Then and now: A reconsideration of the first corpus of scientific English

John M. Swales
The University of Michigan

Abstract

The subtitle of Huddleston (1971) reads *A syntactic study based on an analysis of scientific texts*; this volume thus represents the first carefully designed and substantial corpus of scientific English. In this paper I re-examine a selection of his findings based on the science and engineering half of Hyland's corpus of 240 research articles. Features selected were variation in the passivization of individual transitive verbs, the paucity of instances of V + V-*ing* structures like "He continued working", and the meaning of the modal *must* in research prose. In all three cases, Huddleston's findings were largely confirmed in a database constructed about 35 years later, thus suggesting that English research writing in the sciences is, at least in grammatical terms, fundamentally stable. In the closing section, I contrast this linguistic stability with the rapid technological development of corpus linguistics. I instance a recent co-taught experimental course in which international senior doctoral students from the health and social sciences were able, with relatively little training and guidance, to construct paired corpora of their own research writings and of published articles from their own specialities and then conduct precisely the kinds of analysis that only a highly professional linguist could, with considerable more labour, conduct nearly forty years ago.

Key words: corpus linguistics, scientific texts, linguistic stability, doctoral students writing.

Resumen

El subtítulo del libro de Huddleston (1971) dice *A syntactic study based on an analysis of scientific texts*; este volumen, por lo tanto, representa el primer corpus cuidadosamente diseñado e importante de inglés científico. En este artículo me propongo reexaminar algunos de sus resultados basándome en la mitad del *corpus* de Hyland de 240 artículos de investigación. Las variables seleccionadas han sido la variación que existe en la frecuencia de voz pasiva en los verbos transitivos, la escasez de casos en que se producen estructuras V + V-*ing*, como en "He continued working", y en el significado del verbo modal *must* en la prosa de investigación. En los tres casos, la mayoría de los hallazgos de Huddleston han sido confirmados mediante una base de datos construida unos 35 años más tarde, por lo que se deduce que la escritura del inglés científico es, al menos en términos gramaticales,

fundamentalmente estable. En la sección final, establezco un contraste entre esta estabilidad lingüística con el rápido desarrollo tecnológico de la lingüística del *corpus*. Pongo como ejemplo un curso experimental que he compartido con estudiantes de un programa de doctorado pertenecientes a las áreas de la salud y de las ciencias sociales. Estos estudiantes pudieron, con una relativa mínima dirección y preparación, construir *corpus* paralelos de sus propios trabajos de investigación y de artículos publicados en sus especialidades, y a continuación realizar precisamente el tipo de análisis que sólo un lingüista altamente cualificado podría producir, con una cantidad considerablemente superior de trabajo, casi cuarenta años antes.

Palabras clave: lingüística del *corpus*, textos científicos, estabilidad lingüística, textos de alumnos de doctorado

Introduction

In the middle 1960s the British Government's Office of Scientific and Technical Information funded a research project into the linguistic properties of scientific English. The project was carried out between 1964 and 1967 at University College London. It was undertaken by three linguists and a computer programmer, and all three of the linguists involved, Rodney Huddleston, Richard Hudson and Eugene Winter, would go on to be important figures in their chosen linguistic fields, Huddleston and Hudson as syntacticians, Winter as a pioneering discourse analyst. The final 1968 report was entitled *Sentence and Clause in Scientific English* but was only produced in a few mimeographed copies. There used to be one copy in the archives of the British Council library at its London headquarters, but where it is now, or whether it has survived the vicissitudes of the British Council library policy, I do not know. (There may be a copy lodged at the UK National Lending Library for Science and Technology (or its successor) under the rubric of O.S.T.I Report No. 5030.)

The corpus contained 27 texts of 5000 words each, drawn from three strata: nine high-level texts taken from specialist journals; nine mid-level texts taken from undergraduate textbooks; and nine "low brow" science texts addressed to the educated layperson. This last group was taken from journals like *Scientific American, New Scientist and Discovery*. The extracts selected from specialist journals and textbooks were sub-classified into equal numbers of texts from physics, chemistry and biology, but this sub-classification was not attempted for the "more popular" texts. Thus, the database for the 1968 report consisted of a corpus of 135,000 words composed of

what might be described as the central sciences in the scientific register spread across three genres.

Because the 1968 report is virtually inaccessible, our information about this pioneering study has to be derived from Huddleston's 1971 volume entitled *The Sentence in Written English: A Syntactic Study Based on an Analysis of Scientific Texts*. Even this substantial volume (published by Cambridge University Press) is quite hard to find because it has been out of print for many years; however, secondhand copies are occasionally available, most typically as university library discards. Perhaps because of its comparative rarity, the book (henceforth SWE) has been little referred to in the development of ESP or EST or in the large number of studies devoted to the evolution, form and structure of scientific texts. There are, for example, no references to SWE in most of the major treatments of scientific rhetoric, and only fleeting ones in Nwogu (1990) and Valle (1999). There is just a little more in my *Episodes in ESP* -even though I unfortunately misspelled the author's name (!):

> The theoretical framework is mainly that of transformational grammar and restricted to a consideration of the syntax of single sentences. Reviewers and commentators have not found it of major or direct help in preparing EST courses, but it is a valuable source of reference, being rich in data and subtle grammatical distinctions. At the time of writing, out of print. (Swales, 1988: 16)

As that "time of writing" was the early 1980s, SWE has been out of print for at least 20 years.

Huddleston (1971), as the above quotation suggests, is something of a hybrid. Indeed, the author's opening sentence in his introduction states "I have had two complementary aims in view in preparing the present book: to give a selective grammatical description of a corpus of some 135,000 words of written scientific English and to investigate certain areas of the grammar of 'common-core' English - the grammar that is common to all varieties of the language (except possibly a few highly restricted ones)" (Huddleston, 1971: 1). The descriptive parts of SWE are mainly restricted to the clause level (because that had been Huddleston's particular responsibility in the 1964-67 project) but, even so, he hopes that the book will be useful to applied linguists preparing courses in scientific English. He also notes that "Until further comparative work … is done one cannot of course tell how far the

statistical properties of the corpus reported in the present work are peculiarly characteristic of written scientific English and how far they are generalizable to other varieties; I hope, however, to have provided a solid basis for such comparative study" (Huddleston, 1971: 2-3).

After the introduction, SWE has seven substantial chapters, with a shorter eighth, the chapter titles communicating something of the flavor of the volume:

2 Mood
3 Transitivity and Voice
4 Complementation
5 Relativization
6 Comparison
7 The Modal Auxiliaries
8 Theme

This paper will re-examine certain of Huddleston's accounts, especially some of those that have substantial quantitative data, in the light of the science and engineering components of Ken Hyland's (2000) corpus of 240 research articles. This comparative sub-corpus, of texts published in the 1990s, consists of 30 research articles each from the fields of physics, cell biology, mechanical engineering, and electrical engineering. The corpus totals just over 475,000 words, and is thus about three and a half times larger than that used by Huddleston. Unlike that in SWE, it is restricted to the single genre of the research article, while its disciplinary coverage, on the one hand, is a little broader because of my decision to include engineering, but, on the other, a little narrower because of the absence of chemistry texts. Another difference is that the texts were published 30-35 years apart. However, I am somewhat less interested in teasing out any particular differences between scientific texts from the 1950s and 60s and those from the 1990s, and somewhat more interested in seeing how and where a considerably larger and more genre-specific corpus might require some modification of the interesting findings about the syntax of scientific English found in SWE. (And in this, of course, I have a considerable advantage over Huddleston in that I have the *Wordsmith Tools* [Scott, 1996] concordancer program at my disposal.)

In his third chapter, Huddleston has some very interesting tables about the propensity of particular lexical verbs to occur in the passive. At one extreme, in the

SWE data, there are verbs such as *associate* that only occurred in the passive, while at the other, there are transitive verbs such as *acquire* that only occurred in the active. However, SWE's numbers are small. Would these dramatic -and pedagogically useful- differences hold up in a larger and more contemporary dataset? Other areas worth re-investigating are whether the *ing*-complementizer is as rare as SWE's data suggests, and whether in the larger corpus, it remains the case that the "obligation" use of *must* is more common than the "logical necessity" use. I will discuss these three grammatical features in order and then, in the final section, resituate these kinds of analysis in an EAP pedagogical context.

Passive and active verbs

Pages 120-126 of SWE consist of complex tables listing all the verbs that occurred nine times or more in the corpus. These tables are arranged in terms of declining percentages of passive occurrences, along with other information such as whether the passives are followed by a *by*-phrase of some sort. Although Huddleston does not specify which verb forms he included, it is fairly clear from the italicization in his examples that he included both finite and non-finite verb forms. The one exception appeared to be pre-nominal modifiers such as "*purified* liquid hydrogen"; however, "bare" participles occurring after the NP ("The pressure *shown*") were apparently included, presumably because of their more "verbal" character (Bolinger, 1973; Swales, 1981). In what follows, I have adopted Huddleston's practice.

According to SWE, four verbs always occurred in the passive, *associate, attach, derive* and *distribute,* while in the case of one other, *connect*, the passive percentage reached 95%. These are, I submit, remarkable findings, especially when we recognize that we can easily construct common active uses of these five verbs in our minds (such as "I'll distribute the flyer for you"). So what also does the Hyland sub-corpus have to say about these? Table 1 gives the total occurrences and passive percentages for these verbs in the two corpora:

Verb	SWE #	%	Hyland #	%
associate attach	24	100	165	96
derive	16	100	36	67
distribute	17	100	124	69
connect	10	100	23	85
	22	95	68	63

Table 1. High Percentage Passive Verbs in SWE and their Hyland Equivalents

First, if we bear in mind that the Hyland database is some 3.5 times larger, we can see that *associate* and *derive* are proportionally more common in the later corpus, *attach* and *distribute* less common, while the frequency of *connect* is approximately the same. I do not have any clever ideas for accounting for these differences, except that in Hyland the noun *distribution* is far commoner than its verbal counterpart. At least here, there has apparently been an increase in nominalization. Second, although all five verbs are still more likely to be found in the passive, their passive percentages have dropped, somewhat in the case of *associate* and *distribute,* considerably in the case of the other three. However, when we recognize that overall only about a quarter of the verbs in research articles are in the passive, the findings still indicate that the verbal behavior of these verbs is unusual. For example, we could look in a little more detail at the verb *associate* since it has the highest frequency of passives. Here the prevailing passive pattern takes the form of *be + associated with;* in contrast, of the eight active examples (out of a total of 172), only one is finite:

[1] We *associate* this effect with a particular set of large Fe-clusters, …

The remaining seven are non-finite, five being in the infinitive, as in:

[2] The first model endeavors *to associate* the use of inputs, …
[3] The purpose of the CPEN is *to associate* input (stimuli) with output (resources).

There is one other verb worth examining in Huddleston's list of verbs occurring more than 75% of the time in the passive because it is by far the most frequent verb in the category as a whole. This is the verb KNOW, which occurred 62 times in the SWE corpus, 79% of the time in the passive. There are 151 instances of this verb in Hyland, 120 in the passive. Amazingly, but doubtless coincidentally, this also amounts to a passive percentage of 79%! The commonest pattern here is *known to* (30 instances), followed by *known that* (24 examples), and then by *known as* (12 tokens). An example of each is given below:

[4] In particular, SHG *is known to be* extremely sensitive to the presence of inversion symmetry.
[5] *It is known that* the predicted rolling forces are higher than the experimental values.
[6] The behavior in the high-temperature limit *is known as* Curie's Law, as described in many textbooks.

At the other extreme, Huddleston lists verbs that never occurred in the passive in his corpus. These included, as we might suspect, a number of common intransitive verbs, such as *appear, consist, occur* and *seem,* but the list also contains a number of other verbs that at least have the potential to passivize. In the table below, I give the numbers and passive percentages of some of these verbs in the two corpora:

Verb	SWE #	% Passive	Hyland #	% Passive
act	32	0	60	2
help	22	0	16	0
reveal	16	0	78	9
acquire	9	0	24	46
agree	9	0	19	0
enter	9	0	41	0
imply	9	0	63	5

Table 2. All-active verbs in SWE and their Hyland equivalents

As the table shows, three verbs continued to be found only in the active in the 120 articles from science and engineering collected by Hyland. Even so, it is not so difficult to construct passive examples:

[7] The old lady *was helped* across the busy road.
[8] *It was agreed* that the contract should be signed.
[9] The data *was entered* into an Excel spreadsheet.

Next up the scale of frequency, there was only one passive use of the verb act, and this also was non-finite:

[10] As a second example, consider the simple harmonic motion of a rigid body (moment of inertia l) *acted upon* by a torsional spring…

Of the 63 examples of IMPLY, just three are passive and, again, they are all non-finite; here are two:

[11] In order to generate the corresponding representations for Q5, we should now include the action of X *as implied by* the coset expansion (25) and the…
[12] Even then, no serious attempt was made to modify the conservative assumptions *implied in* the Code rules, …

In contrast, five of the seven examples of passive REVEAL were finite, as in:

[13] Based on this approach, the kinematic meaning of induced construction parameters in spatial motion *are revealed*.

Finally, the Hyland figures for ACQUIRE (13 active and 11 passive forms) show that the absence of passives in SWE for this verb is probably a statistical fluke -*acquire* is commonly and easily passivized.

In this section, I have attempted to update the very interesting figures for active-passive use in the Huddleston corpus. In so doing, it has become very clear that although passive forms occur on the whole about a quarter of the time in scientific RAs, this broad generalization disguises the fact that *individual* lexical items vary greatly in their propensity to passivize. The new data from Hyland shows that these differences are not always quite as dramatic as in the older, smaller SWE corpus, but they are nevertheless both substantial and of relevance for junior researchers with limited English language proficiencies. It would also seem that some of the verbs that do not go easily into the passive, such as *imply,* may have a greater acceptability in non-finite forms, particularly when used in reduced relative clauses. The following table summarizes the findings given above, along with three further verbs (indicated by a *) with high rates of the passive:

Verb	Percentage of Passives
associate	96
*attribute**	92
*arrange**	92
distribute	85
*relate**	84
derive	69
attach	68
acquire	46
reveal	9
imply	5
act	2
agree	0
enter	0
help	0

Table 3. Transitive verbs with various passive percentages in the Hyland sub-corpus

These differences are, I believe, sufficiently large to be able to speak for themselves.

The occurrence of *-ing* verb forms in three corpora

Prima facie, one of the surprising features of Huddleston's corpus is the rarity of the *-ing* complementizer following what Huddleston -and others- have called a matrix verb and Palmer (1965) and Mindt (2002) a 'catenative'. Some of the few examples from the Huddleston corpus are illustrated below:

[14] One technique… *involves measuring* the extent to which the individual spectral features…

[15] …after transplantation the patient *requires nursing* in a unit in which the air is filtered of all bacteria.

[16] …designers *anticipate being* able to save power, …

[17] In Fig 18-2 we *see* a force F *acting* at a point r.

If we leave aside those governed by a preposition, particularly *by verb-ing* operating as an agentive, there are in SWE merely 14 instances of the *-ing complementizer* -in contrast to over 700 where the complementizer is *to* (a ratio of 1: 50). The matrix/catenative verbs and their numbers are shown below:

With intervening NP		Without intervening NP	
picture	1	*continue*	3
see	1	*involve*	2
		see	2
		anticipate	1
		go on	1
		include	1
		resent	1
		stop	1

Table 4. Matrix verbs followed by the *-ing* complementizer in SWE

These figures probably strike the reader as being extremely low in a corpus of 135,000 words, especially as there is evidence (Rudanko, 2000) that the *-ing* complementizer has become increasingly common in English. As a result of both of these perceptions, I confidently expected that the relative frequency of this structure would be considerably increased in the larger and more recent Hyland sub-corpus. However, the raw numbers for the larger Hyland database are four tokens for *involve,*

single instances of *continue, anticipate* and *include,* and no instances of any of the others listed in Table 4. Here are two examples:

[18] It *involves creating* specialized cells…
[19] …then the machine *continues processing* parts from this buffer…

At this point, it can be correctly observed that these figures might simply result from the fact that the matrix verbs themselves are uncommon. So let us consider a few other verbs that have the potential of taking this structure. Take the case of the lemma REQUIRE. This occurs in Hyland 55 times, nearly always followed by to, and only once by the *-ing* form of the following verb:

[20] Changing the number of energy units *requires also adding* more lines to the spreadsheet…

The verb *propose* occurred 115 times in the sub-corpus, but again only once in the target structure:

[21] Pudney (1989) *proposes modelling* this sort of situation using discrete random preference regimes.

The lemma START occurs 64 times in Hyland, 12 times with *to* but just twice with *-ing*; *begin* occurred on 49 occasions, 16 times with *to,* but not once with *-ing*. In fact, the only verb I found that occurred more than 30 times with 10% or more of those occurrences using the *-ing* complementizer was AVOID, with 49 tokens, six of which were in the following form:

[22] …we can *avoid dealing* with the awkward form of the potential…

Standard reference grammars of the English language, such as Downing and Locke (1992), give considerable play to this structure, especially its use with verbs of liking and disliking. They are also fond of contrasting the different semantics of the *to-* and *-ing* complementizers with certain verbs:

[23a] Try to publish in *Ibérica* (attempt to publish in *Ibérica*)
[23b] Try publishing in *Ibérica* (experiment with publishing in *Ibérica*)

There is also usually some discussion of other verbs, such as *begin* and *start,* which can also take both complementizers, but this time with no apparent difference in meaning.

However, for the *science writer,* this V + V-*ing* structure seems to be extremely marginal, except perhaps with the verbs *avoid* and *involve*. This useful, albeit negative, evidence has been gleaned from a close reading of Huddleston (1971) and has been confirmed from searches in the Hyland corpus of scientific research articles. As all leading corpus linguists have observed, our intuitions about frequencies are often at odds with reality. Although nobody might have expected (or intuited) that the structure discussed in this section would be that common in academic writing, I think very few (very much including myself) would have expected it to be that rare.

The question now arises as to whether it is also rare in academic and research speech, and for this I have turned to the 36 speech-events (dissertation defenses, colloquia, seminars, research group meetings etc) collected in the MICASE research sub-corpus, and totalling about 450,000 words. In the table below, I give first the overall numbers for selected verb lemmas, followed by the numbers of them followed by the *-ing* complementizer, and then this number expressed as a percentage of the total.

Verb	Total	Total *-ing*	Percentage *-ing*
anticipate	22	0	0
avoid	40	5	12.5
begin	67	5	7.5
continue	55	2	3.6
hate	24	2	8.3
involve	97	6	6.2
propose	29	1	3.4
require	60	1	1.7
start	420	123	29.3
stop	77	27	35.1
suggest	70	1	1.4
try	275	6	2.2

Table 5. The *-ing* complementizer in the MICASE research sub-corpus

With the exception of the catenatives START and STOP, it does not look from these figures as though this structure is of great utility to research speakers either! Even so, here are some examples:

[24] i would probably, want to *avoid having* a very heterogeneous group…

[25] we would also like to *continue evaluating* our stochastic model,
[26] i *hate applying* for summer money.
[27] the operations *proposed requiring* only paper and pencils…
[28] we never *tried tuning* inside the oscillator…
[29] has anyone *tried reading* this book?

I also included *suggest* among the group of targeted matrix verbs even though it is not normally associated with this structure, because it is not difficult to attest in conversation structures like "I suggest leaving a bit after ten". However, as Table 5 shows, there was but a single instance:

[30] so, the book *suggests, using* a picture to, characterize the, values of the other two variables.

The second most common verb in the list is TRY (cf. examples [28] and [29]); here there are six examples of the *-ing* complementizer, but 211 with the *to* complementizer, suggesting that the "experiment" meaning associated with [23b]) above is only rarely invoked. In contrast, the most common lemma (START) shows a distinct preponderance of the *-ing* complementizer, with double figure numbers for *start* looking*, *start* talking* and *start* thinking*. The percentages for the *-ing* and *to* complementizers in MICASE following START were respectively 69% and 31%. The first figure is close to Mindt's percentage of 71 for *-ing* for spoken conversation (Mindt, 2002), but Mindt's figure for expository prose of 48% is much higher than the Hyland research corpus figure of 14%, a discrepancy for which I have no explanation at present –although the Hyland numbers are small (2/12).

The Modal *must*

My third and last 'reconsideration' of Huddleston (1971) concentrates on the two-page discussion toward the end of his book where he deals with the modal auxiliary *must*. There were 116 of these in his corpus, but he only discusses in detail a sub-set of 37. As is well known, this modal has two distinct meanings; one of obligation ("You must study harder"), the other of logical necessity or logical conclusion ("You must be joking!"). Huddleston, however, says that at least in his analysis of the science texts, there might be a case for a third category –that of "necessary conditions" (p. 312). One of the examples he discusses is the following:

[31] Calcium carbonate is deposited wherever there is a mucilage layer and an aggregation of the chloroblasts, but apparently these conditions *must* be fulfilled before lime is laid down.

As he notes with regard to [31], "unless the conditions are fulfilled, lime cannot be laid down" (p. 312). However, in the end he does not pursue this third option and classifies this use of *must* as obligation.

A total of 24 of his 37 analysed instances Huddleston places under obligation and the remaining 13 under logical conclusion. Although he does not comment on these numbers, they would likely strike many people as somewhat surprising. After all, conventional wisdom about science and engineering would suppose it to be –at least in its written manifestations– a world of empirical calculation and logical reasoning. In such a universe of discourse, one might further suppose, the logical necessity or conclusion use of *must* could be expected to predominate, rather than the often-interpersonal obligation sense. Would then the Hyland corpus confirm the distributions found in SWE or would they confirm our commonsense expectations?

There were 210 instances of *must* in the Hyland science and engineering corpus, proportionally therefore somewhat less frequent than in the multi-genre data in SWE. However, this use of *must* was unevenly distributed in terms of discipline since 91 of the 210 tokens occurred in a single field –that of electrical engineering. Although Huddleston claims that it is comparatively easy to sort examples into the obligation and logical conclusion meanings, I experienced greater difficulty and I have left 10% uncertainly classified; most of these were of the "necessary conditions" type referred to earlier. Of the remaining 189, only 42 (or 22%) strike me as reflecting logical necessity or conclusion, thus leaving a clear majority (78%) to exhibit obligation. Here are three examples of the former, followed by two examples of the latter:

[32] it was concluded that all five genes *must be similar*…
[33] evidence suggests that such splicing factors *must exist*
[33] fails utterly to describe this system (*as it must*, since it applies only to degrees…
[34] …i.e., the angles *must be carefully aligned* to be in the vicinity…
[35] As such, the design team *must analyse* the alternatives…

As might be expected, there was a clear tendency for this major obligatory use to be followed by a verb of action, often in the passive, while the logical use was more likely to be realised by an equative verb or an adjective, such as *equal* or *similar*. So, Huddleston's initial discussion of the two uses of *must* is more than confirmed by the findings from the larger and later corpus. Indeed, it is also confirmed by Biber et al. in the *Longman Grammar* (1999: 495), which states: "The modal *must* is particularly intriguing here because its distribution runs counter to the expectation of personal involvement: the extrinsic meaning of logical necessity is most common in conversation, while the intrinsic meaning of personal obligation is most common in academic prose". And this we also find in research articles.

Conclusions and Applications

It might be thought that the material presented in the last three sections offers a rather arid or at best a rather academic exercise in register or genre analysis. After all, the "then and now" aspects of the study have shown that Huddleston's findings produced in the first decade of English for Specific Purposes have largely been confirmed by a larger database produced in the last decade. Those findings show that transitive verbs do indeed vary greatly in their propensity to passivize, that *-ing* complementizers following catenative verbs are much rarer than descriptive grammars would have us believe, and that in research articles *must* most of the time expresses obligation rather than logical conclusion –and this despite the potential threats to face that this strong sense of the modal tends to invoke.

However, it is precisely the kinds of investigation depicted in this article that have played the major role in an experimental elective EAP course taught at the English Language Institute at the University of Michigan in winter term 2004 by David Lee and myself. The course was entitled "Exploring your own discourse world" and was designed for senior international doctoral students from across the university. All of the classes (until the last two weeks) took place in one of the university's computer classrooms (with projection facilities), where were installed *Wordsmith Tools*, the Michigan Corpus of Academic Spoken English (MICASE), Hyland's full corpus of 240 research articles, and where, with David's help, permission for web-access to the British National Corpus (BNC) was obtained. By design it was a small class: there were two Chinese participants from pharmacology, one from biostatistics, and one from educational technology, plus a Russian sociolinguist from German Studies, and

a Fulbright scholar from Pakistan working on a genre analysis of computer science research articles. As can be seen, only the last had any previous direct involvement with this kind of research on this kind of material.

The participants had on entry impressive computer skills, very considerable abilities in quantitative analysis in their own particular fields, a strong desire to be successful published younger scholars, and, generally speaking, pretty high levels of English language proficiency. In the first half of the course, they learnt how to take advantage of the full facilities that *Wordsmith Tools* offers and undertook precisely the kinds of analyses I have discussed in this paper on the corpora made available to them. In effect, they became quick and effective corpus linguists, especially in searching out the collocational patterns of research prose and (prompted by the instructors) testing out the claims made in standard textbooks and grammars. In an interesting recent paper entitled "Empowering non-native speakers: The hidden surplus value of corpora in continental English departments", Mair (2002: 121) trenchantly notes: "Corpora empower learners because they provide a means of independently corroborating or disconfirming native judgements, and of determining their scope in cases where two of them are in conflict". It was this empowerment that we wanted to transfer to the participants.

In the second half of the course, the students were encouraged to develop their own corpora, which they all did except for the German Studies participant. In three cases, the pharmacologists jointly, the biostatistician, and the educational technologist in fact developed two corpora, one of their own research writings, and one of published articles in their subfields. Generally speaking, they managed to put together about 10 texts of their own, and about 30 published articles. Constructing these corpora was quite time-consuming –especially if it meant converting pdf files to textfiles and stripping out tables, figures, references etc.– but they embarked upon these tasks with great enthusiasm. In an interview, it became clear that, although they found MICASE "quite interesting", and Hyland, "very useful", these were much less exciting than being able to explore and compare texts that specifically represented their own precisely-targeted discourses. The groups then worked on final projects, the details of which will be discussed elsewhere; these projects were, in the last week of the course, presented to an invited audience of EAP practitioners, who were clearly struck by what these non-linguists and non-English specialists have been able to achieve, one outsider commenting that they were more interesting than nearly everything she had

heard at the recent TESOL convention. It is perhaps especially telling that three participants have bought *Wordsmith Tools* for themselves, so that they can take permanent advantage of what they have learnt and of the databases they have put together. The biostatistician has now graduated and in September 2004 will take up an assistant professor position in the biostatistics department at the University of Pennsylvania, a prestigious Ivy League school. Like Mair's English students and teachers in Germany, her corpora (and her *Wordsmith Tools*) will empower her, as a non-native speaker of English, when she deals with the research drafts written by her own graduate students, whether Americans or internationals.

The "then" of this paper was the 1960s when Huddleston was analysing the sentences of written scientific English at the University of London, and I was just beginning to teach scientific English in the College of Engineering at the University of Libya in Tripoli. At that time, neither of us would have dreamed that forty years later (in the time of "now") that a pair of Chinese pharmacology doctoral students, with very little training, would have themselves put together a corpus of over 100,000 words partly to be able to demonstrate that the definite article has been declining in English language medical research articles and in ways that standard grammars do not countenance, or that a Chinese educational technologist can show that her choices of reporting verbs were consistently less evaluative than those used in the published research articles in her field. This recent experience would seem to suggest, at least for certain high-flying advanced learners of English, that a personal investment in becoming, on occasion, a corpus linguist can lead, especially within their circumscribed universes of discourse, to considerable linguistic and rhetorical consciousness raising, which can then lead to greater confidence and competence in both their written expressiveness and their capacity to reflect upon it.

Acknowledgments

I would like to thank Ken Hyland for making his corpus available, and David Lee for co-planning and co-teaching the experimental course and commenting on a draft of this paper. The usual disclaimers apply.

REFERENCES

Biber, B., S. Johansson, G. Leech, S. Conrad & E. Finegan (1999). *Longman Grammar of Spoken and Written English*. Harlow, UK: Longman.

Downing, A. & P. Locke. (1992). *A University Course in English Grammar*. London: Routledge.

Hyland, K. (2000). *Disciplinary Discourses*. Harlow, UK: Longman.

Huddleston, R. (1971). *The Sentence in Written English: A Syntactic Study Based on an Analysis of Scientific Texts.* Cambridge: Cambridge University Press.

Mair, C. (2002). "Empowering non-native speakers: The hidden surplus value of corpora in continental English departments" in B. Kettemann & G. Marko (eds.), *Teaching and Learning by Doing Corpus Analysis,* 119-130. Amsterdam: Rodopi.

Mindt, D. (2002). "A corpus-based grammar for ELT" in B. Kettemann & G. Marko (eds.), *Teaching and Learning by Doing Corpus Analysis,* 91-104. Amsterdam: Rodopi.

Nwogu, K. N. (1990). *Discourse Variation in Medical Texts: Schema, Theme and Cohesion in Professional and Journalistic Accounts*. University of Nottigham: Monographs in Systemic Linguistics, 2.

Palmer, F. R. (1965). *A Linguistic Study of the English Verb*. London: Longman.

Rudanko, J. (2000). *Corpora and Complementation: Tracing Sentential Complementation Patterns of Nouns, adjectives and Verbs over the Last Three Centuries.* Lanham, NY: University Press of America.

Scott, M. (1996). *Wordsmith Tools*. Oxford: Oxford University Press.

Swales, J. (1981). "The function of one type of particle in a chemistry textbook" in L. Selinker, E. Tarone & V. Hanzeli (eds.), *English for Academic and technical Purposes*, 40-52. Rowley, MA: Newbury House.

Swales, J. (1985). *Episodes in ESP*. Oxford: Pergamon.

Valle, E. (1999). *A Collective Intelligence: The Life Sciences in the Royal Society as a Scientific Discourse Community, 1665-1995*. University of Turku, Finland: Anglicana Turkuensia No. 17.

John M. Swales is Professor of Linguistics at the University of Michigan, where he was also Director of the English Language Institute from 1985-2001. His most recent books are a second edition of *Academic Writing for Graduate Students* (with Chris Feak) (University of Michigan Press, 2004) and *Research Genres: Explorations and Applications* (Cambridge University Press, 2004).