzweil are the makers of a piece of hardware commonly used for Optical Character Recognition Eds) from The English Historical Review that l am currently collecting as part of a much larger study than the present. This currently comprises around 45,000 words and, while short in comparison with lexicographic corpora, constitutes a useful set of examples of formal native speaker writing.

Queen Mary College

A small corpus of student writing was compiled while preparing an in-sessional EAP course in Queen Mary College. This was initially keyed in so as to provide a resource that could be drawn on for reformulations and discussion of different aspects of nonnative speaker academic writing both at macro and micro level. Samples of work carried out for departments and as set essays were included giving a short text file of around 4,500 words taken from ten extracts and complete essays.

Longman l Birkbeck Corpus of Learners' English (LBCLE).

This is a new corpus currently being compiled for research, lexicographic and materials writing purposes. It brings together an international collection of students' writing that is coded in terms of text type, first language, nationality , level and so on. It is currently accessed using the Oxford Concordancing Program and exists as ASCII code which makes it possible to use text from it directly through DOS or a word processor. The version l have been working with is an early prototype containing around l00,000 words taken from over 250 extracts. Most of these are short essay and epistolary texts.

Procedural lexis

Narrowing the view of procedural lexis to that developed by E O. Winter (l978) and commented on in Carter and McCarthy (1988) an initial survey of the available corpus material was carried out using OCP. On the basis of a word list a comparison was made between the three text sources available and, when the data from LBCLE was studied, there was very little discernible problem noted in the use of the lexical items being sought (See Fig.l). This initial study also threw up many problems regarding the type of study l was trying to make especially in that l was only able to look for things that existed in the text rather than things that had been missed out! What did emerge was, broadly speaking, that the non-native speakers used the vocabulary set l was scanning for appropriately in terms of discourse function and also in terms of grammar and collocation. l had the feeling that the words matched concepts or categories that existed in Ll and, therefore, presented few problems in their use in L2 discourse. Obviously there is much that could be done in further investigating this area. Given the time available, however, it seemed more fruitful for my purposes to focus on other aspects of the texts that did seem to be throwing up interesting contrasts. While it was proving difficult to find any significant areas of difficulty with regard to a tightly delimited set of procedural vocabulary, it was very clear that a range of non-content items which emerged incidentally in the process of reading the concordance printouts was interesting. An unexpected contrast between native speaker and non-native speaker usage was observable and this did seem worth following up especially as it receives no attention in standard learner's guides such as Swann (1980), nor specific comment in so far as usage is concerned in LDOCE (1987) or COBUILD (1987).

Non-Content Vocabulary

The analytic procedure which was next adopted was to perform a generalised scan of the most frequently used vocabulary in the native speaker material (EHR) taking any item with a frequency greater than five (this number being arbitrarily selected but reflecting experience gained in the use of OCP less than five gets unwieldy). Once this word list was constructed (about four hours of computer time on my PC) it was then possible to deal with the output by extracting every item that could be considered as ''non-content" This was a very quick and dirty process that I would wish to refine at a later date. It was, however, from this list that the categorised lexical sets in were drawn up and the categorised sets have seemed useful. Working with the derived set of logical relationship markers it became clear that there was a significant difference between the ways in which native speakers and non-native speakers writers used certain sentence conjuncts. Notably these included :

- however; moreover; nevertheless;
- thereby. therefore; thus; while; yet.

This difference can best be described with reference to the corpus. Of the 31 instances of "however" obtained from the EHR, 25% were in sentence initial position while in the LBCLE 81% of the 48 occurrences occupied this position. This type of proportional differentiation was maintained across all of the samples drawn form the two text corpora and was supported by a concordance of the same items run on the small Queen Mary College file of student academic scripts

Application / Further Work

One of the main areas in which l hope to exploit the ideas with which l have currently been working will be in writing courses run at the Bell Language Institutes and Queen Mary College. l have already found it particularly fruitful to give students the opportunity to engage in the sort of analysis considered in this paper. By giving students the responsibility for assessing and constructing models for effective academic or formal writing l have found a way of avoiding the imposition of prescriptive and (frequently) inappropriate modes of expression and creating an enhanced awareness of the meaning potential of English as well as bringing about an improved performance in writing tasks.

Style in Academic Writing

The ultimate aim of the project which underlies the present work is to construct a comprehensive description of style in academic writing. The use of Sentence Conjuncts will constitute a small part of this description the way in which theme-rheme patterns are disrupted by an excessive attribution of thematic status to the conjunct itself rather than to the referential content of the argument itself. The major part of the work will involve the design of computer programs that will be able to allocate words to grammatical classes (or to attempt to draw on the experience of those working at Lancaster and Leeds with the TAGGIT and CLAWS programs) and then use this tagged text in comparative studies that will consider historical change in text style in a wide range of disciplines across a 70 or 80year period and contrast across a synchronic axis between disciplines. The eventual aim is to attempt to evaluate current models that are presented to EAP writing students and construct a pedagogy that may be better suited to the purposes of such students than that presently available.

Bibliography

Acerson, Karen L. 1988. *WordPerfect The complete reference*. London: Osborne Mcgraw-Hill

Butler, C. 1985 *Computers in the Humanities*. Oxford: Blackwell

Carter, R & McCarthy, M. 1987 *Vocabulary and Language Teaching*, Harlow: Longman

Emmer. Mark B. 1985 *Snobo14+, The Snobo14 Language for the Personal Computer User*, Catspaw Inc, Salida, Colorado, USA

Hockey S. 1985 *Snobol Programming for the Humanities*, Oxford: OUP

Widdowson, H. G. 1983 *Learning Purpose and Language Use*, Oxford: OUP

Winter, E.O. 1978. A look at the role of certain words in information structure. In Jones, KP and Horsnell, V. (eds.) *Informatics* 3: 1:85-97