Gledhill, Christopher. 1998a. Learning a Genre as opposed to learning French. What can Corpus Linguistics tell us? In W. Geertz, & L. Calvi (eds.), *CALL, Culture and the Language Curriculum*, 124-137. Berlin : Springer Verlag. ISBN 3-5407-6192-6. **(PDF)**

CALL and Cultural Sudies in the Language Curriculum.

Chris Gledhill, Aston University

Learning 'a genre' as opposed to learning 'French'. What can corpus linguistics tell us?

* Corpus linguistics for French language studies.

* Learners and high frequency grammatical items.

* Ideological underpinnings of genre.

Extended Abstract.

The starting point for this paper is the realization that even the most basic grammatical items cause stylistic problems for advanced learners of French. At undergraduate level in the UK, learners of French are expected to produce a wider variety of genres and registers than they are prepared for by their entry qualifications. This causes particular difficulties which on the surface have little to do with formal knowledge of grammar and tend to be put down to 'style'. We argue here that corpus analysis and an empirical perspective to language development can bridge this gap and also unlock much wider issues, in particular the ideological perspective that underpins the language of a particular discourse community. While regular patterns that are not commonly included in expository writing syllabuses may prove to be useful 'set phrases' for the student in a particular assessment, we emphasise the more general role of variation in phraseology and possible exploitation of genre-specific idioms in the language syllabus. Secondly, the grammatical features of expository writing in French are different to those of the general language, and reveal broader issues of the extent to which French education and the media engender and reproduce their own discourse structures.

We examine some of the problems learners encounter when they attempt to transfer English structures in unfamilar communicative roles. Our main finding echoes that of Stubbs (1996): as the aims of a genre are more delimited so the lexico-grammar is restricted in ways that are mostly unconscious to the observer. Stubbs' argument in particular rests on the assumption that the observer is not normally able to examine large amounts of corpus evidence. Our corpus evidence suggests that French has systematically different uses of such words as *is* (not always transferable to French usage of 'est') and other false cognates. Recent corpus work on the grammatical properties of genres (Gledhill 1995) has revealed that grammatical items in different genres and registers behave quite differently to the general language, a fact sometimes obscured by the lexical approach (Nattinger and De Carrico 1992) and the factor analysis approach taken by Biber

(1988). This has led some researchers in the field (Willis 1992, 1996 and forthcoming) to suggest the development of specific learner corpora, comprising a specific task-based approach that attempts to direct learners directly to the structures they appear to need. From a corpus of texts written by French-speaking learners of English, Granger (1996) has shown that successful learners can identify key lexical phrases. But she is cautious about syllabuses aimed at simple phrase-orientation. We discuss the implications of this debate as far as high frequency items are concerned, especially in the light of our specific focus on grammatical items.

While long-range behaviour of lexis is recognised as integral to lexical phraseology, for example a large corpus can reveal the differences between *persuader* and *convaincre* on the one handand *persuade* and *convince* on the other (Johns 1990), there has been less research on the large scale patterns of grammatical items, especially in French. We argue that patterns of grammatical rather than lexical items are intrinsically interesting to the syllabus-designer and CALL specialist. The assumption of late has however been the inverse: not much evidence is needed to distinguish the uses of *dans* as opposed to *en*, and so research on grammatical items in specific corpora has been expected to reveal little of interest. On the contrary, our own research has suggested that the analysis of grammatical phraseology is a more efficient way of ascertaining generic structure and the main rhetorical tendencies in technical, scientific English (Gledhill 1995). Aside from this analytical debate, any hypotheses must be mitigated by the quality of the corpus: a journalistic corpus of French does not reveal patterns that are typical of the kind of expression that traditional syllabuses aim at producing and we echo Granger's (1996) caution.. Aside from analysis of grammatical items, this paper argues for the proper use of open-ended software with learner corpora and for greater sensitivity to the context of use of such corpora as Le Monde: these are powerful reference tools but are in no way representative of 'the language'.

Introduction.

In his discussion of linguistics in language teaching, Wilkins wrote:

> "Even the extremely proficient foreign speaker is still likely to be marked out as a non-native speaker if in his [or her] speech he [or she] seems to avoid the collocations that would be characteristic of the native speaker... Familiarity with collocations is normally considered a high mark of proficiency in a foreign language. (1972 128-9)

Wilkin's comments are extremely valid, and much writing on the 'lexical approach' to language teaching and syllabus design (Willis 1996) is reminiscent of this, although at the time, Wilkins was writing about the teaching of vocabulary and was pessimistic about our ability to get at mush data on the subject. Here we extend Wilkin's argument beyond simply the knowledge of lexical collocations, to the knowledge of the properties and the complexities of high frequency grammatical items. Much corpus work has emphasised the relevance of idioms and lexical phrases to teaching (Nattinger and De Carrico 1992, Willis 1992). Less work has been carried out

on the collocational properties of grammatical items largely because of the vast amounts of data involved but also because grammatical items are often considered to have few collocational properties (as Halliday and Hasan once maintained 1977). Sometimes, their properties are considered to be too well-known and too long range for further attention to be given to them (one can point here to the decreasing size of grammatical item entries from one edition of the Cobuild-Collins dictionary to the next). While these are prefectly reasonable arguments, especially for the general language, grammatical items have also been seen as the most important anchors in longer expressions (Moon 1987) and their increasinly regular use in special corpora allows for a quick characterisation of most of the regular expressions in that corpus. Both Hasan (1989) and Hoey (1992) identify the importance of lexical relations in text in terms of the overall cohesive texture that lexical chains provide. But it can be argued that grammatical items also have an implicit role in cohesive relations, since grammatical items can enter into regular collocational relations with relatively fixed meanings. For example in collocational frameworks it is difficult to interpret the NP1 in the framework *a(n) NP of NP2...* as anything other than a quantity or collective noun, or the item ? between *too ? to V* as anything other than a negative complement (Sinclair and Renouf 1991). Our previous work on technical English (Gledhill 1995) depends on the fact that prepositions and relative pronouns will 'dredge up' regularities across longer expressions, which are themselves often infrequent and statistically difficult to identify as salient. Thus grammatical items often reveal regularities that are in part hidden by lexical variation. But these regular expressions are not just traces of formulaic language. As Stubbs (1996) has argued, the choice of phraseology in a given corpus may throw up an ideological perspective which is very specific to to the genre and to the discourse community that uses that genre. For example, it is likely that in scientific English a high incidence of *to* used as complementizer in article introductions is indicative of a general tendency to introduce modalized statements of aims (This study aims *to*, In this study it is important *to:* Gledhill 1995). While in abstracts the tendency not to use this stucture and to prefer relative clauses (The study which resulted in..., The patients who sucessfully received...). This tendency is indicative both of increased compaction and of the need to state and define general terms (in defining relatives) to ensure greater coverage in the short abstract. But this is also, and quite independently, indicative of the typical expression of an underlying ideoology: patients are always expressed as those ‘who receive drug X... who receive

administration... who take part in studies' and the lack of agency of the scientific observer is more evident in the abstract than anywhere else in the corpus. Yet the fact that no writer could identify this as an explicitly 'known' aspect of their writing suggests that an underlying phraseology is at work (Gledhill 1996). Naturally, to get this kind of data requires a very delimited and documented corpus. This is also the case for learner corpora, where the range of functions and activities involved in the texts limits the complexity of the use of grammatical items and makes their analysis more manageable. The lexical orientation, then, has led to a revolution in the way we see traditional categories such as word class (Sinclair 1987), but for the purposes of extracting regular expressions, it may be better to reduce the size and the scope of the corpus and look for regular anchor-words, as we have seen in the case of research articles.

Before turning to the phraseology of grammatical items in French, we need to discuss our even more fundamental assumption that idiomatic expressions are dependent on genre or register. 'Genre' (Swales 1990) is essentially a text-type from the point of view of a discourse community, a group of specialists who achieve their common aims by the interaction of an agreed set of convetions. One of the explicit conventions is the mechanism of discourse: the medium of exchange being limited to, say, priviledged academic journals, another implicit and possibly hiddden convention is the phraseology of the group. 'Register' is very much a text-type from the point of view of the linguist, a series of texts defined by their linguistic characteristics, although Halliday (1985) and systemecists also use the term to describe the text in its functional situation as recognised by the wider speech community. Thus one may distinguish between the 'genre' of the academic research article and the 'register' of science writing for journalism. In practical terms, the disctinction is not so simple. What concerns us here is that students of French are asked to recognise a range of registers which native speakers would recognise and would be able to identify as authentic examples. On the other hand, we have argued that discourse comunities are not well aware of all the linguistic characteristics of their genres and this must also be true of expository writers, whether journalists, academics or other professional writers. Students are therefore being asked to identify and assimilate structures that are difficult to identify in the first place, especially without the benefit of a computer-held corpus of texts. While native intuitions may remain the best guides in these instances, the use of corpora should be considered in the

developmnent of syllabuses which aim to exploit this aspect of proficiency in a register or specific genre.

Syvliane Granger (1996), has been working on what she calls the use of 'prefabs' in Louvain's International Corpus of Learner English (ICLE). The 'prefab' is a complex expression which has become an acquired element of a learner's interlanguage. Prefabs range from fairly restricted lexical collocations (such as: *highly probable, bitterly cold*) (Benson et al.'s term 1996) to longer more discourse-oriented formulae (e.g. *as a matter of fact, the fact that*) (lexical phrases in Nattinger and De Carrico 1992). While it is assumed that non-native learners tend not to use idiomatic language as much as natives, as Wilkins suggested, Granger's learner corpus reveals that French non-natives actually have a large stock of expressions which they actively use, although they also tend to overuse a limited set of preferred expressions. What is interesting about this is that the learners have built up their own stock of ready-to-use expressions for expository texts (the kind of argumentative student essay demanded in the humanities) and that these expressions are not used in the way typical natives use them. Implicit in Granger's comments is a cautious rebuke of the Birmingham school's insistence on the 'lexical syllabus', the idea that learners be presented with as much data as possible without the benefit of pre-analysis by the teacher. According to Nattinger and DeCarico and Jane and Dave Willis, the principles of the lexico-grammar can only efficiently be acquired by the innovative analysis of examples using 'hypothesis testing', and that inclusion in the syllabus should be on the basis of frequency.

The term prefab itself clearly indicates the prepared nature of the expression, to be retrived at the learner's convenience. At times this limted set of learner prefabs consists of rare expressions which are cognate in both the L1 and L2 (e.g *highly civilized = hautement civilisé*). At other times the prefabs are unusual or inappropriate (e.g *I think that the fact that* is an inappropriate combination of two valid expressions). Interestingly, some favoured expressions by the NNS appear to be completly normal but are dispreferred by the native speakers (*one could notice that*). The point here is that learners can be seen to be learning fixed expressions in a formal environment: for comparative purposes Granger has the learners complete the same exercises,

otherwise the comparison of expressions would be forfeit, because, as we have argued above, collocations are usually very register-specific.

The purpose of this discussion is to demonstrate that on one level learners are expected to ‘learn’, that is actively consciously be aware of and be made aware of the pereferred phraseology in the language, and on the other hand we have discousre communities, albeit extremely specialised ones, where the accepted way of saying things (the phraseology) is subconscious and unknown, and unlearnt for the most part: it is instead ‘acquired’. I am not saying that scientists are not consciously checking their language and indeed much of the dynamics of scientific communication depends on the direct competition of differing ‘phraseologies’, but to a large extend, the very common patterns found in technical English and in learner’s French are difficult to confirm without the help of a corpus. To some degree scientists ‘learn’ in an evolutionary way: having peieces rejected and accepted for publication is a clear example of this, but many of the scientists in the corpus are non-natives and there must be some explanation for their ability to ‘acquire’ structures that are not identifiable or made explicit in general or tehnical English courses.

This is not too far away however from the position of the student of French. For the English student of French, the difficulties because of the difference between spoken and written registers are well known. These problems are exacerbated by the frequent use of ‘spoken style’ in the written register, and the formal educational and cultural expectations required by dense, informative prose in newspapers such as *Le Monde*. In British universities students of French are expected to write in ‘serious style’, and are stereotypically asked to read the ‘press’ to improve their range of language styles. In written language syllabuses, it is common for students to be exposed to such registers as literary reviews, political discourse and news reportage. On less traditional courses, the expectation is less discoursal, and students are asked instead to write and assimilate written versions of social surveys, marketing reports and advertising.