_lognostics
tools for vocabulary research

Tintin and the World Service: a look at lexical environments.
Paul Meara

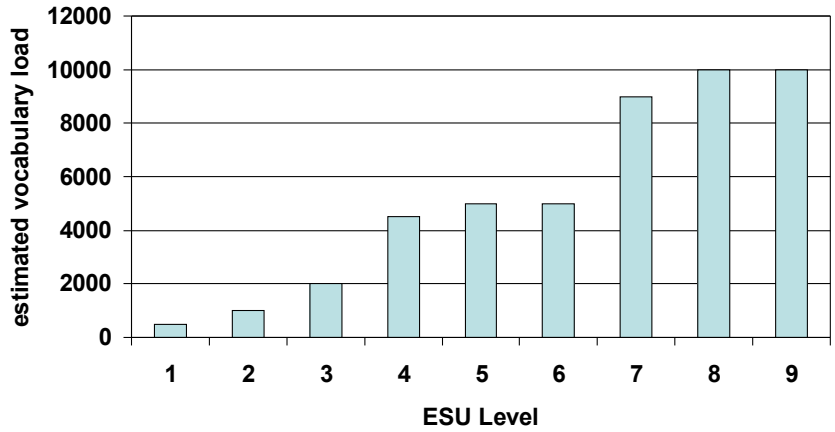
A version of this paper was given as the Hornby Trust Lecture at IATEFL in Swansea. I never actually met A.S. Hornby, but I did hear him talk once in London. At one point during his talk, he mentioned how impressed he had been by some Chinese learners that he had recently met. These learners had had no real access to English, other than classes in China, but in spite of this they were extraordinarily fluent in English. Hornby had asked how they managed to be so good, and was surprised by the reply: young learners simply learned songs off by heart, while more advanced learners used the same method with rather more complex texts. Hornby paused, smiled, and wondered whether there might perhaps be something in these apparently unfashionable methods. I remember thinking how awful it would be if he was right.

Chickens like this have a nasty habit of coming home to roost at unexpected times, and we'll come back to this particular chicken later. Before that, however, I want to describe a project that I carried out for BBC English in 1991.

BBC English, the English teaching arm of the BBC's World Service, have for some time been working on a BBC Core Curriculum - a set of standards and objectives that could be used to describe the materials distributed by the BBC. Broadly speaking, this curriculum was intended to be linked with the English Speaking Union's Framework (Carroll and West 1989). This framework describes 9 levels of proficiency in English, ranging from level 1, beginner, to level 9, mastery. Each of the levels comes with a detailed description of the types of behaviour that one can expect of learners at that level. The description for Beginners, for instance, reads: "knows a few words or phrases such as *good morning* and can understand some public notices or signs. At the lowest level can simply recognise which language is being used." The description for lower-intermediate students, level 4, reads: "has a basic range of English sufficient for familiar and non-pressurising situations. Weaknesses in accuracy, fluency, appropriacy and organisation mean that communication and comprehension is restricted." Carroll and West were chiefly interested in providing a common framework for describing English Language examinations, but their scheme is obviously more general than this, and it is easy to see how it could be applied to grading text-books and other similar materials. At the time they approached me, BBC English had already asked Felicity O'Dell to produce a grammatical syllabus based around the framework. The task I was given was to look at the lexical implications of this work.

At first sight, this looked like a relatively easy task. My research students and I have been working for some time on a set of standardised vocabulary tests. One of the advantages of these tests is that they are extremely simple, and we can test huge amounts of vocabulary in a very short space of time. The tests are good enough that they allow us to make a rough estimate of how big a learner's vocabulary is. (Meara and Jones 1988, 1990). Generally speaking, learners with big vocabularies turn out to be better at most language tasks than learners whose vocabularies are quite small: certainly, as far as listening comprehension goes - the main focus of interest in BBC English - there is a fairly close relationship between vocabulary size and performance in English as a foreign language. Our experience with the vocabulary tests suggested that it ought to be fairly straightforward to produce rough lexical guidelines that would link in with the ESU framework. As a rule of thumb, we reckon that students at First Certificate level score around 3500 on our test. Students at Proficiency level score around 7000 words. Quasi-native speakers score around 9500 words, which is almost 100% on our tests. Given these rough guidelines, we thought it would be a straightforward task to analyse the vocabulary load of any given BBC English broadcast, and to relate it to the expected level of the intended listeners. The exact nature of this relationship was something that we thought would emerge from the research, but our expectation was something like the sketch in figure 1. Most basic courses, (levels 1-3) seem to rely on a very small vocabulary of around 2000 words. After that, there is a sudden jump to around 5000 words for levels 4-6. Then another sudden jump in difficulty seems to occur around level 7, what the ESU framework refers to as "advanced".

Figure 1
How vocabulary size relates to the ESU Levels



At this stage, we were interested in finding a way of measuring the lexical load of a broadcast, and relating this lexical load measure to what we knew about the vocabulary size of the intended listeners. Surprisingly, perhaps, there isn't an agreed way of measuring lexical load. There is some theoretical work on lexical richness, which is not quite the same thing (cf. Ménard 1983), but this work is largely in French, and it isn't clear that it applies straightforwardly to the kind of English that learners are typically exposed to. So in practice, we had to devise our own ways of measuring lexical load, and after some discussion we decided to measure two different aspects of the texts we were working with. These measures were (a) how many different words did each text contain, and (b) how difficult were they?

At first glimpse, these questions look as if they are really straightforward. In fact they aren't. The first problem that you are faced with is what do you count as a word? Do HAPPY, UNHAPPY, HAPPINESS and UNHAPPILY count as four different words, or just one: HAPPY? What about BE, AM, ARE, IS, WAS, and so on? At the time, I was heavily influenced by some ideas about word families, which suggested that sets like this ought not to be treated as instances of separate words, but reduced to their basic forms (Nagy et al 1989). I also felt that we ought to be looking at word TYPES rather than word TOKENS in our analyses. In English, a small handful of words accounts for a very large proportion of what we hear and read in everyday language. The figure usually quoted is that some 2000 words account for about 80% of the words we meet. But this figure only holds if you count every occurrence of a word - each word TOKEN - separately. That is because some very common words, like A, THE, IS, occur over and over again in any text. Rare words, like RHEUMATISM or GLIDER are not nearly so common, and once they have occurred once in a text, they will often be replaced by pronouns or phrases for stylistic reasons. If you count each different word only once - each word TYPE - then you get a rather different picture. We decided that we were interested in using word types as an index, and not word tokens, because it should be easier to distinguish between texts in that way.

However, we were still left with the problem of deciding how difficult these word types were. Again, rather surprisingly, there is no agreed scale of difficulty for vocabulary in EFL. There is a general agreement that rare words are more difficult than common ones, and this assumption often finds its way into textbooks as a way of structuring the vocabulary to be taught. Even then, however, things are not straightforward: some very rare words in English are actually very simple for speakers whose L1s share vocabulary with English. Spanish and Italian speakers, for instance, will often find rare words in English easier than common words, because of the way English has borrowed words from Latin. In the end, we decided to use a word list produced by Paul Nation to judge the difficulty of our words (Nation 1986). Nation's lists are based on earlier frequency counts, but take other considerations into account as well. They fall

Table 1: examples of words from Nation's 1986 lists

Nat0 a all another any both each either enough every ... noun verb vocabulary lesson ... one two three ...

Nat1 about accent accident act action add address advertise ...

Nat2 able abroad absent absolute absorb abstract accept access...

Nat3 abandon abolish abrupt absurd academic accelerate ...

Nat4 anvil funnel rheumatism dregs porridge ...

into four bands, which are illustrated in Table 1.

Nat0 comprises a set of about 500 words, mostly function words, or closed class words, of high frequency. This set includes prepositions, pronouns, quantifiers, numerals, days of the week, months, common greetings. It also includes a set of words like NOUN, VERB, SENTENCE, LESSON, CLASS and so on. These words are not particularly common in the language as a whole, but they are used very frequently in language learning contexts. Nat1 is a set of approximately 1000 words, the words most likely to be encountered by learners outside the Nat0 list. Nat2 comprises approximately the second thousand most frequent words in English. Nat3 is based on a list of words which occur very frequently in academic contexts - basic scientific words, discourse connectives, and so on. Nat4 is my term for items which do not occur in any of the earlier lists.

We planned to use this simple classification as a way of looking at the lexical difficulty of BBC English broadcasts.

BBC English provided a set of computer discs containing transcripts of a number of programmes, and in theory, all we needed to do was to analyse these transcripts in terms of their lexical complexity. In practice, things were not that simple. The actual transcripts looked like the extract in Table 2. It will be immediately obvious that the raw transcripts contain a lot of information that is important to the producer, but not to the person listening to the programme. All this information, and the "stage directions" had to be stripped out by hand, as the transcripts were not consistent enough for it to be done automatically. Once this process was completed, we had a stripped transcript which COULD be processed automatically, and we wrote a set of computer programs to do that for us. The program read each transcript, and worked out how many different words

Table 1
Examples of words from Nation's 1986 lists.

Nat 0	a all another any both each either enough every noun verb vocabulary lesson one two three
Nat 1	about accent accident act action add address advertise
Nat 2	able abroad absent absolute absorb abstract accept access
Nat 3	abandon abolish abrupt absurd academic accelerate
Nat 4	anvil funnel rheumatism dregs porridge

Table 2
Extract from a raw transcript

BBC ENGLISH BY RADIO

OLYMPIC ENGLISH

PROGRAMME 2

Written by: Don Anthony (TO) Producer: Hamish Norbrook (S) RPA Lorraine Selwyn (S)

Taking Part: Don Anthony (TO) Amanda Carlton (REP) V1: Barrie Shore (REP)
V2: Simon Parish (REP)

Tape No: 8R/28/K002/K Tx Date: Saturday 21 May 1988. Recorded: Friday 5 February 1988.
Studio C27 (10-30-1245) Duration 12'20"

SIG: DUR: 0'20

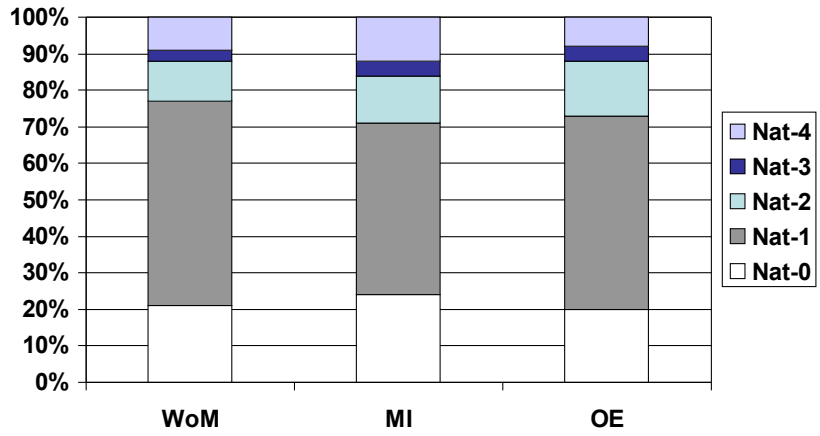
V1 "We present Olympic English. V2 English for the 24th Olympiad. V1 Olympic English. With Don Anthony and Amanda Carlton.

AMANDA: (FADE UP, & DOWN SOUND OF SKIPPING) 37,38,39,40, (COUNTS TO 51)

DON: Hello and welcome back to the second programme of Olympic English. Amanda is over in the corner of the studio trying today's exercise. We'll join her in a moment...

each one contained. Another program reduced these word lists to word families, and a third program compared the word family list to Paul Nation's frequency list. The result was a lexical profile for each series. Typical program profiles look like Figure 2.

Figure 2: Lexical profiles of three BBC English series



The first profile, WoM, comes from a series called *The World of Musicals*. Each program in this series briefly describes the story of a musical, and discusses some of the themes it relates to. The profile in Figure 2 shows that just over 20% of the different words in the text come from category Nat0. Just over half the words come from the Nat1 list. A further 10% come from the Nat2 list, while only a handful of words come from the more technical Nat3 list. Only about 10% of the different word types appearing in this transcript are "difficult" in the sense that they were classed as Nat4 words by our program. Each broadcast in this series contains about 260 different word types, and each broadcast lasts for about 13 minutes. 10% of difficult words therefore works out at approximately 26 difficult words in each broadcast, or just about 1 difficult word every 30 seconds. In terms of running text, this works out at about 1 difficult word in every 200 words. To my mind, this figure is rather low, much less than one might expect in a demanding written text. Broadcasts with this level of lexical difficulty ought to be easy to listen to, as long as the listener is familiar with the basic vocabulary of English. This should mean that *The World of Musicals* ought to be a series that could be handled by listeners with a fairly low level of English, say around level 4 or 5 of the ESU Framework.

The question we can ask now is this: are programme series different in the number of difficult words they present to the listener? The short answer is no, they are not, and you can see this in the other two profiles in Figure 2. These figures show lexical profiles for

two different series broadcast by BBC English. These are *Olympic English*, which deals largely with sport, and *Mission Improbable,* a soap-opera format, in which two characters find themselves in a number of very improbable situations that enable them to discuss the finer points of English grammar. We actually looked at 15 different series, some judged easy by BBC English, others judged more difficult. For each series, we did a detailed analysis of five programmes, and averaged the results. In all cases, the end results looked very much like the data in Figure 2. There was very little difference between any of the series, in fact, with difficult words accounting for around 10% of the total, pretty much irrespective of what the programmes were about. Some series hadslightly more difficult words on average, but individual programmes within a series varied more than the series averages did.

This result came as something of a surprise, and it really brought the project to an abrupt halt. On the face of it, there was nothing to distinguish the lexical profile of an easy programme from a more difficult one: they all had roughly the same proportion of difficult words. In fact, some of the programmes that were aimed specifically at teachers had a slightly simpler lexical structure than programmes aimed at ordinary learners.

Of course, although the number of difficult words in our texts is quite low, they are VERY important. You can see this from the list of words in Table 3, which shows a list of Nat4 words from one broadcast.

Table 3 which musical do these words occur in?

accent amaze award awful bet bloody candle chauvinist cockney delight dialect enchant gutter playwright mud professor rhyme screech triumph utter violet

Of the three series in Figure 2, you can tell immediately that this programme is NOT an *Olympic English* broadcast, since it contains no sports vocabulary at all. It is actually a broadcast from *The World of Musicals* series, and if you know anything at all about musicals, you will have no difficulty in identifying exactly which musical is being discussed here. The answer is at the end of the paper. What these data show is that a great deal of the the content of a broadcast is carried by a relatively small number of words.

Earlier, I commented that the number of difficult words in a broadcast generally worked out at about one every 30 seconds. In fact, things are more complicated than that. Some of the words that our programs class as "difficult" may in fact be words that the listener knows already, even if s/he is not very advanced. Not many learners acquire the 2000 most frequent words in English and no others! Even in cases where the vocabulary is

genuinely new, a lot of help is provided for the listener. The broadcasts often pick up difficult vocabulary, repeat it several times, and often explicitly tell the listeners what these words mean. For instance, if you have a tape of John Black, the Smith, describing how he makes horse-shoes, you might get a dialogue that goes something like this:

> J.B. ...so when the iron is hot, I take it over to the anvil and hammer it.
> Int: Anvil. Anvil. That's the big piece of iron that you put the horse-shoe on and hit it.
> J.B. Yes, the anvil.

This gives you four repetitions of **anvil**, in the space of a few lines of text. In addition, parts of this explanation may also get repeated a second time as the presenter goes over the recorded material in the broadcast. This type of repetition obviously has the effect of reducing the psychological load of the new vocabulary, and effectively reduces the figure of 10% difficult words to very much lower levels.

Another factor which contributes to lessening the psychological load of the new vocabulary is that many of the difficult words that occur in a series occur in a number of different programmes. This means that the cumulative new vocabulary is much smaller than the raw figures imply. You can see this in Figure 3.

Figure 3:
difficult vocabulary in the series *Pop Talk*

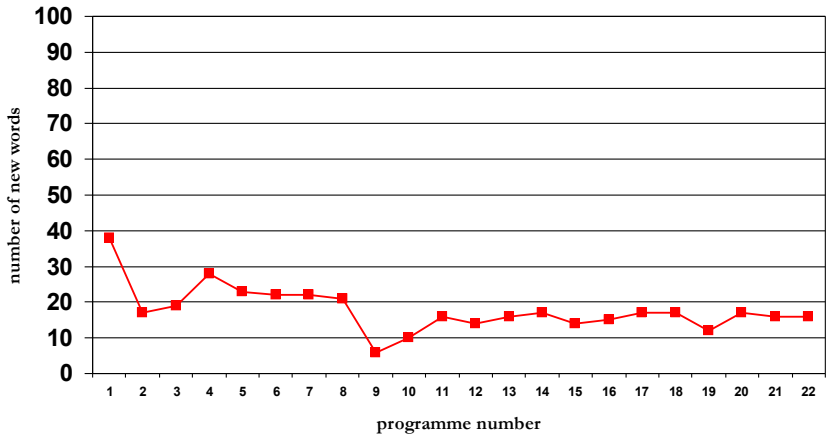


Figure 3 shows data from a series called *Pop Talk*, in which two interviewers discuss a range of topics with a well-known pop singer. The figure shows how many of these "difficult" words occurring in each broadcast are genuinely new in the sense that they haven't occurred before in the series. Obviously, "difficult" words like **album, gig,**

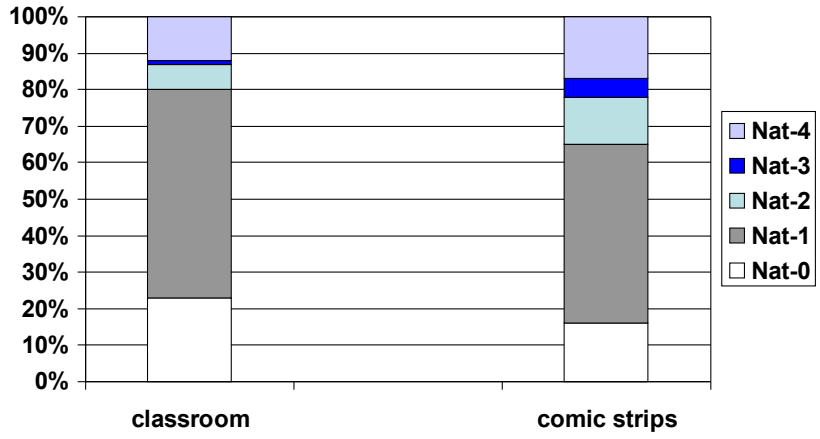
studio, producer, and other items which are important in the recording industry, occur in all the programmes. If we only count as new words words that have not occurred in any of the previous broadcasts, then the vocabulary load of each series tends to get lower as the series progresses. For *Pop Talk*, for instance, by the time we get half-way through the series, each programme introduces only 15 difficult words in each broadcast - a maximum of about one difficult word a minute, or about one difficult word in every 200 words of running text.

With figures as low as this, it is would be easy to conclude that lexical difficulty is not making a significant contribution to the difficulty of BBC English broadcasts. My hunch is that this conclusion is not correct, however. The programs we wrote to do the analyses reported here make a number of assumptions that might not apply to real learners listening to real broadcasts. One of these assumptions is that non-beginners have no difficulty recognising simple words, and that we only need to consider "difficult" words when we measure the lexical load of a broadcast. My guess would be that many listeners actually have quite a lot of difficulty with easy words, especially if the sound quality of the broadcast is not very clear, as is often the case on short-wave radio. The second assumption our programs make is that it is sensible to reduce complex words to their base forms for the purpose of counting. I am less happy about this now than I was when we started this project. It obviously makes sense to treat SONGS as an instance of SONG, and SINGER as an instance of SING. But not all cases are as easy to resolve as these are. It may well be that some intermediate learners can tell that DISENTANGLED is a special instance of TANGLE, but some morphologically complex words are much harder to disentangle than this one. Even a quasi-native speaker would be hard pressed to sort out the meaning of ANTI-DISESTABLISHMENTARIANISM, despite the fact that its morphological structure is entirely regular. Morphological decoding skills obviously change with proficiency (and L1 background), and it may be that what counts as a single word family for a beginner is quite different from what an advanced learner would count as a single family. This idea has been developed by Ringbom (1983) and by Bauer and Nation (1993), but it obviously needs a lot more work.

In spite of these uncertainties, it is still possible to compare the output of BBC English with other types of texts, and I have done this in Figure 4, Figure 5 and Figure 6. The first profile in Figure 4, and the data shown in Figure 5 are based on transcripts of immersion English classes in Quebec. We broke these classroom transcripts down into segments of about 1500 words of teacher talk, or about 15 minutes of class time, so that each segment was roughly comparable in length to a BBC English broadcast, and then we ran the same analyses on these transcripts as we had done for the BBC data. The profiles for the classroom data and the BBC programmes are surprisingly similar. If anything, the classroom data shows a slightly higher proportion of very easy words (Nat1), and a

slightly higher proportion of difficult words (Nat4), at the expense of words of intermediate difficulty (Nat2 and Nat3). Actually, the Nat4 figure is slightly misleading, as the word counts include instances of items like **popsicle**, and a whole set of technical terms used in ice-hockey. These words don't appear in Nation's word count, but are obviously of high frequency in Canadian schools. The cumulative word count for the classroom data shown in figure 5 is also broadly similar to the data for BBC English: although the raw number of "difficult" words in each segment varies quite a lot (the upper line in Figure 5), the rate at which genuinely new words are introduced (the lower line in Figure 5) remains remarkably constant over the period covered by these transcripts, and is very close to the figure we found for *Pop Talk*. It would be nice to interpret this as a kind of convergence: there is probably an optimum rate for introducing new words to learners, and experienced teachers know instinctively what these optimum rates are. If this idea is correct, then it is very impressive that the broadcasters are able to do the same, even in the absence of immediate feedback from their audiences.

Figure 4: Lexical profile and vocabulary difficulty



By way of a contrast, we also ran a series of analyses on some simple comic strip books aimed at young readers (Hergé, 1953, 1954). These texts, which are actually English translations of French originals, turned out to be not nearly so simple as we had anticipated. A typical Tintin story contained about 15,000 words of running text, and a total of around 2000 different words. This is a very rich lexical environment, and a very large proportion of it consists of "difficult" vocabulary (see the right most p[rofile in Figure 4, and Figure 6.). One of the characters, Captain Haddock, who has a habit of using odd profanities like "Blistering Barnacles", makes some contribution to this total, but even

Figure 5:
difficult vocabulary in the classroom data

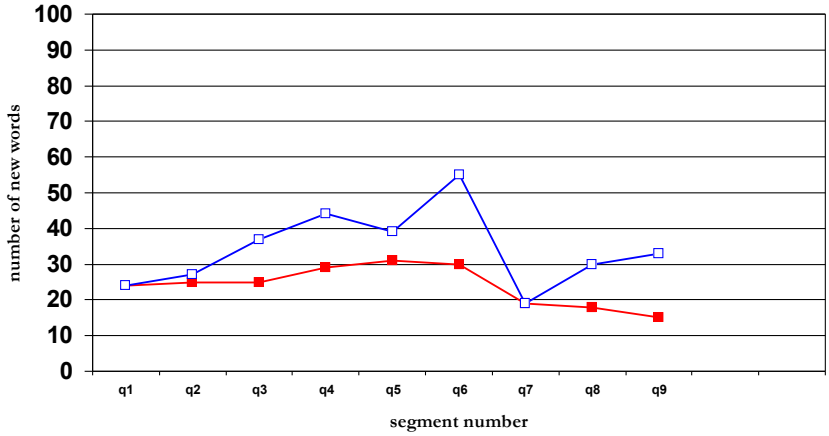
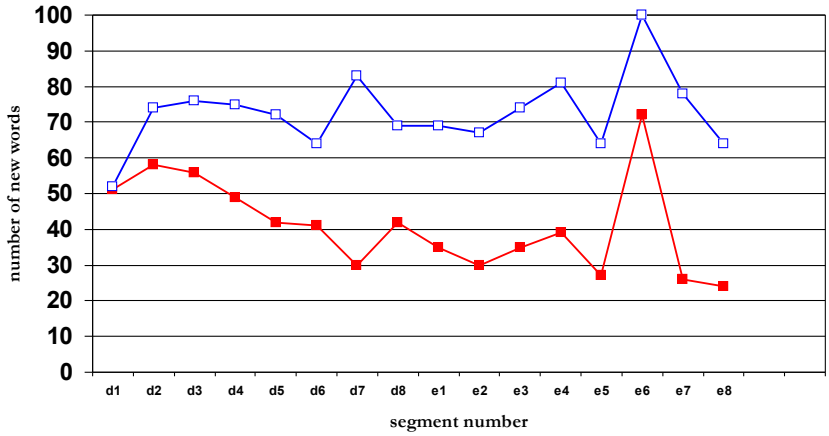


Figure 6:
difficult vocabulary in the comic strip data

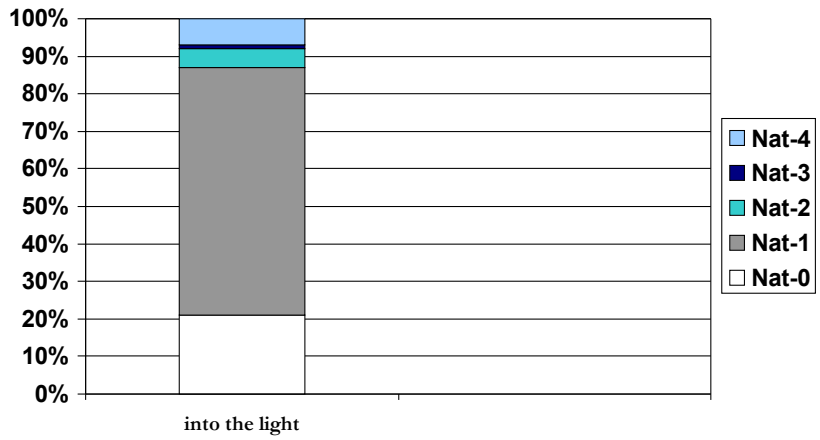


allowing for this, the texts as a whole have a difficult vocabulary in excess of 35% of the total number of different words. Breaking the texts down into 1500 word segments so that they compare with the BBC broadcasts reduces this figure a little, but we are still dealing with more than twice the number of difficult words that we find in the broadcast material. The differences are even more marked when we look at the cumulative vocabulary. Figure 6 shows that the number of difficult words per 1500 word segment ranged from a low of 52 words to a high of 100 words. The cumulative new

words figures, the bottom line in figure 6, start at 50 and decline very slowly towards the end of the text. Even at the very end of the text, the new word rate is considerably higher than anything we have met in the broadcasts or the classroom data.

Figure 7 shows a lexical analysis of a very much simpler text, a set of songs in an album by Chris de Burgh *Into the Light*. The interesting thing here is that the profile differs quite remarkably from anything that we have seen so far. The total number of words is roughly the same as the 1500 word texts that we have been dealing with throughout this paper, but here the proportion of Nat4 words sinks to less than 10% of the total of different words. So too does the list of Nat2 words. Almost all the vocabulary in this album comes from the most frequent items of English: about 20% function words and other similar items, and nearly 70% from the 1000 most frequent words of English. Readers who are familiar with de Burgh may remember that there are a couple of songs in this album which have explicitly religious themes, and use a slightly peculiar vocabulary as a result. In fact it is these songs which account for almost all of the 10% Nat4 words. This pattern is very different from the vocabulary loads that we have noted in other genres, and very much simpler than anything we found in the BBC texts.

Figure 7:
Lexical profile Chris de Burgh *Into the Light*



Conclusion

In one sense, the work we did for BBC English was something of a failure. We failed to show that there was any straightforward relationship between the lexical profiles of BBC English programmes, and the level of learners they are aimed at. As is often the case with research of this sort, though, we ended up with much more interesting questions than the ones we started out with. For me, the most important thing was that the work reinforced my view that we do not know very much about the way learners

acquire vocabulary, and the sorts of lexical environments that they operate in. Krashen (1989) has made much of the idea that teachers pitch their discourse at a level which just stretches the comprehension skills of their students, and that this stretching provides an optimum environment for acquiring new words. The data I have presented here are not incompatible with that view, but the reality is that we do not really know very much about classrooms as lexical environments. In fact, we lack even the most basic tools to investigate these questions. We do not even have an agreed word list that might be used as a reliable index of lexical difficulty. This is obviously a problem that needs to be addressed with some urgency.

The one striking fact to emerge from our analyses is that pedagogical texts - whether classrooms or broadcasts - seem to have very different lexical profiles from texts whose main aim is not pedagogical. I do not know if the de Burgh album is typical of lyrics in this genre, but it is very hard to avoid the conclusion that this type of material is ideally suited to beginners anxious to build up their grasp of basic vocabulary. Similarly, the vocabulary of the Tintin texts seems to be far richer than anything we have found in the pedagogical material. It is difficult to avoid the feeling that more advanced learners, anxious to build up their vocabulary to a level beyond basic, would benefit a lot from serious exposure to texts at this level of difficulty. These suggestions bear an uncanny resemblance to Hornby's Chinese teacher recommending that beginners should learn songs by heart, and that more advanced learners should learn whole books off by heart. I don't know whether Hornby himself would have agreed with these suggestions, but I guess that they might have made him smile.

References

Bauer, L and ISP Nation. 1993.
Word families. *International Journal of Lexicography* 6(1993), 253-279.
Carroll, BJ and R West. 1989.
ESU Framework: performance scales for English language examinations. Harlow: Longman. 1989.
Hergé. 1953.
Destination Moon. London: Methuen. translation by L Lonsdale-Cooper and M Turner. 1959.
Hergé. 1954.
Explorers on the Moon. London: Methuen. translation by L Lonsdale-Cooper and M Turner. 1959.
Krashen, S. 1989.
We acquire vocabulary and spelling by reading: additional evidence for the input

hypothesis. *Modern Language Journal* 73(1989) 440-464.
Meara, PM and G Jones. 1988.
Vocabulary size as a placement indicator. In: P **Grunwell** (ed.) *Applied Linguistics in Society.* London: CILT. 1988
Meara, PM and G Jones. 1990. *The Eurocentres Vocabulary Size Tests*. 10KA. Zurich: Eurocentres.
N**agy, W, R Anderson, M Schommer, J Scott and A Stallman .** 1989.
Morphological families in the internal lexicon. *Reading Research Quarterly* 24(1989) 262-282.
Nation, ISP. 1986.
Word Lists (revised edition). Wellington: Victoria University, English Language Centre.
Ringbom, H. 1983.
On the distinction between item learning vs system learning, and receptive competence vs productive competence in relation to the role of L1 in foreign language learning. In**: H Ringbom**, (ed.) *Psycholinguistics and foreign language learning.* Åbo: Åbo Akademi.

acknowledgements
Thanks to Patsy Lightbown who provided the transcripts of the Canadian immersion classes, and Randall Halter who helped with the analysis of this material.

answer to table 3
The musical in question was *My Fair Lady*, based on George Bernard Shaw's play *Pigmalion.*

This paper was the Hornby Trust Lecture given at IATEFL in 1993, and appeared in the IATEFL Annual Conference Report (1993) pp32-37.