Corpus Based Language Description in ESP Course Design

The Catholic University of Sao Paulo (PUC/SP)

June 28, 1998

Those who teach languages depend on those who describe them for their basic information and its dependability and reliability.
(Sinclair 1997)

1 Introduction

This paper describes research into the application of corpus derived data to the language description (LD) which is fundamental to ESP (English for Specific Purposes) course design. LD forms, along with the analysis of learner needs and wants, the pragmatic-humanist foundation of ESP. It is, necessarily, the analysis of language in use and, as such, it cannot be restricted to description of linguistic forms independent of the purposes that these forms are designed to serve in human affairs. Furthermore, LD should organize language into a useful format which provides for easy access to and application of language data.

Sinclair, arguing for the use of corpora in language description, states that:

> [...] fashionable ELT methodology has paid little attention to the state of language description, behaving as if the facts of English structure were no longer in dispute. In practical terms this has led to the growth and maintenance of a mythology about English [...] which language teachers take for granted, but much of which is challenged by corpus evidence [...] (1997:30).

With that in mind, this research set out to show how a corpus organized around the concept of interpersonal meaning, i.e. MOOD system (Halliday 1994, Eggins 1994), could be used to describe the relationships between language users and how the language they use reflects that relationship. The final step in the process would see the application of the data to the design of a computer mediated course for improving e-mail writing skills.

Four stages of methodology were examined: creation of the corpus, creation of a tagging system, application of the tagging system, analysis of the data and its influence on decisions in syllabus design. This paper is based on *"The complete consort dancing together...": Interaction in E-mai*l (Wyatt 1997), a Masters' Dissertation in Applied Linguistics, successfully defended in April, 1997 at The Catholic University of So Paulo (PUC/SP).

2 Method

2.1 Research Universe and Corpus

It was determined, early on, that the principal beneficiaries of the suggested ESP based course would most likely be academics and professionals who are interested in participating in e-mail discussion lists. These individuals are believed to have certain characteristics and needs in common which would allow them to benefit from computer mediated English courses at a distance. These are outlined below:

All list participants can be considered to be competent computer users. Participation in nearly all international discussion lists demands English language writing skills. A significant number of list participants in Brazil do not participate actively (lurkers). It is assumed that there is a significant number of Brazilian academics and professionals who are computer users and who would participate in lists if they had an appropriate level of English language competence and confidence. The groups formed by Internet users with like interests are frequently called lists. Their principal medium of exchange is the e-mail message. They exhibit the "six defining characteristics" necessary for being identified as discourse communities, according to Swales (1990: 24-7):

A discourse community has a broadly agreed set of common public goals. A discourse community has mechanisms of intercommunication between its members. A discourse community uses its participatory mechanisms to provide information and feedback. A discourse community utilizes and hence posses one or more genres in the communicative furtherance of its aims. In addition to owning genres, a discourse community has acquired some specific lexis. A discourse community has a threshold level of members with suitable degree of relevant content and discoursal expertise .

Three academic discussion lists which represented typical target language users were chosen. Archived messages were downloaded and organized by discussion list. Message headers were reduced to subject fields, only, and the names of individuals were replaced by generic identifiers. Place names and institution names were not changed. A sample of 100 messages was created from the archives.

2.2 Unit of Analysis

In order to analyze speech functions, a unit of analysis was defined, based on the notion "CLAUSE COMPLEX: a Head clause together with other clauses that modify it", which "enables us to account in full for the functional organization of sentences (Halliday 1994:215)."

Halliday uses the term "CLAUSE NEXUS" to identify a group of clauses related by taxis (1994:218).

> The clauses making up such a nexus are PRIMARY and SECONDARY. The primary is the initiating clause in a paratactic nexus, and the dominant clause in a hypotactic; the secondary is the continuing clause in a paratactic nexus and a dependent clause in a hypotactic (Halliday 1994:218).

The Clause Complex and the system of taxis offer precedents within Systemic Functional Grammar (SFG) for Head + Modifier type organization. By applying it as a paradigm, a unit of analysis, Function Complex (FC), was created. The FC describes a distinct SFG entity and avoids the possible confusion inherent in using terms like "clause" and "group", which have their own specific meanings in Halliday (1994). While all meanings (textual, experiential and interpersonal) have equal weight in a Clause Complex, an FC is intended to represent only interpersonal meaning, as described by the MOOD system (Halliday 1994), and just as the logical component of Clause Complexes is understood through the "primary + secondary" relationship of taxis, FCs demonstrate a similar, parallel organization within MOOD. An FC has a "head (primary) function + modifying (secondary) function(s)" structure. In the analysis, each clause complex was broken into its constituent clauses. The Head of the "primary clause" and the MOOD constituent of that Head were identified. Then, the MOOD constituent was treated as the Head Function and its FC classification applied to the whole complex.

2.3 Tagging System

To facilitate the analysis, a speech-function tagging system was developed from Eggins figure "Speech function system (discourse-semantic stratum) (1994:216)". The system provides a framework for analysis which breaks sample texts into their respective Speech Functions with enough delicacy to reveal functional-semantic preferences and tendencies. By following the system from first-level to second-level choices, it is possible to identify speech functions step-by-step.

The use of tags offered more refined access to the data. Through them, all the varieties of language employed for a particular Speech Function were gathered, quantified, alphabetized and compared quickly and efficiently. Database searches were performed combining both tags and specific language items. This method permited detailed searches and cross comparisons between functions too.

A tag consists of eight digits: (00000000). The first three (000ooooo) represent first level choices: giving OR demanding, goods&services OR information, innititing OR responding. The remaining five (ooo00000) mark the ways in which the speech functions are realized. Each digit can contain a value form zero to four. Zero indicates that the choice is not available in that instance. The first level choices are always non-zero, while the others are determined by the particular speech function which has been chosen: offer, command, question or statement. For example, the eighth digit which marks a response as "supporting" or "confronting" is not used to describe speech functions which initiate exchanges.

Tagging system organization

TAG= 8 digits (e.g. 12100010) 0=Not Applicable 1, 2, 3 etc.=Valid Choices

First-level Choices

First digit:	1=Giving	(INDICATIVE)
(role)	2=Demanding	(IMPERATIVE, INDICATIVE)
Second digit:	1=Goods and Services	(PROPOSAL)
(commodity)	2=Information	(PROPOSITION)
Third digit:	1=Initiating	(offer, command, statement, question)
(move)	2=Responding	

Second-level Choices

Fourth digit:	1=Subjective Orientation	Indicative: modulated interrog. Subj.: I
(giving+g&s)	2=Addressee Orientation	Indicative: modulated interrog. Subj.: you
	3=Indirect Orientation	Indicative: declarative
Fifth digit:	1=Inclusive	Let's + imperative
(demanding+g&s)	2=Exclusive	Imperative
	3=IndirectD	Indicative: declarative
	4=IndirectQ	Indicative: modal interrog. Subj.: you
Sixth digit:	1=Querying	Indicative: wh/interrog.
(demanding+info.)	2=Questioning	Indicative: polar interrog.
	3=Declaring	Indicative: declarative/modal declar.
Seventh digit:	1=Polarized	clause + polarity (is/isn't, do/don't)
	2=Marked-modal	clause + modality (Modalization/Modulation)
		[Modulation: Obligation/Inclination]
		[Modalization: Probability/Usuality]
Eighth digit:	1=Supporting	(acceptance, compliance, acknowledgment, answer)
(response)	2=Confronting	(rejection, refusal, contradiction, disclaimer)

Figure 1 Tagging system choices described by position in tag.

In table 1 the structure and meaning of the tag which describes an answer to a polarized question, \12200011\, is shown.

Table 1 Speech function tag for an answer: "\12200011\".

Digits	1	2	3	4	5	6	7	8
Tag	**1**	**2**	**2**	**0**	**0**	**0**	**1**	**1**
Meaning	giving	+ info.	+ respond				polar declar.	+ support
Levels	First Level Choices			Second Level Choices				

2.4 Application of Tagging System

Whole e-mail messages were evaluated as units for first level choices to determine whether a function could be assigned to each entire message. Thus, a message which requested an e-mail address was classified as demanding + informatiom + initiating.

Attention was then focused on message content. The messages were broken down into speech-function complexes and each complex given a tag, along with an identification number, which indicated in which message a complex was found and its position within that message.

Table 2 FC with identifier and tag

identifier(msg./pos.)	**tag**	**FC**
\"004.03"	\"12100010"	\"I received a copy of the..."

Finally, the resultant complexes were grouped by tag, alphabetized and written to text files. Each file represents what can be thought of as a family of function complexes whose constituents serve the same basic functional objective. An analysis of the functions was then performed to reveal common interpersonal content shared among the FCs. The results were used to create a description, or profile, of the language used by list participants which became the basis for all decisions on course content and sequence.

In each case the HEAD (Halliday 1994) of the primary clause was used as the identifying element of the FC. The process of identifying and tagging an FC is described below.

Example 1

Example of tagging

FC: *Somebody asked (sorry, but the delete finger is fast) why _The Pearl_ was banned in the orient.*

This FC is a clause complex consisting of five distinct clauses: a primary clause (with HEAD highlighted)

Somebody asked *why _The Pearl_ was banned in the orient.*

whose MODIFIER is a secondary clause

why _The Pearl_ was banned in the orient.

whose modifier is also a secondary clause

banned in the orient.

separated from the head by an interjection which is, likewise, a secondary clause

[I am] Sorry, but the delete finger is fast.

whose modifier is, likewise, another secondary clause but the delete finger is fast.

For the purpose of identifying FCs, only the HEAD of the primary clause is tagged.

Somebody asked (something)

- polar declarative: no response required

- giving information

- initiating exchange by restating

question tag = 12100010

3 Results

3.1 Message Types

Not surprisingly, the great majority (92.68%) of the messages analyzed were for the exchange of information. They fell into two broad categories: requests with their respective responses and unsolicited notices. The latter include: calls for papers, announcements for conferences, virus warnings, job offers, book and periodical publication notices, and the like. The remaining message types combined to make up less than 8% of the total and were, consequently, not given further consideration. Of principal interest was determining what should receive priority in the course syllabus. Table 3 Message Types Message typesPercentage of Corpus (221) demanding + information + initiating35.45% (122) giving + information + responding33.33% (121) giving + information + initiating23.90%

3.2 Function Complex Types

In order to determine FC types, the corpus was treated as a collection of FCs, not messages. The 96 messages analyzed produced 803 FCs. 39 of the FCs were discarded for obvious reasons. This left 764 FCs to work with. It was found helpful to combine FCs which were type initiating with similar FCs of type responding, since they tended to have more similarities than differences. For example, FCs of type: giving + information + initiating were combined with type: giving + information + responding to form a class of type: giving + information.

Table 4 Function Complex Types

FC type	Quantity	Percentage of Corpus
giving+information	569	74.5%
demanding+goods&services	122	16.0%
demanding+information	56	07.0%
giving+goods&services	6	00.7%

3.3 Correlation of Message Types with FCs

In table 4, message types are correlated with their constituent FC types. The percentages show the amount of each FC type in a message type. For example, under the heading for Message Type giving + information + initiating, the first entry shows that FCs of type: giving + information + initiating make up 73% of the content of those messages.

Table 5 Message composition by % of FC types.

		Message Types			
FC Types		giv+inf+ini	giv+inf+res	dem+inf+ini	dem+inf+res
giv+inf+ini		**73.0%**	**76.0%**	**52.0%**	**42.0%**
giv+inf+resp		01.0%	11.6%	06.4%	00.0%
dem+inf+ini		01.0%	02.0%	**25.0%**	**14.0%**
dem+inf+res		00.0%	01.0%	00.0%	**14.3%**

4 Discussion

The data derived from the research assumed two roles in the course design process. First, it helped determine what language the students would need to master in the academic e-mail environment and second, it provided a rich language resource which they would be able to utilize in several ways.

The course syllabus was centered around the demands of information exchange, the principal activity in the corpus. It was found that statements which had a nominal group as their subject outnumbered those which had "I" as their subject by a ratio of 2:1. The predominance of that type of third person subjects, as opposed to pronouns, is a strong indicator of the level of indirectness maintained in the messages. Consequently, the ability to produce that type of sentence was given priority. Certainly, students would have to do more than that, but the overall consistency of the way list users passed information allowed the choices of what to teach first to be narrowed considerably.

Example 2 SUBJECT (NG) + "is" + ATTRIBUTE

"12100010"\"Enrollment is limited ...

"12100010"\"FUN101 is a free, non-credit course ...

"12100020"\"GEOPOL is very fast and ...

"12100010"\"The language is so similar to many of our students' ...
The corpus also provided an overview of the levels of politeness and indirectness used by list participants. These became the main focus of the course, since the subject matter of the three lists differed, but their general manners showed little variation. Students are frequently concerned about being able to adjust their language to suit the formality or intimacy of the situation. The corpus indicated in a clear and concise manner just what list users expect from each other, in terms ofpoliteness and personal distance. Requests for information were divided between polar interrogatives (47%), declaratives (40%) and WH/interrogatives (13%). It is important to mention here that the polar interrogatives constitute an indirect manner of requesting since the responses they sought were not simply "yes" or "no", but something beyond. Thus, 87% of the information requests could be considered indirect and, as such, polite. The skills needed to produce these were also given a high priority in the syllabus.

Example 3

Indirect requests using modal polar interrogative

"\"004.09"\"21103000"\"Can you help?

"\"006.05"\"21103000"\"Can anyone provide information...

"\"019.01"\"21103000"\"Can anyone who has a computer lab recommend...

Example 4

Declaring

"22100310"\"suggestions on how to prepare for this adventure will also be appreciated inasmuch as ...

"22100310"\"Right now I will take any general information on the ...

"22100320"\"Any info on newspaper articles, [...] will be greatly appreciated. As a resource for learners, the corpus represents a space to be explored. By providing the tools to access and organize the information there, the course designers enable students to build their own information structures and customize their learning to their personal needs and preferences. The course design process involved a search for appropriate software. As of this writing, a final decision hasn't been made, but indications are that it will be a Perl based search engine on the course Web server.

5 Conclusion

This work has tried to demonstrate the usefulness of language description based on interpersonal meaning. The results, although tentative, indicate that a corpus tagged for speech functions can serve as a method for organizing linguistic data in such a way that researchers and students alike may compare and evaluate the ways in which language is mediated to satisfy interpersonal demands while still accomplishing the objectives of the speaker/writer.

References

[1] Eggins, Suzanne. 1994 An Introduction to Systemic Functional Linguistics. London. Pinter Publishers Ltd.

[2] Halliday, M. A. K. 1994. An Introduction to Functional Grammar, 2nd ed. London. Edward Arnold.

[3] Sinclair, .John M. 1997. "Corpus Evidence in Language Description". In Wichman A., Fligelstone S., McEnery T., Knowles G. (eds) Teaching and Language Corpora. London and New York. Longman. pp. 27-39.

[4] Swales, John M. 1990. Genre Analysis: English in academic and research settings. Cambridge. CUP.

[5] Wyatt, .Robert D. 1997. "The complete consort dancing together...": Interaction in E-mail. Masters' Dissertation. So Paulo. Pontifcia Universidade Catlica.