_lognostics
tools for vocabulary research

The Mathematics of Vocabularies

Paul Meara
University of Wales Swansea

This paper is a very exploratory piece about the mathematics of vocabularies, intended to provoke discussion rather than to provide answers. Like many British linguists, my own mathematical education is fairly rudimentary, and over the many years that I have been working on vocabulary acquisition, I have increasingly found this to be a problem. In other areas of applied linguistics, it is still possible to work with informal models - indeed, a great deal of British Applied linguistics is work of this sort. However, when it comes to working on vocabularies the limitations of informal models make themselves apparent very quickly.

The main problem with vocabularies is that they are big: even on the most conservative estimates of vocabulary size in monolinguals, we must be dealing with 15,000 or more items in the average person's lexicon, and it is, frankly, difficult to see how we can talk sensibly about something as large as this without using some sort of mathematical simplification. Most people get round this problem by simply treating the lexicon as if it were a simple, monolithic whole, happily ignoring the fact that words have remarkably different properties, that different parts of the lexicon behave in remarkably different ways. This can sometimes lead to wild generalisations about 'the lexicon', often based very limited data. Most of us would be able to think of a dozen or so papers where very strong claims about lexicons have been developed on the back of experimental evidence involving a mere handful of different lexical items.

When it comes to bilingual lexicons, we are dealing with larger and even more complex systems, and you would expect people to be more cautious about the generalisations that they make. In fact, here too, we have claims made about 'the bilingual lexicon' based on very small amounts of data. Stroop tests, (for example Kiyak 1982) make up a surprisingly large fraction of all the studies on bilingual lexical storage, and they employ an experimental method which uses at most a dozen words from a particularly specialised part of the lexicon. It is difficult to see how data from studies of this sort can be generalised in a meaningful way.

This paper is principally concerned with a very specific problem in lexical studies: the question of how vocabularies grow, and how we can describe this growth. Growth is a phenomenon which has attracted a lot of attention in other disciplines, and models for most types of growth and decay are available. The growth of populations, the growth of

economies, the spread of disease among a population and so on, all have analogues in the area of the lexicon, and in principle, it ought to be possible to apply these models to the way people's vocabularies grow and decline throughout their lives. As far as I am aware, however, this very basic property of lexicons has not been modelled seriously.

There are, of course, a number of unwritten assumptions about vocabulary growth in the research literature, though it is often difficult to disentangle just what these assumptions are because of difficulties in measuring vocabulary size anyway. In general, we all accept that growth in monolinguals begins very slowly, and increases throughout childhood. After that, vocabularies increase in a more-or-less linear fashion at a rate of about 1,000 words per year, perhaps with spurts around the age of seven and fourteen. This growth seems to tail off in adulthood, and vocabulary may even decline in old age. Where bilinguals are concerned, the general assumption seems to be that learning a second language does not seriously impair your first language in any way; that learning vocabulary in childhood is probably easier than learning as an adult; the more languages you learn, the easier it seems to be to acquire vocabulary in further languages, but the price you pay for this is that you forget them more easily; once learned, a second language vocabulary can quickly regenerate under the right conditions. I do not think that many people would quarrel with these statements, though there might be some quibbles with the detailed figures. In fact, it is very difficult to find anywhere where they are made as explicit as I have done here. They represent a set of commonly held views that we all take for granted, and which don't seem to be in need of an explanation.

Important as these assumptions are, there is very little work which examines them in any depth, and the only serious example of modelling in this area that I know is an article published by Klaus Riegel in 1968. This article, a wide-ranging, almost rambling piece, published shortly before Riegel's early death, deals with several different aspects of bilingual performance, with special reference to the lexicon. I first came across Riegel's work shortly after it was published, and it impressed me a lot - mainly because it seemed to talk about mathematical models of the bilingual lexicon that were extraordinarily simple, and yet far more comprehensive than anything else that was available. At the time, my own rather limited mathematics made it difficult for me to take Riegel's model any further than his original sketch did, although one could not help realising that this was an important new direction for research. Surprisingly, Riegel's paper seems not to have generated any interest among linguists. When I searched the standard databases in preparation for this paper, I was able to locate only a handful of people citing this work. This handful was mostly restricted to Reigel's immediate friends and colleagues working in the field of Transactional Analysis, and publishing in the journal *Human Development*. As far as I am aware, no-one has seriously taken up Riegel's ideas about the mathematics of lexical growth, and the way the growth patterns in two languages interact in bilinguals.

Riegels's ideas in this respect are basically very simple, and they can be summarised very quickly. Riegel's initial assumption is that people are exposed to the total lexicon of a language in a way that can be described in terms of a few simple parameters: the total

number of words in a language (A), the current size of their vocabulary (N), and a factor (m) which describes the richness of the linguistic environment, or the rate at which new words are experienced. This suggests that the rate at which you encounter new words in a vocabulary will gradually slow down over time: the more words you know, the fewer words there are left for you to learn. A plot of cumulative exposure to words against time will show a rising curve that gradually levels off.

Growth patterns of this sort can be described by simple exponential equations like the one shown in example 1:

$$dN / dt = m(A-N) \qquad \text{eq 1.}$$

where $N=A(1-\exp(-mt))$.

This equation states that the change in N over time gradually decreases as N gets larger, and that this slowing in the rate depends on the value of m. Setting m=0.07, for instance, and measuring time in years, produces a curve implying that normal monolingual speakers would have been exposed to about half the vocabulary of their L1 by the age of 10, just over three quarters by age 20, and 90% of their vocabulary by age 33. Readers may recognise the mathematics here: it is the same sort of argument that is sometimes used in lexico-statistics to describe the accumulation of lexical types in long texts. Setting m at other values makes the slope of the graph steeper or shallower, but doesn't change its basic shape, or affect the other arguments that depend on it.

Riegel goes on to show how the basic equations can be adapted to cover some simple bilingual situations. In particular, he shows that introducing a second language into a linguistic environment changes the patterns of exposure that people experience. Depending on how old you are when you begin a new language, and depending on how much time you devote to it, Riegel's model is able to predict how the two languages will interact.

Riegel distinguishes between two different types of bilingual environment. The first of these is an 'independent' condition, in which a proportion of the total language input occurs exclusively and consistently in a second language. He suggests that this condition occurs when a learner takes regular classes in the L2, but the model covers a whole range of cases, including the extreme one where a complete change of language occurs - for instance the case of children who leave their native country and never speak their native language again. Riegel's second condition, which he calls 'confounded', occurs when two languages are randomly encountered in the environment.

Riegel suggests that two similar sets of equations describe the way these two conditions affect the type of linguistic exposure that a bilingual experiences. The confounded condition is the more complex of the two, so we will leave it out of account here. In the independent condition, equation 1 describes what happens up until the point where the second language is introduced. After that point - call it t1 - Riegel suggests that two

equations are required, one for each language. Suppose that input in language A is reduced to a proportion p, and that the second language B is introduced at proportion q where q=1-p, then exposure to the two languages can be described by equations 2 and 3 respectively:

for Language A:

$$N=(A*1-\exp(-mp(t1/p + t-t1))) \qquad \text{eq 2}$$

for Language B:

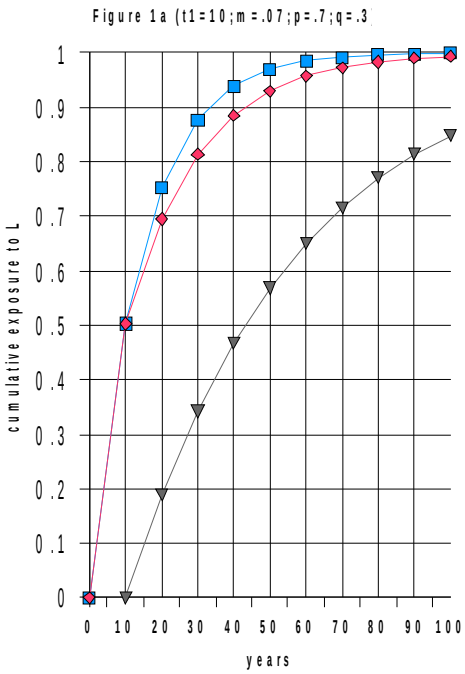
$$N=(B*1-\exp(-mq(t-t1)) \qquad \text{eq 3}$$

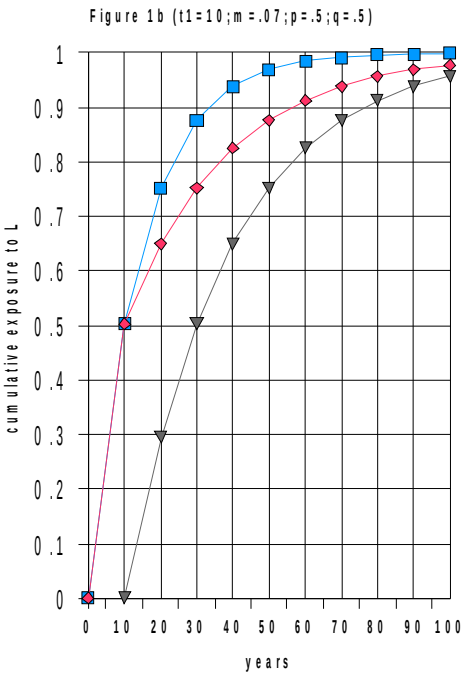
These abstract formulae will make more sense if you look at Figure 1, which shows how these equations perform with some reasonable assumptions about p,q and t1. Models a-c in Figure 1 are taken from Riegel's paper. In these graphs, t1 is assumed to be 10 years, and the proportion of L2 input is varied from .3 to .7. Riegel notes that in all three cases, substantial exposure occurs in the L2, with only a very slight loss in the L1 relative to the monolingual condition. Even in Riegel's extreme case, where the L2 receives 70% of the total exposure after age 10, and the L1 only 30%, it still takes a very long time for the cumulative exposure to L2 to exceed the cumulative total for L1.

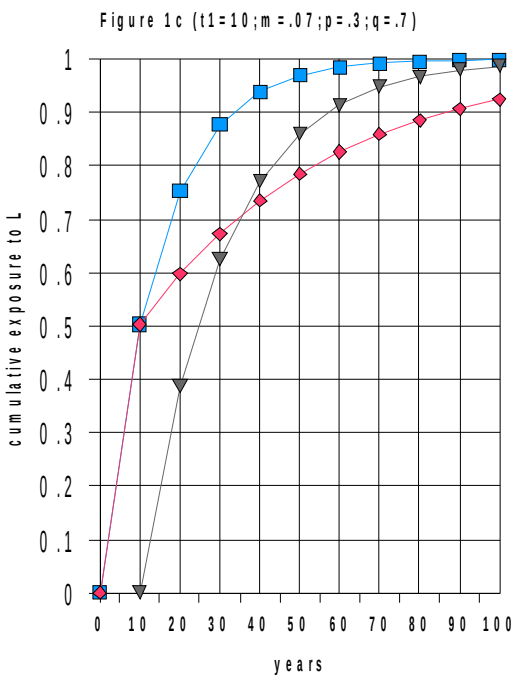
It is important to stress here that the curves in Figure 1 do not represent vocabulary growth directly. Riegel's model is a model of the linguistic environment in which learners find themselves, and it is explicitly not a psycholinguistic model. Later on in this paper, we will look at some ways in which the model can be adapted in the direction of a psycholinguistic model. For the moment, however, let us introduce a couple of bench marks, which will make the discussion slightly more concrete. It is obviously not the case that language proficiency varies directly with exposure to vocabulary, but we can reasonably establish some plausible achievement levels based on what we know about L1 speakers' ability to perform in their L1. An obvious benchmark of this sort seems to be the sort of competence level achieved by a five-year old child. A second useful benchmark is the level of competence achieved by a young adult, a person aged 15, say. We do not need to put actual numbers to the vocabularies at these benchmark levels: we can simply say that the level reached as a result of five years' exposure to the language produces a certain level of competence, call it C5, while further exposure up till age fifteen results in a higher level of competence, which we can call C15. In Figure 1 models a-c, these levels seem to correspond roughly to exposure to .25 and .6 of the vocabulary respectively. All three of Riegel's illustrative cases reach the benchmarks, though it may take a considerable time to achieve this.

Not all possible cases have this happy outcome, however, and some other plausible cases are shown in Figure 1 models d-h - a set of more interesting cases not discussed in Riegel's paper, but implicit in his model. Figure 1, models d and e represent the classic case of a person who takes up a second language late in life, and spends a strictly limited amount of

Figure 1: examples of bilingual environments with different values for the parameters p, q and t







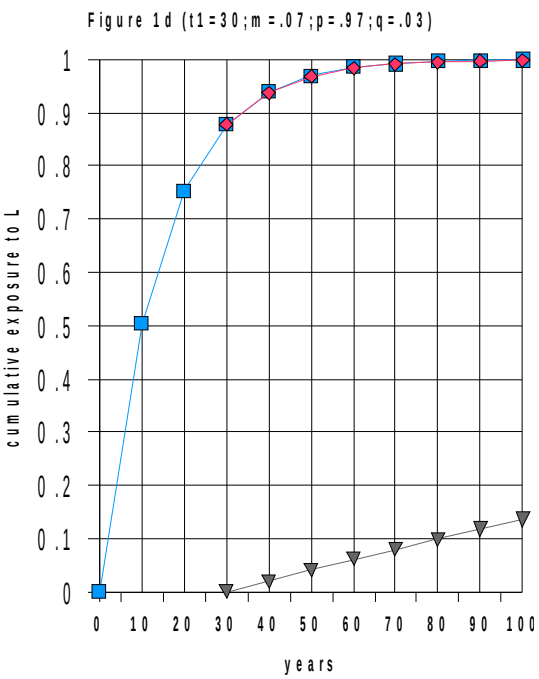
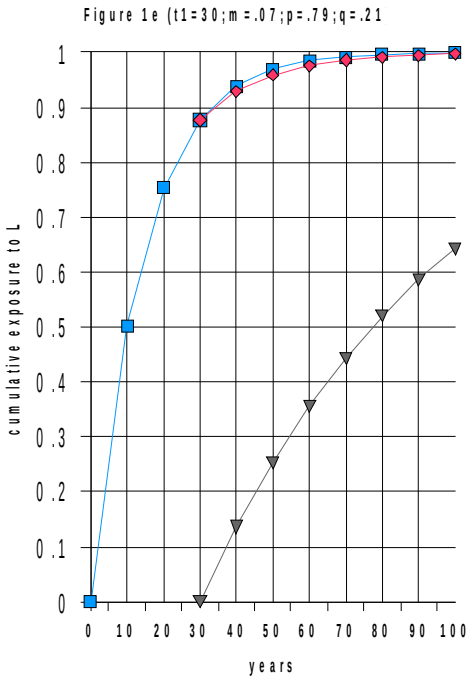
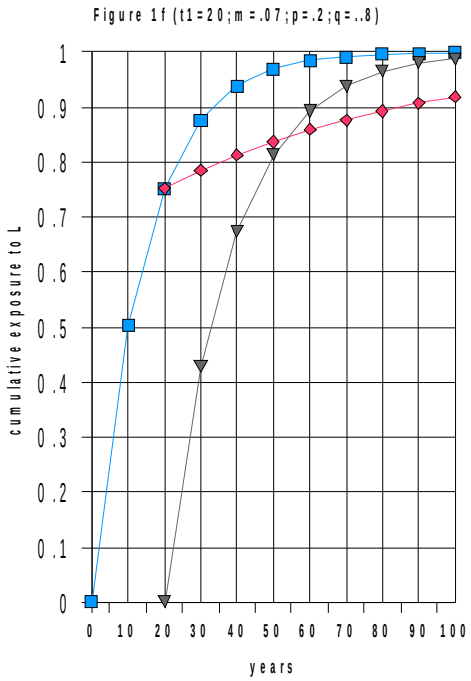
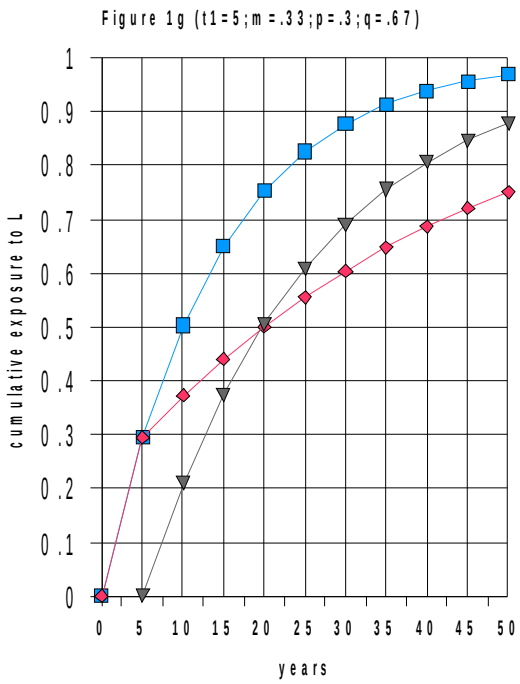
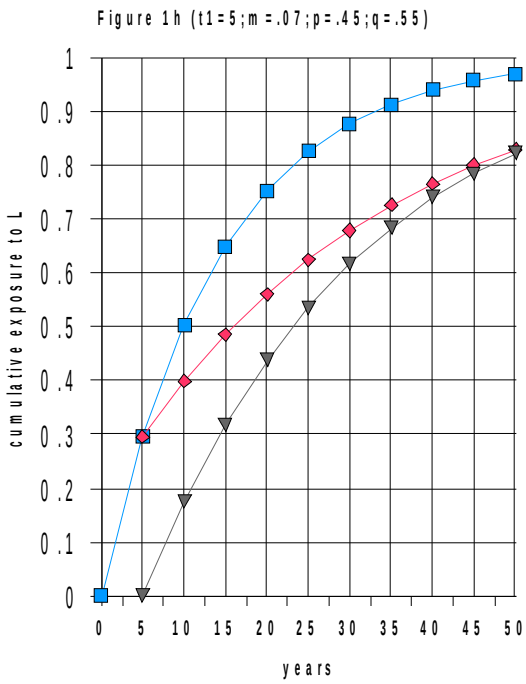


Figure 1 (continued)









time working on it or in it. Case 1d - the three hours a week learner - fails to get anywhere near our first competence level, C5: the cumulative exposure at this level of input just isn't enough for this level to be achieved. The two hours a day learner shown in Figure 1 model e does considerably better. Figure 1 model f represents a more extreme case, a young adult aged 20, who settles in a new country, and spends four fifths of their time exposed to the L2. In this case, both benchmarks are easily reached. In due course, cumulative exposure to the L2, and presumably linguistic competence, overtakes exposure to the L1. Figure 1 model g and model h show two cases of child bilingualism, where a child is brought up monolingually till the age of five, but subsequently goes to a school where the language of instruction is an L2. These two graphs show the effects of relatively small differences in the amount of L2 exposure in this situation.

As we noted earlier, it is important to stress that Riegel's model is a model of how people are exposed to the vocabularies of their L1 and L2, and not a model of how they acquire these vocabularies. Nevertheless, even working at this simplistic level, one or two points come out very strongly. The first point concerns the lasting impact of the L1 in any bilingual situation where the L2 is introduced after the L1. Even when the L2 is introduced quite early on, and when it receives greater exposure than the L1, cumulative exposure to the L2 lags behind cumulative exposure to the L1 for a very long time. The second point is that that the greatest difference between the L1 bilingual condition and the monolingual condition are longer term differences, rather than shorter term ones. For instance, in Figure 1 model b, the greatest difference between the monolingual condition and the L1 bilingual condition occurs around t=40. After that point, the differences decline and eventually disappear.

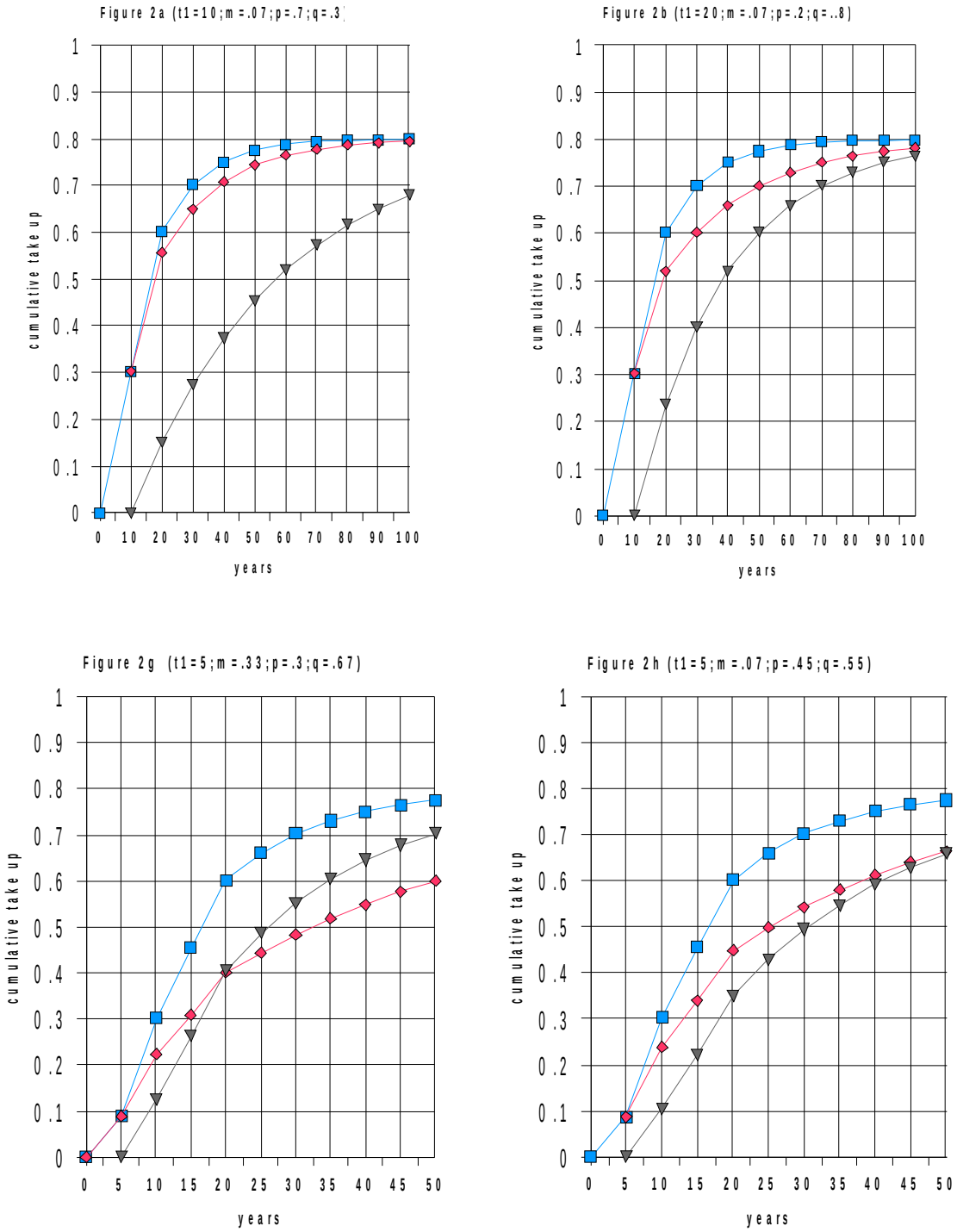
It goes without saying that Riegel's model is a grossly oversimplified and idealised picture of the way real people become acquainted with their language. It assumes, for example, that the language people are exposed to does not vary with their age – all the words are available all the time; it assumes that the languages being learned are stable, and do not themselves vary over time; and so on. Taking these simplifications into account, however, it is clear that Riegel's model throws some interesting light on the assumptions that we listed earlier. In particular, Riegel's model suggests an explanation as to why learning a second language doesn't appear to affect your first language: the answer is that it does, but only in the medium term, and only in the marginal areas of the lexicon. If exposure is the principal determinant of uptake, then only the very rarest of words will be affected by learning a second language. It follows, of course, that we would not expect to find great differences in the vocabularies of monolingual children and bilinguals with only a few years of exposure to their L2 - though this is precisely the stage where most people have attempted to look for differences of this sort. Even bilingual children with equal exposure to both languages from birth show very little deviation from the monolingual pattern over most of their school years on this model. A further prediction that we can derive from Riegel's model is that the richness of the linguistic environment is the crucial factor that determines vocabulary growth. Very small changes in the parameter *m* make very large differences in the rate at which new words are met, and these difference are particularly

important in the early stages of learning a language. A lexically rich environment can make up for a relatively small amount of exposure time. At the same time, the effects of a lexically poor environment can be compensated for in the long run, as long as the exposure to the linguistic environment is not too restricted. Combining a limited proportion of exposure time with a deliberately limited exposure rate - the pattern that we normally find for adults learning a language on their own and using text books with limited vocabularies – does not look like a regime that will generate a high success rate in the long term.

How important is it that Riegel's model deals with lexical environments, rather than with vocabulary acquisition per se? On the face of it, this simplification is a very important one, but on closer consideration, it might be less so. The crucial insight in Riegel's model is that it attempts to describe the learning environment in terms of a few simple parameters, and we can adapt this idea to make the model more plausible from a psychological point of view. The main problem with the model as it stands is that it implies that vocabulary growth is fastest at birth, and declines steadily after that. This is obviously false. Very young children don't have any vocabulary worth talking about, and vocabulary growth remains small until age two or three. However, we can make the model behave more plausibly by including a single extra parameter in the model, one that varies with age, and is particularly sensitive to very small ages. Let us call this parameter a take-up factor, and let us assume that in very young children the take-up factor is very small, and that it reaches a maximum around 18 years. At first sight, you might expect that introducing a factor of this sort would have a very dramatic effect on total vocabulary size. In practice, this isn't the case. Introducing a factor of this kind doesn't actually make a great deal of difference to the way Riegel's model works, except in the very early stages. Imagine, for example, a take up factor that starts off very small, reaches 60% by age 10, eventually levels off at, say, 80% at age 20, and after that doesn't vary very much. In effect, a take-up factor of this sort slows growth at the very early stages of language acquisition, but once the asymptotic level is reached all it does is impose a maximum level on the proportion of the vocabulary in the environment that is acquired. You can see this in Figure 2 models a-c, where the data plotted in Figure 1 models a-c has this additional take-up factor imposed upon it. The similarities between the two sets of curves will be apparent. Growth in the early stages of acquisition is slower than in Figure 1, and the final level of achievement is limited by the value of the take up factor. Since the take-up rate is affected mainly by age, and the L2 is introduced when the take-up rate is already high, the acquisition of L2 words is relatively fast, and the large differences between L1 and L2 recorded in Figure 1 are considerably reduced. The take-up factor also affects the size of vocabulary associated with our benchmark levels, C5 and C15, but perhaps not as markedly as we might have expected. If we plot out the early stages of these curves in more detail, there are some differences in the way the L1 and the L2 develop, but these differences have few obvious long-term effects. The main difference seems to be that take-up in an L2 acquired in adulthood will not be constrained in the same way as acquisition of an L1 in early childhood.

The take up factor is one that we might expect to vary a lot between individuals - an

Figure 2: Examples of Riegel's model, with the addition of an uptake parameter.



obvious candidate for a study of individual differences among second language learners. There is in fact some evidence that our ability to pick up words from the environment might be a stable trait, and that it strongly affects our ability to learn second languages (Skehan 1993; Gathercole, Hitch, Service and Martin 1997). This suggests that despite its simplicity, and despite the reservations that Riegel is at pains to make, the model may actually be more powerful than it looks. Furthermore, it is easy to see how the model might be developed to include other factors which it currently does not. At the moment, for example, the model assumes that the number of words in the linguistic environment is constant: it would be very easy to model a changing linguistic environment that included the sudden increase of vocabulary that children encounter when they go to school, or the massively large vocabularies that come into play in Higher Education environments. It would also be easy to introduce an attrition factor into the model, with very low levels of exposure resulting in an overall loss of the underexposed language.

The general point to be made here is that simple mathematical models can be surprisingly powerful, and given the fact that large lexicons are difficult to explore in an experimental context, there might be much to be gained from using models of this type in the study of vocabulary. Models of this sort can sometimes suggest fruitful areas of research that don't become apparent on their own – as we have seen here with the idea of take-up rate. Furthermore, there is no reason why models of this sort should be limited to the realms of theory. We now have a number of very large-scale studies of young children learning two languages in reasonably well-understood environments (e.g. Verhallen and Schoonen 1993), and it ought to be possible for us to use this data to evaluate models like Riegel's and to work out approximate values for the parameters his model uses. This would obviously be a significant step forward, opening up new and important fields of research.

Phenomena like growth and decay are well-studies in other disciplines, particularly biology (cf. Thompson 1961), and in these fields, simple models like the one I have outlined here have proved to be surprisingly productive. Despite its power, the mathematics that underlies these models is not difficult, so why is it that no work building on Riegel's model has been undertaken since the publication of this paper 25 years ago? The answer seems to be that simple mathematical modelling of this sort does not normally form part of the usual range of training provided to applied linguists. My personal view is that this is a serious problem: it puts applied linguistics at a serious disadvantage relative to other social sciences, by severely limiting the types of questions we can ask and the kinds of solutions we can propose to those questions. In the long run, we are going to have to address this issue, and make sure that formal modelling, if only at an elementary level, becomes part of the training of at least some young applied linguists, in much the same way as it has always been part of the training given to psychologists. As far as vocabulary is concerned, if we don't do this, then we will end up working with superficial models that greatly oversimplify the nature of vocabularies, but at the same time manage to miss out on the important general properties that vocabularies share with other large systems that grow, decay and reorganise themselves over time.

Bibliography

Gathercole, SE, GJ Hitch, E Service and AJ Martin 1997
Phonological short-term memory and new word learning in children. *Developmental Psychology*, 33(1997), 966-979.

Kiyak, **HA.** 1982
Interlingual interference in naming colour words. *Journal of Cross-cultural Psychology*, 13,1(1982), 125-135.

Riegel, K. 1968
Some theoretical considerations of bilingual development. *Psychological Bulletin*, 70,6(1968), 647-670.

Skehan, P 1993
Foreign language learning ability: cognitive or linguistic? *Thames Valley University Working Papers in English Language Teaching*, 2(1993), 151-191.

Thompson, D'A W 1961.
On growth and form. Cambridge: Cambridge University Press. 1961.

Verhallen, M and R Schoonen
Lexical knowledge of monolingual and bilingual children. *Applied Linguistics*, 14,4(1993), 344-363.

This paper first appeared in: **M Gill, A Johnson, L Koski, R Sell and B Wårvik** (eds.) *Language, Learning , Literature: studies presented to Håkan Ringbom.* Åbo: Åbo Akademi. 2001. 151-167.