Lexis for Young Learners: Are we heading for frequency or just common sense?

Thomaï Alexiou¹, Nikolaos Konstantakis²
¹Hellenic Open University, ²Swansea University
¹t_alexiou@hotmail.com, ²nikonst@hotmail.com

Abstract: There have been limited studies to show what junior level language learners are taught and why. Just how useful are the words that Greek young learners of about 7 to 8 years old have acquired by the end of their first ‘junior’ year?

The present paper presents a qualitative and quantitative analysis concerning the vocabulary used in a number of recent coursebooks for Junior A level. More specifically, the degree to which these words overlap with the first two thousand most frequent words in English is ascertained, according to wordlists from the General Service List (GSL) and the British National Corpus (BNC). In addition it reveals the topic areas within which the vocabulary of these coursebooks falls.

Key words: vocabulary, young learners, frequent/infrequent words, word lists, thematic areas

1. Introduction

Linguists and researchers have not managed so far to reach a consensus concerning the advantages of an early start in foreign language learning. The majority claim that a teacher’s abilities and/or training, the educational system, and continuity of input are of overall importance, but they all seem to agree that there is no harm in starting young. In particular, in terms of vocabulary, there have been limited studies to show what junior level language learners are being instructed and why. As yet, there is a vague picture of the kind of vocabulary that Greek young learners in particular are being taught.

How much vocabulary does a second language learner need? Nation (1990) suggests that the most frequent 2,000 words are essential for any language use, and, according to Nation & Waring (1997:11), a second language learner needs 3,000 high frequency words. More specifically, when reading, knowledge of 3,000 words provides a coverage of at least 95% of a text (Liu Na and Nation 1985). On the other hand, knowledge of 2,000 words gives a coverage of only 80% in a text which means that one word in every five is unknown (Nation & Waring 1997). According to Laufer (1998), though, a learner needs to know words that cover 95% of the text to be able to understand it. As far as speaking is concerned, 2,000 word families provide almost 95% lexical coverage in spoken discourse (Adolphs and Schmitt 2004).

Native 5-year olds have a vocabulary of 4,000-5,000 word families (Schmitt 2000:3) and they add 1,000 more each year which is actually quite impressive. Second language young learners attending schools with English as the language of instruction can also add 1,000 word families per year but do not reach up to the 4,000-5,000 word families native children have at that age. Studies showed that second language young learners appear to have a vocabulary of 1,000-2,000 word families after 5 years of English language learning (Nation 1990). However, this study was conducted in India and Indonesia so results in other countries may greatly differ.

According to Cameron, it is realistic to think that children could learn 500 words per year (Cameron 2001:75). This is very interesting but the challenge here is really to find out what sort of words young learners are actually taught. Also, what are the criteria for
teaching vocabulary to young learners? Needs analysis is usually necessary so as to know why and where young learners are learning the language. Frequent words could also form a criterion bearing in mind the statistics mentioned above.

What is of overall significance though is to make sure that the words to be instructed are appropriate at age and cognitive level as this may well influence their willingness and motivation to learn them. It is also important that these words are in accordance to their interests and daily life (e.g. classroom language, instructions) and so they are words or chunks they can use often, which also boosts up their confidence. A good idea is to start by instructing words which are used internationally and which learners are already familiar with (e.g. taxi, basketball, computer). This also promotes self-confidence.

One can easily derive from all these that what is most crucial when teaching at a young age is the psychological factors involved. Hence, experts or coursebook writers should select vocabulary items to be taught that foster young learners’ motivation and interest for the foreign language.

1.1. Young learners of English: The Greek context
Commonly, learners in Greece start when they are 8 to 9 years old. Exposure to the English language may vary according to learners’ background and often their parent’s socio-economic status (Mattheoudakis & Alexiou, forthcoming). Therefore, some learners might have started English during their pre-junior course (usually lasting an academic year and involving subconscious teaching strategies through games, stories and songs) but the majority tends to start at age 8 and they have lessons three times a week. Each lesson lasts about 45 minutes. Since 1991, English became the compulsory foreign language in all the primary schools. More recently, from 2003, English is also taught in the last four years of primary education. Apart from these classes, most young learners attend courses offered at private language institutes that are not compulsory and are mostly exam-oriented (Mattheoudakis and Alexiou, forthcoming). These lessons take place three times weekly and also last for about 45 minutes.

When comparing coursebooks of that level, it seems that children learn about 5 words per lesson (or more) during their first year, learners are likely to have completed about 100 hours and are expected to have an active vocabulary of around 300 words (CYLET handbook 2003).

When choosing coursebooks teachers opt for coursebooks that include many TPR (Total Physical Response) songs, rhymes, language twisters and they closely examine the methodology to fit their age. Activities are supposed to match with the interest and cognitive and intellectual level of the learners at age to facilitate re-labelling, as Aitchinson (1987) suggested. Words in the coursebooks have to be concrete and not abstract and it is a fact that these words often include infrequent words (such as balloon, alien, kite, zebra).

2. The study
The purpose of this study is to examine the vocabulary used in 5 of the most recent (2005-2006) and updated coursebooks for Junior A level. In particular, the study aims at examining the degree to which these words overlap with the first two thousand most frequent words in English, according to the wordlists of the General Service List (GSL) and the British National Corpus (BNC). Topic areas within which the vocabulary of these coursebooks falls are also investigated.

More specifically, the research questions are:
• Do the words instructed during the first year overlap with the first two thousand (2,000) most frequent words in English?
• Do we instruct highly infrequent words at that level?
• Are there any common words between the coursebooks?
• What thematic areas do these words cover?

Both qualitative and quantitative analysis were employed for this purpose.

3. Methodology and Procedure

3.1 The word lists

The word lists that were used in order to test the vocabulary of the coursebooks were the General Service List of English Words (GSL) (West, 1953) and frequency lists from the British National Corpus (BNC). The GSL comprises the 2,000 most frequent words in English based on frequency counts from a corpus of 5 million written words. The frequency lists from the BNC were developed by Nation (2006). The BNC is “a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written.” (http://www.natcorp.ox.ac.uk/corpus/index.xml). Nation developed 14 lists of 1,000 words each. For the purposes of this study, only the first two lists were used, i.e. the first 2,000 words, in order to be consistent with the GSL which consists also of 2,000 words.

The words of both the above lists have been expanded by Nation (Heatley, Nation and Coxhead, 2002) to their word families up to and including level 6, according to Bauer and Nation (1993). As a result, both the above lists contain in fact 2,000 headwords each. The actual number of word forms contained in these lists is much greater.

3.2 The software

In order to carry out this research, a software called RANGE (Heatley, Nation, and Coxhead, 2002) was used. The software may use as files any kind of word lists. One or more texts can be loaded and the program compares the text(s) against the word lists. The texts to be fed into the software must be in a txt format, and for this reason the coursebooks for this study had to be scanned. The results are presented in a different txt file that the software generates.

4. Results and Discussion

A point that had to be decided upon before continuing with the evaluation of the results was the unit of measurement that should be used. Vocabulary can be measured in tokens, types, word families, or lemmas (Nation 2001:6-8). The shortlisted candidates for this study were tokens and types. When counting in tokens, one counts every single word form, whereas when counting in types, one counts again every single word but when the same word form appears it is not included (Nation 2001:7). As a result, it was decided to use types and not tokens because many words in the texts kept repeating. This would not have given an accurate count of each new word form presented in the coursebooks, which was possible only by using types. The total number of types for each coursebook is shown in Table 1.
Table 1. The total number of types that appear in each coursebook

<table>
<thead>
<tr>
<th></th>
<th>Book 1</th>
<th>Book 2</th>
<th>Book 3</th>
<th>Book 4</th>
<th>Book 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of types</td>
<td>402</td>
<td>401</td>
<td>438</td>
<td>343</td>
<td>532</td>
</tr>
</tbody>
</table>

Regarding the number of types found in each coursebook, there appears to be some inconsistency. The first two coursebooks have almost an identical number of types presented, i.e. 402 and 401 respectively. This is not the case, however, with the remaining three. Coursebook 3 has 438 types, not too many more than the first two, but there is a sharp decline in the fourth coursebook which has 343 types. This is almost one hundred types less than coursebook 4. Finally, there is a steep increase in the fifth coursebook, which has 532 types, almost two hundred types more than coursebook 4, and one hundred types more than coursebook 3.

It is apparent that there is variety in the number of types presented in different coursebooks of the same level. It is important here to remember that this is level one, which means that these are the very first words that children learn in English. The average seems to be something in the area of a little more than 400 types, but there are coursebooks which present significantly less or significantly more types in the first level. This may have something to do with the thematic areas that the authors choose for each coursebook. The main thematic areas for young learners for Junior A level can be seen in Table 2.

Table 2. The thematic areas covered in the 5 coursebooks.

<table>
<thead>
<tr>
<th>Common thematic areas in the 5 coursebooks</th>
<th>Other thematic areas found in the 5 coursebooks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colours</td>
<td>Sports and leisure-time activities</td>
</tr>
<tr>
<td>Animals</td>
<td>Clothes</td>
</tr>
<tr>
<td>Numbers</td>
<td>Days, months, seasons</td>
</tr>
<tr>
<td>The body and the face</td>
<td>Comic / fairytale words</td>
</tr>
<tr>
<td>Family, friends and ourselves</td>
<td>Time</td>
</tr>
<tr>
<td>School and the classroom</td>
<td>Weather</td>
</tr>
<tr>
<td>The home</td>
<td>Musical instruments</td>
</tr>
<tr>
<td>Food and drink</td>
<td>Places</td>
</tr>
<tr>
<td>Toys</td>
<td></td>
</tr>
<tr>
<td>Celebrations</td>
<td></td>
</tr>
</tbody>
</table>
These are the thematic areas covered in the coursebooks. The first column shows the common thematic areas used in all 5 coursebooks. These areas are by no means exhaustive as, according to the writer’s idiosyncrasy or the publishers’ demands, these may differ. As can be seen there is relative variety but the core of areas are by and large similar. This of course comes as no surprise. Such areas (colours, animals, food, home, classroom, activities etc) are very close to young learners’ ‘heart’ and reality and the concepts included are already familiar to them. It would be quite strange trying to instruct, for example, scientific/psychological/medical terms that are remote from their interest, or topics related to religion and politics as these areas would surely include ideas and concepts second young learners have not yet met in their first language.

Table 3. The percentage of coverage that the GSL and BNC lists respectively give for each of the coursebooks.

<table>
<thead>
<tr>
<th></th>
<th>Book 1</th>
<th>Book 2</th>
<th>Book 3</th>
<th>Book 4</th>
<th>Book 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC</td>
<td>85.08</td>
<td>85.03</td>
<td>80.14</td>
<td>80.76</td>
<td>80.64</td>
</tr>
<tr>
<td>GSL</td>
<td>84.83</td>
<td>84.79</td>
<td>80.14</td>
<td>82.22</td>
<td>78.95</td>
</tr>
</tbody>
</table>

Turning now to the quality of these words, Table 3 shows the percentage of coverage that the 2,000 words of GSL and BNC lists provide in these coursebooks. There is not great discrepancy between the results that the two lists provide within the same coursebook. For the first two coursebooks, the difference between GSL and BNC is not more than 0.25, which is almost negligible. For coursebook 3 the count is actually identical. This means that we can safely assume that the percentage of coverage provided by the 2,000 most frequent words for these three coursebooks is at that level, whatever the measurement word list. As with Table 1, i.e. the total number of types presented, in Table 3 we find the first two coursebooks, although belonging to different publishing houses, to be in complete harmony regarding the number and also frequency level of types presented.

Nevertheless, things in Table 3 look rather different for coursebooks 4 and 5. The difference between GSL and BNC is about 1.5 and 1.7 respectively. A possible reason may be that the authors have chosen to use a rather more ‘traditional’ vocabulary in coursebook 4, since the percentage of GSL is higher than that of the BNC lists, and GSL is a word list coming from the ‘50s. For coursebook 5, it seems that it is quite the opposite; the authors have chosen a more ‘modern’ vocabulary, since the percentage of BNC is higher than that of GSL.

Whatever the difference between the results that the GSL and BNC lists provide, though, what is more conspicuous in Table 2 is that the majority of the vocabulary presented in these coursebooks belongs to the most frequent 2,000 words in English.
Percentages in all coursebooks range from about 79% to 85%. Clearly, the authors have chosen to use the commonest words in English at this first level to a great degree. On the other hand, the remaining percentage of lexis, which ranges from about 15% (coursebooks 1 and 2) to almost 21% (coursebooks 3 to 5), comes from less frequent levels. One would expect that exactly because the level is introductory there would not be infrequent words, or even if there would be, they would not be too many. 15% of the words of the coursebooks being infrequent, however, is not a negligible percentage. It seems that although the level is introductory, use of less frequent words could not be avoided. This also confirms a recent study by Milton and Vassiliu (2000) where it was found that elementary foreign language text coursebooks include large quantities of the most frequent vocabulary at the very outset of learning, but they also include considerable quantities of infrequent vocabulary. Below are some examples of words used for the thematic areas covered in the coursebooks of the present study but which do not belong to the 2,000 most frequent words:

- Animals (*grasshopper, squirrel*)
- The body and the face (*moustache, claw*)
- Clothes (*sweater*)
- Family, friends and ourselves (*godmother*)
- Food and drink (*cabbage, paella*)
- The house (*carpet, fridge, teapot*)
- School and the classroom (*blackboard, notebook*)
- Sports and leisure-time activities (*skateboard*)
- Celebrations (*cowboy, Halloween, wedding*)
- Days, months, seasons (*calendar*)
- Comic / fairytale words (*elf, fairy, dragon, monster, witch*)

Table 4. Total number of types and common types in all five coursebooks.

<table>
<thead>
<tr>
<th>Total number of types in all five coursebooks</th>
<th>Number of common types in all five coursebooks</th>
</tr>
</thead>
<tbody>
<tr>
<td>949</td>
<td>108</td>
</tr>
</tbody>
</table>

Graph 1. Total number of types and common types in all five coursebooks in percentages.

Finally, an attempt was made to ascertain to what degree there is an overlap in these coursebooks regarding vocabulary. Table 4 shows that the total number of individual types was 949. Quite surprisingly, the common types in all five coursebooks were only
108. Graph 1 shows that in percentages these figures translate into only a 10% overlap of the vocabulary used in these coursebooks. This confirms an earlier study conducted by Milton and Benn (1933), who examined the vocabulary in 29 text coursebooks of French and found that they had only 19 words in common. What this practically means is that each coursebook presents different words than the others. If a student, for example, is to use coursebook 1, he/she will learn other words than another student who will use coursebook 3. At this introductory level, this may not be as bad as it sounds. However, what teachers must make certain is that the words that are not presented in one coursebook but in another are encountered by the students in the following levels, if these missing words are deemed significant.

5. Conclusions
This study was an attempt to examine the vocabulary that Greek young learners are instructed at the first year of their English language learning. For this purpose, vocabulary from five coursebooks at Junior A level was collected and then analysed using the RANGE software. The results showed that in beginner level coursebooks there is great variety in the number of new words presented. Some coursebooks present more words than others.

This vocabulary was also compared to the 2,000 most frequent words in English. The results from this comparison showed that most of the words presented at this introductory level do indeed come from the 2,000 most frequent words, but at the same time there are considerable quantities of infrequent vocabulary also used.

Furthermore, common words between coursebooks of the same level seem to be very few. The study showed that there is only a 10% overlap of the types used in all five coursebooks. It seems, therefore, that coursebook writers are highly idiosyncratic in their choice of infrequent vocabulary, which confirms older studies (Milton and Vassiliu 2000, Milton & Benn 1933).

The topics selected are in accordance to their interest which guards young learners’ motivation; however, these topics do tend to be very simple, and, according to Cameron (2001), tasks and language should be demanding (although not too demanding) so as to present a challenge for the younger ones. It is true that, especially at this young age and when it come to words, children are like ‘sponges’ but what use are sponges without water? (Munõz, personal communication). Therefore, much more input is recommended, even if that means that this input will be higher than their present cognitive level of development, so that children’s ZPD (Zone of Proximal Development) will be benefited (Vygotsky 1962; Krashen 1985).

This study unveils some interesting results because it shows that common sense prevails in the choice of vocabulary for young learners. Coursebook authors use not only frequent but also infrequent vocabulary at this first level of instruction. They also use a variety of thematic areas which in turn determines the vocabulary presented. The most surprising finding is the discrepancy between all these words that seemingly have very common thematic areas in coursebooks.

A first future step would be to conduct a follow-up study comparing lexis from the first and the second Junior level and examine frequency issues and other differences. A comparison and analysis of vocabulary coming from the coursebooks used at Greek state schools during the first two years of English language teaching may be a very interesting path to follow as well.
References
British National Corpus, http://www.natcorp.ox.ac.uk/
Nation, P. & R. Waring (1997.) “Vocabulary size, text coverage and word lists”. In N. Schmitt and M. McCarthy (Eds.) Vocabulary: Description, acquisition and pedagogy, Cambridge: CUP, 6-19.