How long does it take to learn vocabulary?

Jong Bor Lee

July 7, 2006

Let's say you are learning a second language. You may then know that grammar, though not necessarily easy, may well be mastered in a few months to a decent level. The real hinderance towards fluency is vocabulary, especially if the language you are learning doesn't share too many cognates with the ones you already know.

You may already know some greetings and the name of things you can find around the house. You have "catched" some words here and there. But it isn't enough to understand. So, now you are serious about learning as much vocabulary as possible. Unfortunately, there is no easy way to this task: you have to work on it daily, a few words at a time, and always review those which you have worked on previously.

So, we are interested in answering the following question: how long does it take to learn vocabulary?

I won't be concerned here with learning methods but with learning rates - it doesn't matter whether you asked the new words to a native, or looked them up in the dictionary, or read a list. I will make a few assumptions about the number of words learned daily by an individual and then I will derive some interesting formulas which answer the questions: how long does it take? How many words will I know on day x?

The assumptions

You will be really committed to learn during the first days, so you will try to memorize some considerable amount of words daily - after all, you want to build a solid foundation which will allow you to feel at ease when using your new language. However, you don't intend to memorize words throughout your life - it must stop some day. You don't want and you won't be able to keep the same learning rate of the first days.

This takes us to the following assumption: you'll learn n words the first day, the second day you'll learn a little less, and even less on the third day. Moreover, we will supose that the amount of words learned on day d is equal to a percentage λ ($0 < \lambda < 1$) of what was learned on day $d - 1$. Thus, if we write w_d for the number of words learned on day d, it is true that $w_d = \lambda^d n$. This formula is true for $d \geq 0$ ($d = 0$ is the first day).

We will make a second assumption: you have a goal, which is to learn N words. Most probably, you don't know how many words you want to learn. However, only a rough estimate is needed. If you intend to acquire a basic, survival vocabulary, $N = 2000$ will do. If you want to converse about any topic in your language, you should take $N = 5000$ or more. Erik Gunnemark says that $N = 8000$ words is all you will ever need[1]. Those aiming to read literature need $N = 10000$ and more.

The minimum effort

Given that $w_k = \lambda^k n$ is the number of words you learn on day k, the number of words you will know after d days is $W_d = \sum_{k=0}^{d} w_k$. Replacing w_k in this formula, we obtain:

$$W_d = \sum_{k=0}^{d} \lambda^k n = n \sum_{k=0}^{d} \lambda^k = n \frac{1-\lambda^{d+1}}{1-\lambda}$$

The last equality is due to the geometric sum formula $\sum_{k=0}^{n} r^k = \frac{1-r^{n+1}}{1-r}$.

No matter how big d is, W_d *won't* be as big as we want. This is counter-intuitive. If we take a big value of d, the the sum should be big too, shouldn't it? If we spend a lot of days learning, we'll be able to learn as much as much as we want, won't we? Not really: λ is smaller than 1, thus, we are suming numbers which rapidly decrease to zero. That is, we learn fewer words each day, until, some day, the number of words learned daily is near zero.

Taking $d \to \infty$, we show that W_d won't get bigger than $\frac{n}{1-\lambda}$. Now, let's recall we have a goal, namely, learning N words. Thus, we need $\frac{n}{1-\lambda}$ to be at least as big as N. This means that $\frac{n}{1-\lambda} \geq N$. Working out the value of λ, we get $\lambda \geq 1 - \frac{n}{N}$. That is: in order to reach our goal, we have to keep the percentage λ as big as $1 - \frac{n}{N}$. This is our first important result:

> In order to reach our goal of learning N words, we need to keep the percentage λ as big as $1 - \frac{n}{N}$.

Recall that n is the number of words you learn on your first day. Recall, too, that you are always learning less words than in the previous day. We suppose that the number of words learned any day is some percentage λ of the number learned on the previous day.

We may say that $\lambda = 1 - \frac{n}{N}$ is the minimum effort you should make in order to achieve your goal.

I can't learn 19,88023984 words a day

Imagine you have decided to learn $n = 20$ words on Monday. Your goal in the long run is to learn $N = 10000$ words. Thus, your minimum effort λ is

[1] I haven't been able to find a English version of the following article (in Russian): http://www.poliglots.ru/articles/gunnemark_vocabular.htm

$\lambda = 1 - \frac{20}{10000} = 0,998$. You like to plan ahead, so you calculate how many words you'll have to learn on Tuesday. This is $w_1 = \lambda n = 19,96$. You'll have to learn $w_2 = \lambda^2 n = 19,92008$ on Wednesday, and $w_3 = \lambda^3 n = 19,88023984$ on Thursday. Of course, this doesn't make too much sense. It *does* make sense from the purely mathematical point of view, but it doesn't when you have to really learn the words: you need to know whether you'll learn 19 or 20 words.

A possible solution to this problem is to learn 20 words daily as long as w_d is bigger than 19.

Let's skip to day 10. On day 10, $w_{10} = 19,6$ words. On day 20 we have $w_{20} = 19,21$ words. On day 30, we have $w_{30} = 18,83$ words. We see that w_d gets smaller day by day. We are interested in finding out which day d does w_d decrease to 19 on. This is easy to find out. We have to write $w_d = 19$ and solve for d. Replacing $w_d = \lambda^d n$ we get:

$$\lambda^d n = 19 \Rightarrow d = \frac{log(19/n)}{log(\lambda)} = \frac{log(19/20)}{log(0,998)} = 25,62$$

where *log* stands for the natural logarithm.

Thus, you should spend 26 days learning 20 words daily. After this 26 days, you'll begin to learn 19 words daily. Then you recall that you are learning fewer words day by day, so some day you'll learn only 18 words daily, and then 17 daily...

You have discovered that your task is to be divided in periods. The first period, you learn 20 words. On the second period you learn 19 words, etc. Each period is to be identified with the number of words you learn daily during it. In order to plan ahead, we want to know how much does each period last.

It is easy to calculate when each period starts. Period p (during which p words are learned daily) should begin when w_d (which is the *teorethical* number of words learned on day d) becomes as small as p. We solve $w_d = p$ for d and obtain $d_p = \frac{log(p/n)}{log(\lambda)}$. The length of period p equals the number of days until period $p-1$ begins (periods are numbered backwards!), that is

$$d_p - d_{p-1} = \frac{log\left(\frac{p-1}{p}\right)}{log(\lambda)}$$

We have found an important formula.

> The length of the period p, during which you learn p words daily, is $L_p = \frac{log\left(\frac{p-1}{p}\right)}{log(\lambda)}$.

How many words will I know when...?

The formula for L_p allows us to do some interesting estimates.

Of course, the number of words learned on period p will be pL_p. Let's call this w_p (forget about the theoretical w_d we used before - we won't need it anymore). It is interesting to know how many words you will have learned after k periods have elapsed. This equals $\sum_{p=n+1-k}^{n} w_p$. Let's call this quantity W_k (ignore the W_d we used before). A formula shown in the appendix allows us to write:

$$W_k = \sum_{p=n+1-k}^{n} w_p = \sum_{p=n+1-k}^{n} p \frac{log\left(\frac{p-1}{p}\right)}{log(\lambda)} \approx \frac{\frac{3}{2} log\left(1-\frac{k}{n}\right) - k}{log(\lambda)}$$

Not a bad formula, but one would rather want to know how many words one will have learned on a given *day*. We are not that far from finding this out, though. Yet another formula which will be shown in the appendix allows us to write that the number of days ellapsed at the end of period k, which we will call d_k (again, ignore any d_p or d_k you may have seen before), equals:

$$d_k = \sum_{p=n+1-k}^{n} L_p = \sum_{p=n+1-k}^{n} \frac{log\left(\frac{p-1}{p}\right)}{log(\lambda)} = \frac{log\left(1-\frac{k}{n}\right)}{log(\lambda)}$$

We may now solve for k. k will be written in terms of d:

$$k = (1 - \lambda^d)n$$

We replace this expression in our formula for W_k, which we now should call W_d. We may change the $\approx$ sign for a $=$ sign just to feel more at ease:

$$W_d = \frac{3}{2}d + \frac{n(\lambda^d - 1)}{log(\lambda)}$$

This formula can be simplified further. In fact, the first order Taylor formula for log says: $log(1+x) = x + \ldots$. Plugging in $x = -\frac{n}{N}$, we get $log(1-\frac{n}{N}) \approx -\frac{n}{N}$, that is $-\frac{n}{log(\lambda)} \approx N$. So, we may write:

$$W_d = \frac{3}{2}d + N(1 - \lambda^d)$$

Let's state it clearly:

> If you learn n words on day 0, and from that day on, the number of words you learn on any day equals λ times the number of words you learned on day before, then, the number of words you will know on day d is given by the formula:
>
> $$W_d = \frac{3}{2}d + N(1 - \lambda^d)$$

This formula has two terms. The first term, $\frac{3}{2}d$, represents a steady learning rate. In fact, this term says you are learning 3 words every 2 days. The second term equals zero when $d = 0$ and equals N when $d \to \infty$. This term may be much bigger than the first one for “reasonable” values of d. Take, for example, $d = 1000$ (about 3 years). Then, $\frac{3}{2}d = 1500$, but $N(1-\lambda^d) = 8649$ if we want to learn $N = 10000$ words in total and we use $n = 20$. We feel inclined to call the second term “the main term”, although it’s a pretty dull name.

Some may regard the first term as unrealistic. This terms increases no matter how big d is. However, we don’t plan to learn words all our life. Our mathemathical formula no longer reflects this fact. But this may be fixed easily by replacing the first term by some function $a(d)$ that decreases to zero when $d \to \infty$. It might even be the case that you regard the term as unrealistic because you are ambitious and instead of learning 3 words every 2 days you choose to learn 1 word every day as your steady learning rate. Then you would use $a(d) = d$. It’s up to you.

We may then state this general formula:

$$W_d = a(d) + N(1-\lambda^d)$$

where $N(1-\lambda^d)$ is the main term, due to your hard work (which consists in learning p words daily during period p) and $a(d)$ is up to you, and depends on how ambitious you are.

Appendix: some nice formulas

The calculation on W_k, the number of words learned after k periods have elapsed, leads us to the calculation of the sum:

$$\sum_{p=n+1-k}^{n} p \cdot log\left(\frac{p-1}{p}\right)$$

while the calculation of d_k, the number of days ellapsed at the end of period k, leads us to the calculation of

$$\sum_{p=n+1-k}^{n} log\left(\frac{p-1}{p}\right)$$

Let’s start with the easiest one, which is the second one, $\sum_{p=n+1-k}^{n} log\left(\frac{p-1}{p}\right)$. It is a telescopic sum:

$$\sum_{p=n+1-k}^{n} log\left(\frac{p-1}{p}\right) = \sum_{p=n+1-k}^{n} log\,(p-1) - log(p) = log(n-k) - log(n) = log\left(1-\frac{k}{n}\right)$$

Now, let’s analyze the first one, $\sum_{p=n+1-k}^{n} p \cdot log\left(\frac{p-1}{p}\right)$. We will start noticing that it is enough to calculate $I_j = \sum_{p=2}^{j} p \cdot log\left(\frac{p-1}{p}\right)$. Then the sum we want to calculate equals $I_n - I_{n-k}$.

Let's rewrite I_j as follows:

$$I_j = \sum_{p=2}^{j} log\left[\left(\frac{p-1}{p}\right)^p\right] = log\left[\prod_{p=2}^{j}\left(\frac{p-1}{p}\right)^p\right]$$

Computing $\prod_{p=2}^{j}\left(\frac{p-1}{p}\right)^p$ for small values of j quickly gives out a pattern:

$$\prod_{p=2}^{j}\left(\frac{p-1}{p}\right)^p = \frac{(j-1)!}{j^j}$$

This may be estimated with Stirling's formula:

$$j! \approx \sqrt{2\pi}j^{-1/2}e^{-j}j^j \Rightarrow \frac{(j-1)!}{j^j} \approx \sqrt{2\pi}j^{-3/2}e^{-j}$$

Thus, we get

$$I_j \approx log(\sqrt{2\pi}j^{-3/2}e^{-j}) = log(\sqrt{2\pi}) - \frac{3}{2}log(j) - j$$

Which implies that the sum $\sum_{p=n+1-k}^{n} p \cdot log\left(\frac{p-1}{p}\right)$ may be estimated as

$$I_n - I_{n-k} \approx log(\sqrt{2\pi}) - \frac{3}{2}log(j) - j - \left[log(\sqrt{2\pi}) - \frac{3}{2}log(j) - j\right] = \frac{3}{2}log\left(1 - \frac{n}{k}\right) - k$$

In conclusion:

$$\sum_{p=n+1-k}^{n} p \cdot log\left(\frac{p-1}{p}\right) \approx \frac{3}{2}log\left(1 - \frac{n}{k}\right) - k$$