Forthcoming in The Current State of Interlanguage
eds. Eubank, Sharwood-Smith, Selinker.
Benjamins.
1995, pg 265-272.

Letter attached at the end

BEYOND 2000 - a measure of productive lexicon in a second language (1)

Batia Laufer - University of Haifa

Abstract

The paper describes a measure of productive lexicon in second language which calculates the percentage of the non frequent words (less frequent than the first 2000) in a sample of the learner's writing. Several uses of this measure are illustrated: in investigating the relationship between passive and active vocabulary, the progress in lexical richness of L2 learners and lexical attrition of immigrants.

1. Learning vocabulary means an increase in the vocabulary size

Second language acquisition has often been described and discussed in terms of the learner's progress along the Interlanguage continuum, from almost a non-existent knowledge of L2 towards native-like competence, without necessarily reaching this level. By the same token, L2 vocabulary development can be regarded as a gradual approximation to the native speaker's lexical system. Such a process will entail qualitative and quantitative changes in the learner's lexicon. Whatever the imprtance of improving the qualitative knowledge, e.g. the word's connotations, non-core meanings, collocations, the obvious mark of progress in lexical development is in the increase in the learner's vocabulary size. It can hardly be disputed that the most striking difference between the vocabulary of a native speaker and that of a second language learner, or between the vocabulary of a beginner and that of an advanced learner is in the number of words each one controls, particularly in free production. When a learner is stuck for a word, or is literally lost for words it is not the case that a specific shade of meaning is unavailable to him/her. It is the word in its core meanings that is unknown. An example

of the gap between the vocabulary of learners and that of native speakers is reflected in the following figures: learners at the Cambridge FCE proficiency level, which is also the level of many high school graduates in English in the Western world, are supposed to know around 3000 lexical items. 18 year old native speakers of English are reported to know about 18 000 words, according to modest estimates (Nation 1990).

2. Researching vocabulary learning requires measuring the productive lexicon

If vocabulary learning means gradual increase in the number of words in the learner's lexicon, then it is reasonable to expect second language researchers to investigate this phenomenon. For example, it would be interesting to find out what conditions (of input, or learner) affect the growth or lack of growth of vocabulary size; whether the passive and the active vocabulary developments are related; whether there is some kind of ceiling beyond which active vocabulary size does not grow; what changes in lexical size occur during the language attrition process.

To answer these questions and similar ones it is necessary to measure the productive lexicon of the subjects. The most popular measures used in the description of the productive lexicon are lexical variation (LV), lexical originality (LO), lexical density (LD) lexical sophistication (LS) and Giraud (1960) index. Other less frequently employed measures are semantic variation, lexical quality, t-unit length and error free t-unit length. Yet all these measures of lexical richness are seriously flawed as research tools as they can be demonstrated to lack validity, reliability, or both. (For a detailed criticism of lexical richness measures, see Laufer and Nation 1993). These problems with measuring the productive lexicon may explain the dearth of studies of lexical development.

3. Lexical Frequency Profile - a measure of productive vocabulary

Because of the above limitations of the various measures of productive vocabulary size, a different measure of lexical richness was devised - the Lexical Frequency Profile, hence LFP.

The LFP shows the percentage of words that a learner uses at different vocabulary frequency levels in his/her writing, or put differently, the proportion of words from different frequency levels vis a vis one another. The LFP is calculated as follows. Let us imagine a composition which consists of 200 word types. Among the 200 word types, 150 belong to the first 1000 most frequent words, 20 - to the second 1000, 20 - to the UWL- University World List (Xue Guoyi and Nation 1984) and 10 are not in any list. To calculate the LFP, we convert these numbers (the number of words at each frequency level) into percentages out of the total of 200 word types. The LFP of the composition is therefore 75% - 10% - 10% - 5%. The entire calculation is done by a computer programme which matches vocabulary frequency lists with a text that has been typed into the computer. The programme can identify each word form as a member of a word family. A word family includes the word's base form, its inflections and the derivational affixes -able, -er, -ish, -less, -ly, -ness, -th, -y, -non-, and un-. (For further discussion of word families see Bauer and Nation (1993). Having matched the text with the word-lists, the programme calculates the profile as described at the beginning of this section. Two profiles are calculated: one - in terms of word families, as illustrated in example above, and the other - in terms of word forms (2). The validity and reliability of LFP as a research tool were shown by Laufer and Nation (1993 and forthcoming).

4. Beyond 2000 - a more convenient measure

The lexical profile mentioned above is a detailed profile, showing 4 types of words used by the learner. We can also use a condensed profile which will distinguish between the basic 2000 words and the 'beyond 2000' words. Such profile is attained by adding up the first two word lists (the first and the second 1000s most frequent words) and by adding the last two lists (the UWL and the 'not in the lists' categories). In the above mentioned study, it was found that similarly to the detailed profile, the condensed profile was also reliable and valid. It remained stable in different compositions of the same learners and it varied significantly across groups of learners at different levels of language proficiency. To avoid the possibility of changes in the profile due to text length, comparisons were carried out on compositions of approximately

the same length. (Some preliminary results of the relationship between LFP and text length suggest that the profile is stable if the lenght is between 200 and 400 words. But definite conclusion about sensitivity of the profile to text length has not been reached yet). The advantages of the condensed profile over the detailed one is in facilitating cross-linguistic and correlational studies. Detailed word frequency lists are not available for all languages. Yet for many languages there can be found 'minimal dictionaries', or a list of most frequent words from which the 2000 basic words can be extracted, as in the cases of Finnish and Hebrew. In Russian, a list of the 2000 most frequent words is available in the form of 'The central vocabulary' (Vogt 1970). Therefore, the description of lexical profile in terms of basic and beyond-basic vocabulary makes it possible to compare profiles of the same subjects using different languages. In addition, if we have one measure rather than several, which numerically expresses the lexical richness of a composition we could correlate it with other variables like the passive vocabulary size, a grade on a test, and so on. The single measure expressing the lexical richness in free production could be the percentage of the beyond-2000 words in the learner's sample.

5. Using the 'beyond 2000' measure

Several studies have been undertaken in which the active vocabulary was measured by the condensed profile. I will briefly describe how it was used in these studies. (The profile referred to shows the relative percentages of *word families* at different frequency levels.

5.1. **Where has all the input gone?** - the effect of comprehension based instruction on active vocabulary.

The purpose of the study was to find out whether there would be an increase in the active vocabulary as a result of comprehension based instruction. Learners that were examined participated in a course of English for academic reading in which the language input consisted of reading comprehension of academic texts. Vocabulary was explicitely taught, but for comprehension purposes only. There was hardly any active use of language during class time.

The subjects were 37 EFL adult learners from the Haifa Institute of Science and Technology. Their L1 was Hebrew and their level of English was a rough equivalent of the Cambridge First Certificate of English. At the beginning of the experiment, all the learners were given the Levels Test of passive vocabulary size (Nation 1983) and were also asked to write a composition entitled 'How important is science to the modern world?' This topic was chosen as it was of interest to the students tested. A composition on a different topic would have produced a similar lexical profile as neither the LFP measure, nor the 'beyond 2000' measure are sensitive to the subject matter of the learner's writing (3). The compositions were analysed by condensed profiles, i.e. in terms of the basic 2000 words and beyond 2000 words. The same tests were repeated at the end of the semester. Table One shows the group means of the passive vocabulary and the mean lexical profiles of compositions at the two testing points.

PUT TABLE ONE HERE

The conclusion of the study was that teaching vocabulary for comprehension will result in an increase in the passive vocabulary, but this increase will not affect the growth of the productive lexicon. (4)

5.2. **Where has all the lexis gone?** - lexical attrition in immigrants' mother tongue

The study attempted to find out whether different sociolinguistic factors pertaining to the use of English and Russian in Israel would result in different lexical change patterns in these languages as spoken by old immigrants to Israel. Two groups of immigrants were compared. One group consisted of native speakers of Russian, the other - of native speakers of English. Both groups left their country of origin in their mid twenties, or their early thirties and have been living in Israel about 20-25 years. All the subjects completed at least secondary education in their countries of origin and are now middle class professionals. The two groups wrote a composition in their L1. The Russian group wrote about the problems facing the new immigrants to Israel, the English group had to argue for or against the government's right to control family size. For purposes of comparison, two groups of young NEW immigrants,

speakers of Russian (age 20-25) and speakers of English (age 18) were given the same compositions respectively. All the compositions were analysed in terms of basic and beyond 2000 vocabulary. Even though Russian and English morphologies differ from each other since the Russian words are heavily inflected, the beyond 2000 measure is not likely to produce very different profiles for the two languages. This is so because it describes the vocabulary in terms of percentages of word families, not word forms. Therefore 'girl' and 'girls' is counted as one 'word' for English, and the 10 forms of the Russian equivalent 'djevochka' (5 case forms in singular and 5 in plural) are also counted as one 'word'.

Table Two presents the comparison of the new and the old immigrants in the two L1 groups.

PUT TABLE TWO HERE

The results showed that the two languages, English and Russian exhibited different patterns of lexical change, probably due to the differences in the opportunities of the two immigrant groups to use their respective mother tongue, the degree of readiness to maintain it and the different status each language has in Israel.(5)

5.3 **When will they ever learn?** - the progress in the lexical richness of advanced learners.

The first stage of the study attempted to find out whether there would be a significant increase in the productive lexicon of advanced second language learners' writing over a period of one academic year.

The subjects in the study were 48 first year University students in the department of English Language and Literature, speakers of Hebrew or Arabic as L1. During the 1st year of study, the learners take twelve to sixteen hours of language courses in English Language and Literature per week with a variety of teachers, most of whom are native speakers of English. The new vocabulary input is received through listening to lectures and through the reading material assigned by teachers. All the learners took an entrance exam to the department in which they wrote a composition. In this composition, they were asked to argue for or against one of the following statements. (Their choice was kept on record in the department).

a. A person cannot be poor and happy, because money is always needed to gain something that is important to that person.

b. It is always what you do not have as a child that is important to you as an adult.

c. In a free country, industry has the right to develop any product that will sell, and sell it to anyone who can pay for it.

At the end of the first semester (14 weeks of intensive study), 23 students were asked to write the same composition again; at the end of the two semesters (28 weeks of study), the remaining 25 students were given the composition of the entrance exam. Table Three shows the changes in the lexical richness in the compositions of the two groups.

PUT TABLE THREE HERE

Table Three shows that the percentage of the beyond 2000 words has increased significantly in the two groups. This increase may look like an indication of the learners' progress in the productive vocabulary. The important question, however, is how the learners' lexis compares with the lexis of the native speakers' argumentative writing. The attrition study described in 5.2 showed that 18 year old English native speakers, newcomers to Israel used 19% of the beyond 2000 words and old immigrants - 21%. (see Table Three). Furthermore, an LFP analysis of several texts in a standard reader for academic purposes (the texts were written by scholars, dealt with general subjects and had the style of argumentative prose) revealed that the percentage of beyond 2000 words in such writing was as high as 35%. When we compare these figures with our learners' performance, we can see that the progress in lexical size which was found in our study, though statistically significant, does not look very impressive from a vocabulary learning perspective. If the ultimate goal of second language learning is a near-native competence, then our learners' journey is far from being over.

In the second stage of the study, we have been looking at the lexicon of more advanced learners in the English department, who are in their 2nd and 3rd years of study and also M.A. students, i.e. in their 4th or 5th year in the department. The data, which is still being collected,

will show if and when the active lexicon of L2 learners will reach a level similar to that of a native speaker.

6. Conclusion

In this paper, it was argued that vocabulary acquisition is reflected first and foremost in the increase in the learner's vocabulary size. Therefore research in vocabulary acquisition has to resort to measures of vocabulary size, particularly the size of the active lexicon.

It was suggested that a possible measure of the active lexicon could be the percentage of non basic words used in a sample of the learner's language, namely words that are not among the 2000 most frequent words in language, hence the name of the instrument - 'beyond 2000'

Examples of studies were given to show how the 'beyond 2000' measure was used to investigate some important issues in vocabulary: the extent of the growth in the active vocabulary over periods of time; the extent of lexical attrition; the relationship betwee a particular input and lexical cange; the lexical differences between different groups, speakers of the same language; differences between similar groups, speakers of different languages. It is hoped that the use of 'beyond 2000' will contribute to finding additional much needed answers in vocabulary development research.

Notes

1. Some of the data discussed in this paper was collected by my students. I am grateful to Ms. Aliza Bar-Shlomo for collecting the data for the 'input study' and to Ms. Edna Collins for the English data in the 'attrition study'.

2. In the word families profile, the percentage of the most frequent words is lower than in the word tokens profile since all instances of the same word, i.e. its inflected forms and its derivations are counted as one occurrence.

3. This is true for subjects of general nature. A topic that requires a highly specialized vocabulary might yield a different profile. Such topics should be avoided in studies of general vocabulary knowledge and acquisition.

4. In the original study, the changes in the composition profiles were also checked for each student. Out of 37 learners, 9 progressed, i.e. received a higher % of 'beyond 2000' on the post test, 3 showed no change in the profile and 25 deteriorated, i.e. had a lower % of the beyond-2000 words at the end of the semester than at the beginning. These results corroborated the 'no progress' results presented in Table One.

5. In the original study, there was an additional group of Russian immigrants who have lived in Israel for 10 years. These did not exhibit patterns of attrition in lexical richness.

References

Bauer, L. and I.S.P. Nation. 1993. Word families. **International Journal of Lexicography** 6: 253-279.

Giraud. P. 1960. **Principes et Methodes de la Statistique Linguistique**. Paris: P.U.F.

Laufer, B. and I.S.P. Nation. 1993. Lexical Richness in L2 Written Production: Can it Be Measured? Paper presented at the 10th AILA Congress, Amsterdam

Laufer, Batia and I.S.P Nation. 1995. (In press) Vocabulary size and use: lexical richness in L2 written production. **Applied Linguistics**

Nation, P. 1983. Testing and teaching vocabulary. **Guidelines** 5: (RELC supplement) 12-24

Nation. I.S.P. 1990. **Teaching and Learning Vocabulary** Rowley, N.Y.: Newbury House.

Xue Guoyi and I.S.P. Nation. 1984. A University word list. **Language Learning and communication** 3/2: 215-229

Vogt, H.O. 1970. **Ryskans centrala ordforrad** (The central vocabulary of Russian). Lund: Sprakforlaget.

TABLE ONE - Passive and active vocabulary change (N=37)

	Vocab. size (passive)		% of beyond 2000 words (active)	
	mean	sd	mean	sd
Pre-test	3000	637	11.6	3.9
Post-test	3300	752	11.1	4.2
Difference	significant		not significant	
	t=2.30, p=.03			

TABLE TWO - Comparison of new and old immigrants' lexis as measured by the % of beyond 2000 words

		Russian			English	
		mean	sd		mean	sd
Newcomers	(n=10)	21	3.57	(n=10)	19	6.05
20 years in Israel	(n=11)	13	5.17	(n=26)	21	12
t value		5.82			0.56	
p		.01 (significant)			.4 (not sign.)	

TABLE THREE - Changes in lexical richness

as measured by the % of beyond 2000 words

Entrance exam		Group 1 (n=23) After one semester		Group2 (n=25) After two semesters				
mean	sd	mean	sd	mean	sd	t	p	
9.96	6.2	13.17	5.85			2.76	.01	significant
8.48	3.87			10.04	4.39	2.27	.03	significant

END OF DOCUMENT - MESSAGE LOG

R3

USERID: RHFL401 ORIGIN: HAIFAUVM CREATED: 04/30/95 08:15:26

FILENAME: BEYOND SCRIPT CLASS: A FORMAT:C

SPOOLID: 847 RECS: 376 COPY: 1 DUPLICATE: 1

PRINTED AT: HAIFAUVM ID: VM3812 AT: 04/30/95 08:16:17

END OF DOCUMENT - MESSAGE LOG

Date: Wed, 25 Jan 1995 23:06:47 -0600 (CST)
From: Eubank Lynn Alan <eubank@jove.acs.unt.edu>
To: Batia Laufer <RHFL401@UVM.HAIFA.AC.IL>
Subject: Re: rsvp
In-Reply-To: <199501250226.DAA05667@jove.acs.unt.edu>
Message-Id: <Pine.SOL.3.91.950125230208.884A-100000@jove.acs.unt.edu>
Mime-Version: 1.0
Content-Type: TEXT/PLAIN; charset=US-ASCII

Ah, good. I'm glad to see that I did indeed get the comments to you! As for publication dates, I'd like to get all of this done as soon as possible, but, of course, it does depend at least in part on the people who only this very week managed to get their papers to us! that means we now wait thru the review process, etc. At any rate, our goal--or, at least, MY goal, is to get this done as soon as I possibly can. (I've got a LOT better things to do!)

now, as for other publication info, our provisional title has been The Current State of Interlanguage. Since the whole thing got started, we've managed to get a lot more cohesion into the volume than what one normally finds with a Festschrift, so it's been possible to consider a bit better overall publication. At present, we're planning to go with John Benjamins. at least, it's JB that sent me all the formatting instructions that I've been using to do the final copy on papers. I guess that's pretty meaningful. I've been letting Mike and Larry handle the "business" end of this whole deal, though, so I'm hardly the expert.

later,
Lynnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn

Accepted (14)