Identifying Chains in Spoken Academic English

David Brett

Introduction

John Sinclair (1991: 109-110) proffers two models for the interpretation of language:

a) the open-choice model, in which the production of language is seen to be a piece-by-piece construction of phrases and larger syntactic entities. The most typical realisation of this approach is the tree-diagram, at each step, represented by the nodes, a large number of choices may be made, and "the only restraint is grammaticalness". Sinclair describes this as being the usual approach adopted by grammars at the time the work was written;

b) the idiom principle, described as follows:

> The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments. To some extent, this may reflect the recurrence of similar situations in human affairs; it may illustrate a natural tendency to economy of effort; or it may be motivated in part by the exigencies of real-time conversation. However it arises, it has been relegated to an inferior position in most current linguistics, because it does not fit the open-choice model.

When evaluating the relative importance of the two models, Sinclair states that neither are sufficient on their own to describe language, and especially spontaneous oral production. To the contrary, both play a fundamental role, however, the second should be adopted as default, the first is to be resorted to only in those cases whereby a phenomenon cannot be explained by the idiom principle.

> For normal texts, we can put forward the proposal that the first mode to be applied is the idiom principle, since most of the text will be interpretable by this principle. Whenever there is good reason, the interpretive process switches to the open-choice principle, and quickly back again.

This standpoint is echoed by Mason (2008: 237):

> Essentially we assume that pretty much everything has been said before, though that is of course an over-simplification. There are indeed new and

> creative constructions, but they are the exception rather than the rule. Most of language will consist of chunks that have occurred before, just as we tend to re-use words and occasionally introduce new coinages. But it is not only the words themselves than we re-use, it is also their contexts, as they are inseparable. And their contexts are effectively multi word units.

According to the open-choice model, the point at which selection from the lexicon may take place is usually at the word boundary, i.e. after a typographical space, barring a number of clear exceptions: the space dividing *out* and *of* is clearly arbritary (cf. *into*); English is particularly rich in phrasal and prepositional verbs; regarding the space after *of* in the discourse marker *of course* as a point at which choice can be made is obviously unappropriate. However, demonstrating and evaluating the presence of the idiom principle in a text implies its being analysed for multi word units (MWUs), which are almost inevitably defined as such on the basis of the number of spaces between uninterrupted strings of alphanumeric characters. The fact that considerable inconsistencies exist in conventional usage with regard to the typographical space is a necessary evil that must be taken in account when conducting analysis (Note Suffice to consider the variations of orthographic representation of compound nouns).

Mason (2008: 233) provides an example of how samples of text can be analysed by way of comparison with a reference corpus to highlight the presence of 'chains', a term which is deemed to be "less awkward than the more general term 'multi-word units'". Eleven short passages of written text pertaining to different genres, such as science fiction, academic prose, news and children's fiction were analysed by way of a search for *n*-grams[1], with *n*=2-7, in the written part of the British National Corpus (BNC). Nine of the samples displayed percentages of formulaicness above 50%, with one sample, a conference call for papers, reaching 100% coverage in the reference corpus. The sample with the lowest coverage, at 11%, was a linguist's (Hoey 2005) rephrasing of a sentence by a travel writer, deliberately wrought to avoid typical lexical relationships. A weighting system was applied in the analysis, whereby significance was given to the chains not only on the basis of frequency but also on the value of *n*: "since short chains are usually more frequent while longer ones occur less often, using frequency alone would favour short chains" (Mason 2008: 233). A similar approach is adopted in this paper; furthermore, an attempt will be made to provide graphic rendering of the data, in order to highlight the degree to which a given token participates in chains to the left and right at the various values of *n*.

1 Materials and methods

1.1 The sample text

The sample text is composed of the transcription of the first five minutes of the second in a series of 26 lectures delivered at the University of California at Berkeley in 2006[2]. This lecture series, "Physics for Future Presidents", was chosen principally because, while it concerns an academic subject represented in the reference corpus, the content is not geared towards experts, instead it aims to illustrate general notions to students from non-scientific backgrounds. In fact, a very low number of vocabulary items in the sample text could be deemed typical of the specific field: *calorie*, *horsepower*, *energy*, *gram*, *kilowatt*, *watt*, *joule(s)*, *molecule(s)*, *motion*, *celcius*, *nutritionist's*, *pound* (as a measurement of weight), *scientific*, *scientists*, *particle*, *vibration*; even the most technical terms, the units of measurement *watt* and *joule* should be familiar, at least as terms (though not necessarily their exact denotation), to anyone with second-level education.

Dudley-Evans (1994: 148) identifies three academic lecturing styles:

1. the reading style, "in which lecturers either read the lecture or deliver it as if they were reading it". Tone groups are seen to be short, and the intonation range narrow, with falling tone predominating;

2. the conversational style, "in which lecturers deliver the lecture from notes and in a relatively informal style with a certain amount of interaction with students". Tone groups are longer and key shifts may occur;

3. the rhetorical style, "in which the lecturers give a performance with jokes and digressions". The intonational range is wide and the high key is often exploited. There are frequent asides and digressions marked by key and tempo shifts.

The sample text can be attributed with substantial certainty to the third category. From a textual point of view, a considerable number of digressions are present in the form of anecdotes from the speaker's personal history, such as: when he was at school; when he was backpacking; and a misunderstanding that occurred when he was cooking with his wife. From a prosodic point of view, analysis of a 73 sec. segment of the sound file yielded the following results: the mean tone unit length was 6.08 tokens (σ=2.47) and the articulation rate (syllables/second excluding pauses) was equal to 5.68 (σ=1.63). Of particular interest was the latter figure, as the speaker was seen on the whole to be speaking rather quickly[3], however, as can be seen from the standard deviation value, the amount of variation was considerable. This variation was seen to correspond closely to the function of the individual tone units: in an anecdote regarding carrying water when backpacking the rate of articulation is in the range of 7-8 syllables per

second, whereas when giving definitions concerning the calorie, this figure drops to 3-4.

The sample text is composed of 959 tokens, with 324 types, resulting in a 33.79 type/token ratio. Comparison may be made with a text of the same number of tokens at the start of a physics lecture in the BASE corpus (pslct034). While the type/token ratio is slightly lower, 28.99, we may note the presence of a number of types which may be opaque to the lay person, at least in the technical sense in which they are used: *holography*, *dimensional*, *phase*, *beam(s)*, *plate(s)*, *fringes*, *interfere/interference*, *amplitude*, *wave(s)/wavefront*, *fourier* and *intensity*.

Finally, the sample text was produced by a native speaker of American English (b. 1944 in New York), who is an expert in nuclear and particle astrophysics, and Full Professor at University of California at Berkeley. Hence, we may safely assume that the text is a representative sample of Spoken Academic English.

1.2 The reference corpus

The reference corpus was composed of text files from the BASE (British Academic Spoken English) and MICASE (Michigan Corpus of Academic Spoken English) collections. The former is constituted by c. 1.6 million tokens from 160 lectures and 40 seminars, video and audio recorded at the universities of Warwick and Reading, respectively, from 2000 to 2005. The corpus was designed to be representative of four broad disciplinary groups (see Table 1 for details), each category being composed of 40 lectures and 10 seminars. While the trascriptions are freely available in plain text and tagged XML format, access to the search tool and multimedia files requires a subscription[4].

The MICASE collection, which has been elaborated at the University of Michigan since 1997, is currently composed of 1.8 million tokens of trascription of spoken American English in various academic settings, including lectures, meetings, seminars and so on[5]. The architecture of the corpus reflects that of BASE to a great extent, in that the four main macro-areas are represented in more or less equal proportions, although the terminology varies in some cases, as can be seen in Table 1. Furthermore, the variety of speech events covered is considerably greater:

> The speech events included in the corpus include: small and large lectures (62), public interdisciplinary or departmental colloquia (13), discussion sections (9), student presentations (11), seminars (8), undergraduate lab sessions (8), lab group and other meetings (6), one-on-

> one tutorials (3), office hours (8), advising consultations (5), dissertation defenses (4), study groups (8), interviews (3), campus/museum tours (2), and service encounters (2)[6].

MICASE also features a fifth category, 'Other', in which content of a miscellaneous nature is stored.

In order to homogenise the two corpora, occurrences of *gonna* in the MICASE files were changed to *going to*. While in the BASE and MICASE corpora the lectures, seminars etc. are each in separate files, for the purposes of this study it was deemed suitable to merge all the files of a given disciplinary category into a single large file. As a result, the reference corpus was composed of nine large text files: two relating to each of the four superfields foreseen by the design of the BASE and MICASE corpora, as well as one file relating to the 'other' category present in MICASE. While this entailed the loss of a certain amount of detail with regard to possible differences in degree of interactivity between lectures, seminars, colloquia etc., it greatly aided the readability of the provenance of occurrences. In other words, totals of occurrence of chunks in the reference corpus were flanked by nine columns breaking down these results to the disciplinary level, the alternative being 400-odd columns, should each file have been taken individually. The large files will be referred to henceforth by way of the abbreviations indicated in Table 1. Statistics concerning the files in the reference corpus, derived using the WordSmith Tools Wordlist program, are reported in tables 2 and 3.

BASE		MICASE	
Arts and Humanities	AHB	Arts and Humanities	AHM
Physical Sciences	PSB	Physical Sciences and Engineering	PSM
Social Sciences	SSB	Social Sciences and Education	SSM
Life and Medical Sciences	LMB	Biological and Health Sciences	BHM
		Other	OM

Table 1. Abbreviations for files in the reference corpus

Text File	TOTAL	AHB	PSB	SSB	LMB
Tokens	1,675,671	439,111	344,358	458,995	433,207
Types	32,093	18,211	10,373	14,522	13,742
TTR	1,92	4,15	3,01	3,16	3,01

Table 2. Statistics concerning the files from the BASE corpus

Text File	TOTAL	AHM	PSM	SSM	BHM	OM
Tokens	1,853,430	436,132	396,471	443,737	360,565	216,525
Types	30,928	16,269	10,714	13,688	12,584	7,268
TTR	1,67	3,73	2,70	3,08	3,49	3,36

Table 3. Statistics concerning the files from the MICASE corpus

1.3 Methods

The sound file of the sample text was transcribed using the freeware application Transcriber[7]. The transcription conventions were used: no punctuation, except for clitics (*aren't*, *here's* etc.); lower case, and division into tone units. In this study the negative particle 'n't' was not counted as a separate token, so 'I don't know' would constitute a 3-gram, whereas in other works (e.g. Starcke 2008: 215) it is classified as a 4-gram.

After this initial procedure, strings present in the sample text were searched for in the reference corpus using a perl script specifically developed for the purpose[8]. The procedure was as follows:

1) the sample text was tokenised by conversion to upper case, and all carriage returns, tabulation characters, double spaces etc. were eliminated;

2) the tokenised sample text was split into potential *n*-grams with an initial value of 2 ('A B', 'B C', 'C D' etc.);

3) the reference files were then taken one at a time, tokenised, and scanned for the occurrence of each of the 2-grams;

4) on finding an occurrence of a 2-gram in the reference files, the value in the slot in the array pertaining to the 2-gram and the reference file was incremented by one;

5) on completion of the scan of the reference files, total occurrences were calculated and exported to a spreadsheet containing token numbers in column 1, 2-grams in column 2, total occurrences in columns 3, and occurrences in individual files in columns 4-12;

6) *n* was incremented to 3 and the script reiterated the process from step 2 (in successive loops, *n* was set to 4, 5, 6, 7 and 8).

7) finally, all the elaborated data were saved to a single XML in order to facilitate subsequent graphic rendering.

2 Results

The results concerning the occurrences in the reference corpus of 2, 3 etc. up to 8-grams formed from the sample text display a somewhat predictable tendency in which the shorter *n*-grams are far more frequent. Nevertheless, the extent to which the shorter strings, particularly the 2-grams (79%), were seen to be present in the reference corpus is quite remarkable, especially considering the fact that the reference corpus is by no means large (3 million tokens is a paltry figure in comparison to that of the corpus used in Mason's 2008 work).

The results concerning the 2- and 3-grams, however, do not display strings which could be purported to be chains as such, possible exceptions concerning the former group being 'you know' and 'I mean', both of which play an important role as discourse markers. Examination of the sample text indicates that three occurrences of 'you know' are of this type, in two instances following other discourse markers 'well you know' and 'I mean you know'. The most repeated 3-gram was 'a little bit' with four instances in the sample text, followed by: 'a lot of', 'this is a', 'a couple of', 'you can do', 'but if you', 'little bit of', 'i i i', 'is this is', 'as you can', 'you should be', 'of course you', 'you could do', 'you you you', 'more than that', all present with two instances.

While similar to the 2-grams, in that the most frequently recurring 3 word strings do not appear to constitute recognisable units of meaning as such, some patterns of regularity start to emerge. First of all, a significant number relate to quantification: 'a lot of', 'a little bit', 'some of the', 'a couple of', 'all the time', 'the amount of', 'little bit of' and 'there's a lot'. Furthermore, there is evidence of chains which suggest the interpersonal function, such as 'I don't know', 'you have to', 'you know that', 'you know you' etc. and text organisation 'I'm going to' and 'this is what'.

When examining the *n*-grams with $n>3$, far more interesting patterns start to emerge, especially in the case of $n=4$ and $n=5$, as the numbers of results obtained with higher *n* values dropped drastically. The number of *n*-grams found in the reference corpus was 153 when $n=4$, and 43 with $n=5$. The proportion to which each file in the reference corpus contributed to these results was observed in order to verify whether patterns emerged with regards to a) geographical location, bearing in mind that the speaker in the sample text is from North America, and b) macro area, i.e. whether the sample text showed characteristics that were more common in speech styles in certain academic disciplines, rather than in others. In each case, the data were normalised to compensate for the varying numbers of tokens in the files taken into consideration. During the comparison of the macroareas,

the 'Other' category, present only in the MICASE collection, was excluded, whereas this collection of texts was included in the diatopic comparison.

With regards to the place of origin of the texts, the differences between the correlation of the sample text with the collections made on both sides of the Atlantic do not appear to be of any significance. The 4-grams showed a small bias towards the North American corpus (BASE = 47.5%; MICASE = 52.5%); while the difference at the level of the 5-grams, showed a swing, albeit negligible, towards the Old Continent (BASE = 50.6%; MICASE = 49.4%).

On the other hand, observance of the degree to which texts from the various macro-areas constituted the results displayed an interesting pattern, which is a marked bias *away* from the Arts and Humanities. Furthermore, the macro-area which contributed most was precisely that to which the sample text belongs: the Physical Sciences, and the data for both 4- and 5-grams, taken as a whole, seem to align along a continuum with the so-called 'hard sciences' at one extreme, passing through the 'soft' sciences, to arrive at Arts and Humanities, the disciplines that are commonly thought to be polar opposites of subjects such as Mathematics and Physics. Table 4 shows these data as percentages.

	Arts and Humanities	Physical Sciences	Social Sciences	Life and Medical Sciences
4-grams	18,4%	31%	23,9%	26,7%
5-grams	18,2%	31,8%	22,9%	27,1%

Table 4. The proportions to which the macro areas contributed to the total count of occurrences

Observation of the individual *n*-grams is clearly the case here in order to investigate what could explain such a marked pattern. To the contrary of what may be expected, these are not composed of items such as 'exceed the speed of light' or 'calculate the gravitational power of', that would be clearly attributable to the specific field of discourse[9]. To the contrary, only one out of the 153 4-grams could be identified in any way as being concerned with Physical Sciences: 'the energy of motion'; and none of the 5-grams displayed such a characteristic. Table 5 shows the top 40 occurrences in the reference corpus of 4-grams from the sample text. Note that two *n*-grams occurred twice in the sample text.

	4-gram	R	S		4-gram	R	S
1	I'M NOT GOING TO	223	1	21	AND THEN IF YOU	29	1
2	I WAS GOING TO	190	1	22	I DON'T KNOW BUT	29	1
3	A LITTLE BIT OF	179	2	23	IS THIS IS A	24	2
4	THERE'S A LOT OF	152	1	24	A FEW YEARS AGO	24	1
5	THIS IS THIS IS	126	1	25	KNOW THAT THIS IS	21	1
6	I'M GOING TO DO	99	1	26	WHAT ARE YOU DOING	20	1
7	I THINK IT WAS	89	1	27	I I I I	20	1
8	YOU HAVE TO DO	81	1	28	TO ASK YOU TO	19	1
9	THAT YOU HAVE TO	74	1	29	ASK YOU TO DO	18	1
10	SHOULD BE ABLE TO	72	1	30	IN A LITTLE BIT	18	1
11	I MEAN YOU KNOW	68	1	31	YOU CAN DO A	17	1
12	TO GET RID OF	62	1	32	SO YOU GET THE	17	1
13	NOW I'M GOING TO	44	1	33	YOU CAN DO A	17	1
14	YOU SHOULD BE ABLE	41	1	34	THAT THIS IS NOT	16	1
15	A LITTLE BIT AND	38	1	35	LITTLE BIT OF A	15	1
16	GOING TO ASK YOU	38	1	36	YOU KNOW IF YOU'RE	15	2
17	THIS IS NOT A	37	1	37	AND SO YOU GET	12	1
18	IT A LITTLE BIT	36	1	38	A LOT OF ENERGY	12	
19	THAT A LITTLE BIT	33	1	39	THAT YOU DON'T REALLY	11	1
20	THAT YOU CAN DO	29	1	40	NOT GOING TO ASK	11	1

Table 5. Top forty occurrences of 4-grams in reference corpus. S= number of occurrences in sample text; R= number of occurrences in reference corpus

A number of categories can be formed from the top forty occurrences of 4-grams in the reference corpus. Perhaps of least interest for our purposes are those which regard quantification (3, 4, 15, 19, 22, 24, 30, 35, 38). Of greater significance are those which concern the interpersonal function (8, 9, 10, 14, 16, 28, 29, 30, 37, 40), many of which appear to contribute to directive speech acts, i.e. the speaker is soliciting physical or mental action from his hearers, the terms 'have to' and 'ask' being recurrent. The pronoun 'you' occurs frequently in other *n*-grams (20, 21, 26, 31, 32, 33, 37 and 39), however, in this case it is more difficult to attribute these chains to the interpersonal function, as they appear to be uses of the general pronoun, rather than that referring to second person singular or plural. The highest scoring category is that concerning text organisation (1, 2, 6, 13). For reasons of space we are unable to examine the context of each of these 5-grams in the reference corpus, but taking just the most

frequent 'you should be able to', the most common R1 collocates are: 'do' (7); 'answer' (2) and near synonyms 'respond' (1), 'give [definitive yes or no answers]'; and 'follow' (2) and near synonyms 'distinguish' (1), 'figure [this out]' (1) and 'tell [the difference]' (1).

	5-gram	R		5-gram	R
1	YOU SHOULD BE ABLE TO	40	11	I I WAS GOING TO	3
2	THIS IS THIS IS A	20	12	A FEW YEARS AGO AND	3
3	GOING TO ASK YOU TO	16	13	YOU SHOULD BE FAMILIAR WITH	3
4	A LITTLE BIT OF A	14	14	A LITTLE BIT AND THEN	3
5	NOT GOING TO ASK YOU	10	15	SHOULD BE ABLE TO UNDERSTAND	2
6	TO ASK YOU TO DO	10	16	DON'T REALLY HAVE TO KNOW	2
7	I'M NOT GOING TO ASK	8	17	BE ABLE TO UNDERSTAND THIS	2
8	THAT YOU HAVE TO DO	7	18	THE AMOUNT OF WORK THAT	2
9	IN A LITTLE BIT OF	4	19	DRINK A LOT OF WATER	2
10	THAT THIS IS NOT A	4	20	YOU KNOW THAT THIS IS	2

Table 6. Top twenty occurrences of 5-grams in reference corpus. R= number of occurrences in reference corpus (none occurred more than once in the sample text)

Examination of the results concerning the 5-grams (Table 6) shows a marked increase in the presence of chains concerning the interpersonal function, which account for more than half of the twenty most frequent occurrences in the reference corpus (1, 3, 5, 6, 7, 8, 13, 15, 16, 17 and 20). Their involvement in directive speech acts is even more evident, with 'ask' and 'have to' being again recurrent, however, a number of chains appear in which the speaker is checking or appraising his hearers' understanding of the topics dealt with (1, 13, 15, 16, 17 and 20).

The numbers relating to the *n*-grams with $n>5$ decrease drastically, both in terms of the chains found in the reference corpus, and the extent to which they are present therein. Table 7 shows the results for 6-grams together with their distribution across the disciplinary groupings. With regards to the latter, the proportions, from highest to lowest are: Social Sciences (8), Physical Sciences (6), Arts and Humanities (5) and Life and Medical Sciences (3), however, the numbers are so low that they can hardly be considered significant. From a diatopic point of view the chains were seen to be present in comparable numbers in both BASE (13) and MICASE (11).

	6-gram	TOTAL	AH	LM	PS	SS	OT
1	GOING TO ASK YOU TO DO	9			2	5	2
2	I'M NOT GOING TO ASK YOU	7	2		3	2	
3	NOT GOING TO ASK YOU TO	4	1	1	1	1	
4	THAT YOU DON'T REALLY HAVE TO	1		1			
5	YOU SHOULD BE ABLE TO UNDERSTAND	1		1			
6	FOR I THINK IT WAS THREE	1	1				
7	KNOW THAT THIS IS NOT A	1	1				

Table 7. All occurrences of 6-grams in reference corpus and disciplinary provenance (none occurred more than once in the sample text)

On the other hand, the chains themselves are strikingly similar, in that the five most frequent concern the speaker directly addressing his hearers, three of which express lack of obbligation (2, 3, 4), while one is an appraisal of his listeners' knowledge. Again we may examine the reference corpus to observe the context of the most frequently occurring chain, 'going to ask you to do'. Two of the concordance lines are preceded by a negative particle, indicating a lack of obbligation. Five concern activities to be done during the lecture/seminar e.g. 'I'm *going to ask you to do* something which has a solution on the back of the sheet' and 'I'm going to play you a little clip and I'm *going to ask you to do* one thing', whereas two concern planning of the learning activity over subsequent weeks, e.g. 'what I'm *going to ask you to do* is, get through as much of the reading as you can' and 'so that's what I'm *going to ask you to do* is at least for the next three weeks which is how long we'll spend, on Marx'.

The final results complete this pattern. Two 7-grams, that are overlapping in the sample, were found in the reference corpus: 'I'm not going to ask you to', 3 occurrences (AHB, PSB, SSM) and 'not going to ask you to do', 1 occurrence (PSB). The sole 8-gram to be covered is merely a concatenation of the two 7-grams 'I'm not going to ask you to do' (PSB).

2.1 Linear analysis of results

The results outlined in the previous section are of a vertical type, i.e. extrapolating and highlighting those aspects which are deemed to be most significant, predominantly on the basis of frequency. However, as noted in the introduction, the viewing of the sample text and correspondences in the

reference corpus in a linear fashion may help to underline variability in the levels of formulaicness of various stretches of text. In order to do this, all the results were collated into a single XML file in order to allow graphic rendering by way of a custom-built application developed using Adobe Flash 8. On loading the data at runtime, the application lays out the sample text, token by token, on the horizontal axis. Above each token are seven boxes, representing the tokens presence in the *n*-grams found in the reference corpus with *n* values from 2 to 8. This presence is rendered on a continuum from white (no presence), through four shades of grey, to black (highly significant presence). The frequency data were weighted on the basis of *n*, to account for the inevitably higher frequencies for the lower *n* values. For example, with *n*=2, the following weights were applied: f>0 and f<10; f<100; f<1000; f<10000; and f>=10000. With *n*=5, far lower frequency parameters were applied: f<2; f<5; f<10; f<20; and f>=20. Furthermore, each box at each level was shaded on the basis of an average value of partecipation in *n*-grams to the left and right. For example, with a hypothetical string 'A B C D E', the box above 'C' in the 2-gram row would be shaded on the basis of the average of the frequencies for 'B C' and 'C D'; in the 3-gram row, the shading would indicate an average of 'A B C', 'B C D' and 'C D E', and so forth. While this does 'blur' the exact rendering of the results somewhat, it is, nevertheless, the only way to display the data in a easily readable format for multiple *n* values.

This method is hence not ideal for identifying boundaries between recurrent chunks which are adjacent in the text: traditional frequency-based analyses, as exemplified above, prove far more suitable for this task. On the other hand, it is useful for contrasting chains with high frequency values with those present in low frequencies, or indeed those that are not found at all, in the reference corpus. In other words, this purports to isolate segments which may be viewed as entailing the open choice model, by comparison with others that are seen to be of a routine nature in the field of discourse under examination. By its very nature, such horizontal rendering of data produces rather voluminous results, and for reasons due to space limitations, only two segments (c. 44 tokens long) from the sample text will be illustrated and discussed. For exemplary purposes we have deliberately chosen segments including longer chains that were seen to be significantly present in the reference corpus as discussed above.

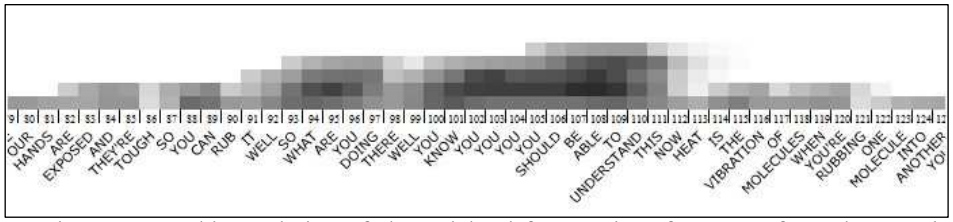

Figure 1. Graphic rendering of the weighted frequencies of *n*-grams from the sample text in the reference corpus. The rows display *n* values from 2 to 8, starting from the lowest. Tokens = 80 to 125

Figure 1 shows the results from tokens from 80 to 125. Observing the lowest row, that rendering the presence of 2-grams, it may be seen that these tend to flow together in a more or less uninterrupted fashion, constituting a sort of 'background noise' of low significance. The 3-grams are slightly more informative, showing null presence at the start and end of the segment, and two peaks: at tokens 95 and 108. Discrimination starts in earnest at the higher tiers, from *n*=4 onwards, where three high frequency chains emerge: 'what are you doing', 'you you you you' and 'should be able to'. At *n*=6, the upper limit of coverage is reached, underlining however a three chains flowing together to form 'you should be able to understand this now'. Of interest is the sharp drop in coverage evident at token 113, suggestive of a relatively sharp interruption of the preceding and successive chains. In fact, from a textual point of view, this token, 'heat', displays a shift in function from the interpersonal to the informative: the speaker has finished commenting on his listeners' ability to grasp the content, and proceeds to provide a textbook definition of 'heat' (note that 'now', in the sound file, belongs to the preceding tone unit, and hence is to be interpreted as temporal deixis, rather than a discourse marker).

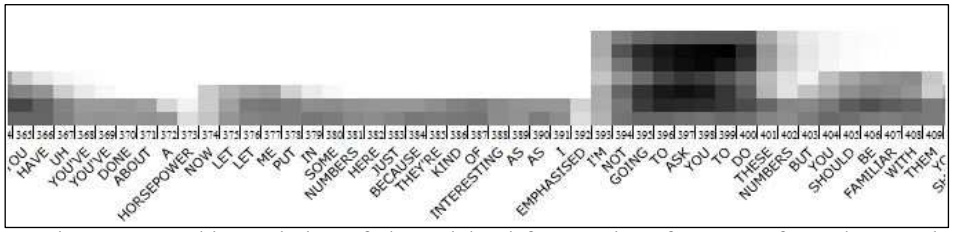

Figure 2. Graphic rendering of the weighted frequencies of *n*-grams from the sample text in the reference corpus. The rows display *n* values from 2 to 8, starting from the lowest. Tokens = 365 to 409

The second segment to be analysed can be seen in Figure 2. In this case an even more extreme pattern emerges. The first tokens form part of a sequence of message fragments: pronoun+auxiliary, followed by filler 'uh'

and a double repetition of the pronoun+auxiliary in contracted form. The speaker here is possibly doing a mental calculation, for which he needs to 'buy time', in fact, the phrase is completed with a hedge, 'about', indicating that the calculation is probably by no means precise. The dearth of coverage for 'horsepower' suggests that this is a potential break from the preceding routinely formed language. Token 374, 'now', introduces a stretch with moderate to low coverage, which continues in a relatively homogenous fashion up to token 392, 'emphasised'. In order to verify whether the gap which appears at this point could actually be attributed to orthographic variation, 'emphasised' v. 'emphasized', the reference corpus was consulted using WordSmith Concord, and indeed seven occurrences were found with '-ise' and 17 with '-ize' in BASE; whereas the sole occurrence in MICASE was with the latter spelling. Hence, trascription inconsistencies in the reference corpus, due not only to diatopic factors, may skew the data somewhat. Nevertheless, the following tokens form an unmistakable chain: 'I'm not going to ask you to do'. True, this was attested as the only occurrence of an 8-gram in the frequency-based analysis above, yet this rendering highlights how significant it actually is in comparison to the co-text. Furthermore, it also suggests that the chain blends into another chain that appears to start from token 403 onwards: 'but you should be familiar with', with a considerable decrease in frequency constituted by 'these numbers'. Therefore, rather than merely isolating high-frequency chunks, the data may also be suggestive of the presence of p-frames, such as: I'M NOT GOING TO ASK YOU TO DO [X] BUT YOU SHOULD BE FAMILIAR WITH [X][10].

Conclusions

This study constitutes an attempt to approach the corpus-based analysis of a short sample of Spoken Academic English from two angles. The first concerns the adoption of mainstay 'vertical' techniques involving the identification of the strings of tokens of various lengths in the sample text that were seen to be frequent in greatest numbers in the reference corpus. The second involves rendering the results in a holistic, linear fashion in order to highlight possible variations in the frequency of occurrence of different segments of the sample text.

With regards to the former, somewhat predictably, the number of results was heavily dependent on the length of these *n*-grams, ranging from c. 80% coverage at $n=2$, to the occurrence of a sole 8-gram in the reference corpus, suggesting that 8 is the upper threshold for which significant results may be

found. The results, particularly at higher values, suggested that the chains in the sample text found at most significant levels in the reference corpus were mainly those involved in the interpersonal function, and to a lesser extent, the text organisational function. In other words, the most significant formulae identified in the sample text, by way of comparison with a large body of texts from a wide range of academic disciplines, were those which dealt with: telling hearers what they are expected to do and what they don't have to do; checking the hearers' understanding; and linking preceding with forthcoming sections of the discourse. Of particular interest is the breakdown of the frequency figures for n=4 and n=5 into their disciplinary provenance: representation was notably higher in the Physical Sciences, precisely the domain to which the sample text belongs. The grouping with the lowest representation was that composed of lectures and seminars in the Arts and Humanities. These findings suggest that certain types of interaction (essentially appearing to deal with the management of tasks) are more common in the former, rather than the latter, disciplines. Finally, the sub-division of the results on a geographical basis (Great Britain v. the USA) revealed no significant differences.

The second approach adopted displays some aspects which are methologically innovative, using original software to illustrate in a linear fashion the high degrees of variability in the frequency with which segments in a sample text are present in a reference corpus. The findings are highly supportive of Sinclair's model of the idiom principle, in that the tokens of the sample text are seen to concatenate on an extensive basis at the level of 2-grams and 3-grams. Interruptions in these sequences may take two forms: occasional gaps in coverage are indicative of points where the open choice model may come into play; at other points coverage extends to far longer chains, and a substantial number of 4-, 5- and 6-grams highlighted the routine nature of these stretches in academic discourse. Furthermore, evidence was observed of long chains which concatenated with high levels of coverage at both extremes and slight decreases in the middle, which may reveal the presence of chains, i.e. formulaic stretches including one or two slots where variation may occur.

Such methodology is not of use only to the linguist, as it may also prove useful to materials writers and foreign language instructors working on specialised courses on academic listening and speaking. Given a sample text, it is possible to identify on the basis of clear parameters which segments are most likely to be heard in other academic texts, thereby allowing the writer or instructor to focus on the textual, pragmatic and prosodic features of the type of language that learners will come into contact with in their future academic careers.

Notes

[1] Stubbs (2005: 5) defines an "*n*-grams" as "a recurrent uninterrupted string of orthographic word-forms".

[2] A video clip of the lecture can be watched at: http://www.youtube.com/watch?v=iGMVQU3sp1s [last visited 30/04/2010].

[3] Goldman-Eisler (1968) defines the medium rate of articulation for English speakers as being between 4.4 and 5.9 syllables per second.

[4] See http://www2.warwick.ac.uk/fac/soc/al/research/collect/base [last visited 30/04/2010].

[5] See http://quod.lib.umich.edu/m/micase/ [last visited 30/04/2010].

[6] Citation from http://micase.elicorpora.info/researchers/about-micase [last visited 30/04/2010].

[7] See http://trans.sourceforge.net/en/presentation.php [last visited 30/04/2010].

[8] Danielsson (2004) identifies perl as being an ideal programming language for the development of *ad hoc* tools for corpus analysis and provides some basic examples of useful scripts.

[9] Biber (2009: 289), in describing lexical bundles in academic prose and conversation, outlines a continuum with "multi-word collocations" at one extremity, and "multi-word formulaic sequences" at the other. The former are typically of a technical nature, being composed solely of lexical/content words, often extended noun phrases, with high Mutual Information (MI) scores, but low frequency ratings. The latter, featuring both content and function words have lower MI scores, but are seen to be far more frequent. The results for the 4- and 5- grams in this work show a distinct bias towards the latter extremity.

[10] Stubbs (2005: 5) uses the term "p(hrase) frame" to describe a "a recurrent n-gram with one variable lexical slot".

Bibliography

Biber, D., 2009 "A Corpus-driven Approach to Formulaic Language in English", in *International Journal of Corpus Linguistics*, XXIV, n. 3: 275-311;

Danielsson, P., 2004 "Simple Perl Programming for Corpus Work", in Sinclair, J.M.H. (ed.), *How to Use Corpora in Language Teaching*, John Benjamins, Amsterdam-Philadelphia: 225-248;

Dudley-Evans, A., 1994 "Variations in the Discourse Patterns Favoured by Different Disciplines and Their Pedagogical Implications", in Flowerdew, J. (ed.), *Academic Listening: Research Perspectives*, Cambridge University Press, Cambridge: 146-158;

Goldman-Eisler, F., 1968 *Psycholinguistics. Experiments in Spontaneous Speech*, Academic Press, London;

Hoey, M., 2005 *Lexical Priming: A New Theory of Words and Language*, Routledge, London;

Mason, O., 2008 "Stringing Together a Sentence: Linearity and the Lexis-syntax Interface", in Gerbig, A., Mason, O. (eds.), *Language, People, Numbers: Corpus Linguistics and Society*, Rodopi, Amsterdam: 231-248;

Sinclair, J.M.H., 1991 *Corpus, Concordance, Collocation*, Oxford University Press, Oxford;

Starcke, B., 2008 "Differences in Patterns of Collocation and Semantic Prosody", in Gerbig, A., Mason, O. (eds.), *Language, People, Numbers: Corpus Linguistics and Society*, Rodopi, Amsterdam: 199-216;

Stubbs, M., 2005 *The Most Natural Thing in the World: Quantitative Data on Multi-word Sequences in English*, paper presented at Phraseology 2005, Louvain-la-Neuve, 13-15 October 2005.