Estimating Native Vocabulary Size in an Endangered Language

Timofey Arkhangelskiy
National Research University
Higher School of Economics,
Moscow, Russia
timarkh@gmail.com

Abstract

The vocabularies of endangered languages surrounded by more prestigious languages are gradually shrinking in size due to the influx of borrowed items. It is easy to observe that in such languages, starting from some frequency rank, the lower the frequency of a vocabulary item, the higher the probability of that item being a borrowed one. On the basis of the data from the Beserman dialect of Udmurt, the article provides a model according to which the portion of borrowed items among the items with frequency ranks less than r increases logarithmically in r, starting from some rank r_0, while for more frequent items, it can behave differently. Apart from theoretical interest, the model can be used to roughly predict the total number of native items in the vocabulary based on a limited corpus of texts.

1 Introduction

It is well known that in the situation of language contact the most easily borrowed part of the language is the lexicon (although there are counterexamples, see e.g. (Thomason, 2001:82)). Typically, for an endangered language or dialect L1 whose speakers are bilingual in another language L2 which is more prestigious and/or official in the area, the borrowing process is overwhelmingly unidirectional. Due to the influx of borrowed stems, words, and constructions from L2, as well as frequent code switching in speech, the size of the native vocabulary of L1 (defined as the set of vocabulary items in L1 which were not borrowed from L2 and are still remembered by the language community) is gradually decreasing. The stronger the influence of L2, the less native items remain in the vocabulary of L1, native lexemes being replaced with loanwords or just being lost without any replacement. Eventually the process may lead to a situation whereby L1 is confined to a small range of communicative situations, retaining only that part of native vocabulary which is relevant in these situations, and ultimately to language death (Wolfram, 2002).

It is interesting to study the vocabulary of a language currently undergoing the process of lexical erosion and search for rules that govern the process. Indeed, the process of native vocabulary shrinkage is not chaotic and turns out to conform to certain rules. In this article, I provide a model which shows how the native lexicon of an endangered language is being gradually lost. The model may be used to roughly estimate the native vocabulary size of the language. Apart from theoretical interest, such an estimate could have practical value for a field linguist, since it helps evaluate the coverage of the dictionary she compiled for the language: if the number of items in the dictionary is significantly less than the estimate, chances are there are vocabulary items still not covered by it.

2 The model and the data

The model is based on two observations related to frequency of vocabulary items. The main observation is that in the situation of extensive bilingualism, the probability of an item being a loanword instead of a native one increases with decreasing frequency of that item in L1: the less frequent the item, the more likely it is to turn out to be a borrowing. This synchronic property of the vocabulary is probably a consequence of a diachronic property of the borrowing process

Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 63–67,
Baltimore, Maryland, USA, 26 June 2014.

whereby the less frequent an item in L1, the higher the probability it will be replaced with a non-native item from L2 in a given period of time. The other observation is that such behavior is characteristic of vocabulary items starting with some frequency f_0, while items of higher frequency may be governed by different laws.

The relation between frequency, rank and other properties of lexical (and other linguistic) items has a long history of study, starting at least from Zipf's work (Zipf, 1949). The idea that the most frequent items can have special properties is also well known (see e. g. (Dixon, 1977:20) for syntactic properties or (Bybee, 2010:37–48) for phonetic and morphosyntactic effects of frequency), and it has been widely used in lexicostatistics and glottochronology since Swadesh (Swadesh, 1955) for estimating the degree to which several languages are related to each other and determining the point in time at which they diverged.

Based on these two observations and on the data from an endangered dialect, I propose a model of synchronic distribution of loanword items in the vocabulary of an endangered language. The model highlights the connection between the rank of an item (i. e. its number in the frequency list) and the probability that the item is a borrowed one. By a borrowed item I understand an item that was borrowed from the language L2 whose influence L1 is currently experiencing. This definition might seem a little arbitrary: what if L1 has a number of items left from its previous extensive contact? But since most vocabulary items in most languages were probably borrowed from another language at some point and since it is often impossible to distinguish between native items and old borrowings, one has to draw a line somewhere, and this seems to be the most reasonable way to do so. According to this model, the fact "item of the rank r is a borrowed one" can be viewed as an outcome of a Bernoulli trial in which the probability of success can be approximated quite precisely by a logarithm of the rank of the item in the frequency list, starting from some (not very high) rank r_0, while for any item with smaller rank it can behave differently:

(1) Pr[the item is a borrowed one] = $a \log(r) + b$, if $r > r_0$,

where r is the rank of that item.

The actual language data, however, makes it difficult to prove the hypothesis in the form presented above. The data the model should be tested against is a list of stems with their frequencies in the corpus and labels saying whether a stem was borrowed from L2. Thus, we have a situation of binary choice, as for every frequency rank the stem corresponding to it is either native, or borrowed. Besides, for great many stems it is impossible to determine their rank precisely, since, however large the corpus, there are always many low-frequency stems that have same frequencies in it (there are, for example, more than 1200 hapax legomena in my case). When several stems have the same frequency, we can determine the segment (r_1, r_2) their frequency ranks occupy, but we cannot say which stem has which frequency rank.

To overcome these difficulties, I first will seek an approximation for the function $P(r)$ defined as the portion of borrowed stems among all stems whose rank does not exceed r:

(2) $P(r)$ = (number of borrowed stems among those with rank < r) / r

As I will show, P(r) grows logarithmically in r, for r > r0, and this approximation is very precise for our data. In Section 4 I discuss why this fact implies the original claim (1).

The data I used comes from the Beserman dialect of the Udmurt language (Finno-Ugric). All speakers of this dialect are bilingual in Russian (and some in literary Udmurt), the number of speakers is at most 2000 and is decreasing steadily. The dialect, unlike literary Udmurt, is endangered, since most fluent speakers are now in their forties or older, and the children usually communicate in Russian both with each other and in the family. Beserman has a number of older loanwords borrowed from neighboring Turkic languages (which are recognized as native by the speakers and will not be dealt with in this article by definition of a borrowed item) and a vast number of Russian borrowings, either incorporated into the lexicon, or spontaneous. My primary source was a corpus of spoken Beserman totalling about 64,000 tokens that was collected in the village of Shamardan, Yukamensk region, Udmurtia, with my participation.

3 The analysis of the data

The items whose distribution was studied were stems, although similar calculations could be carried out for lexemes. I built a frequency list of

all stems, both Beserman and borrowed/Russian, for our corpus of spoken Beserman. Productive derivational affixes were not incorporated into stems, and in Russian stems, aspectual pairs were counted as one stem. The list was manually annotated: each stem was marked as either native or borrowed.

The distribution of native and borrowed stems is plotted at the figures 1 and 2. The only difference between the graphs is that the *x* axis of the plot on Fig. 1 is logarithmically scaled; all the data points and lines are identical at both plots. For each point, *x* stands for the rank of a stem in the frequency list, and *y* denotes the portion of borrowed stems among those with rank less than *x*.

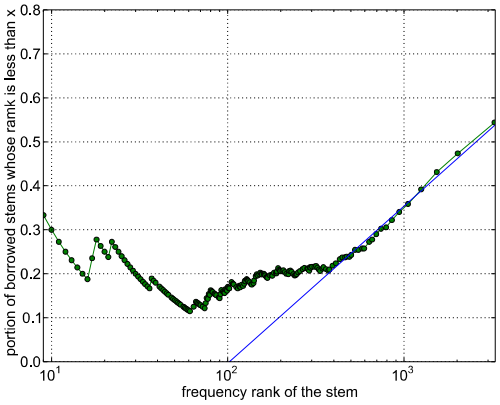

Fig. 1. Portion of borrowed stems with respect to the frequency rank with logarithmic approximation (semi-log plot)

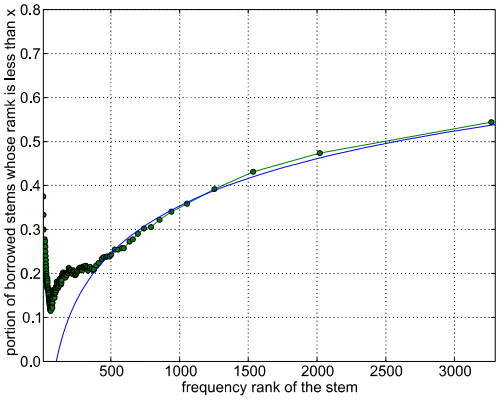

Fig. 2. Portion of borrowed stems with respect to the frequency rank with logarithmic approximation (linear axes)

The data points plotted at the graphs were split in two parts. Starting from r_0 of roughly 350, the data can be approximated nicely by a logarithmic function (a line in the semi-log plot): the blue curves are the approximations of the form $y = a \log(r) + b$ obtained with the least squares method. The peaks and declines in the beginning ot the frequency ranks range, e. g. for $r < 50$, do not provide any real insight into the behavior of the corresponding stems because the denominator in the formula for $P(r)$ is small and every single borrowed stem causes a visible rise of the line. For $50 < r < 350$, it can be seen that the portion of borrowed stems grows with r, but its growth does not conform to the same law which governs the behavior of less frequent items. For $r_0 > 350$, the best fit has the following parameters ($p < 0.001$):

$$a = 0.1550712 \pm 0.000254, \quad (3)$$
$$b = -0.71760178$$

The approximation is quite precise, as can be seen from the picture and the statistics (root-mean-square error 0.0088, coefficient of determination 0.99). One possible point of concern is the fact that the density of data points is much higher on the left part of the plot, so that the result is heavily influenced by the points with low frequency and only slightly influenced by the points with rank greater than 1000. If the items with higher ranks behave slightly differently than those with lower ranks, the difference could go unnoticed and the approximation will be not so precise for items with greater ranks. The only way to overcome this obstacle is testing the model on larger corpora. Another negative effect of such disparity stems from higher variance of the points on the left. However, it seems that for points with $r > 350$, the variance is already small enough for this effect to be significant (note that the y coordinate in such points is an average over at least 350 original observations).

Borrowed stems make up about 0.21 of the first 350 stems, and the behavior of $P(r)$ differs in this segment. The portion of borrowed stems increases slowly until it reaches the level of 0.2 for $r = 150$. For the next 200 frequency ranks or so, $P(r)$ stays at that level until it starts growing again around $r = 350$.

4 Calculating the probability of being borrowed

According to the model I propose, the labels "native" or "borrowed" in the data table can be

seen as generated by independent Bernoulli trials: the stem with frequency rank r gets the label "borrowed" with the probability $a \log(r) + b$, for all $r > r_0$. However, the logarithmic approximation that was derived in Section 3, estimates $P(r)$ rather than the probability of rth stem being a borrowed one. Here I will show how a logarithmic approximation for probability can be deduced from the approximation for $P(r)$.

Suppose the label for the rth stem is an outcome of a Bernoulli trial with probability of success (i. e. getting the label "borrowed") equal to $f(r)$, an increasing function whose values do not exceed 0 and 1. We define $z(r)$ as 0 if the rth item is native or 1 otherwise. Then the expectation of $P(r)$ can be estimated as follows:

(4) $E[P(r)] = E[(1/r) \sum z(i)] = (1/r) \sum E[z(i)] = (1/r) \sum f(i)$

The resulting sum may be estimated by the following inequalities:

(5) $$(1/r)\int_1^r f(x-1)\,dx \;\le\; (1/r)\sum_1^r f(i) \;\le\; (1/r)\int_1^r f(x)\,dx$$

Provided the interval is sufficiently narrow, we can assume that E[P(r)] is approximately equal to the right part of (5). Now, we know that E[P(r)] is well approximated by a logarithmic function $y = c \log(r) + d$ (for points where this logarithmic function is less than 0 or greater than 1, let y equal 0 or 1, respectively). Therefore, the following holds:

(6) $$\begin{aligned} &(1/r)\int_1^r f(x)\,dx = c\log r + d \;\Rightarrow \\ &(1/r)(F(r) - F(1)) = c\log r + d \;\Rightarrow \\ &F(r) = c\,r\log r + d\,r + F(1) \;\Rightarrow \\ &f(r) = F'(r) = c\log r + (c+d)\;, \end{aligned}$$

where $F(r)$ stands for the indefinite integral of $f(r)$. Using the constants obtained in the Section 3, we can estimate the probability as follows:

(7) Pr[the item is a borrowed one] = (0.1550712 ± 0.000254) log(r) − (0.534576 ± 0.000254), if $r > 350$.

5 Using the data for assessing dictionary coverage

The logarithmic model predicts that every item which has sufficiently large frequency rank will necessarily be a borrowed one, as the logarithm crosses the line $y = 1$ at some point. Based on this observation, one can estimate the expected total number of native vocabulary items the language retains. To do that, one should sum up the expected values of y for every r from 1 to the rightmost r for which the probability is still less than 1. In doing so, we assume that the events "the item of the rank r is a borrowed one" are independent and random (they happen with probability (0.1550712 ± 0.000298) log(r) − (0.56253058 ± 0.000298) for $r > 350$ and with probability 0.21 for more frequent stems). Calculations reveal that the point at which the probability curve crosses the line $y = 1$ lies in the interval (23770, 24206), and the expected total number of native stems is between 3603 and 3725 (for $a = 0.1550712$, it equals 3664). These bounds should be further widened as the observed value of a random variable is likely to deviate from the expected value within certain limits. Using Hoeffding's inequality for the sum of independently distributed random variables (Hoeffding, 1963) (8), we get that with 0.99 probability, the number of native Beserman stems should lie somewhere between and 3369 and 3962.

(8) $\Pr[|\sum X_i - E[\sum X_i]| \ge t] \le \exp(-2t^2 / \sum (b_i - a_i)^2)$, where $\Pr[a_i \le X_i \le b_i] = 1$

This estimate is rather imprecise, but nevertheless it provides information on the order of magnitude of the native vocabulary size. At the moment, there are about 2000 native Beserman stems known to us (which yields about 4000 dictionary entries in the dictionary (Kuznetsova et al., 2013)), therefore the model indicates that the list of stems can be significantly expanded and the efforts should be continued.

6 Assumptions and limitations

Apart from the two observations connecting frequency of vocabulary items and the probability of borrowing, there are more subtle assumptions the proposed estimate is based on, which can introduce additional pitfalls to the method.

One of such pitfalls is the assumption of representativeness of the corpus. When speaking of frequencies and frequency ranks of stems or words in the framework of this method, I mean the frequencies of those items in the corpus of

texts. In reality, however, an item is less likely to be replaced by a loanword if it is either frequent in speech in general, or frequent in particular communicative situations. As corpus data is the only means to estimate frequencies, we have to substitute the real frequencies with those found in the corpus. Although in the case of corpora of larger languages for which multiple means of communication are available (books, press, broadcasts etc.), the notion of representativeness is quite vague (Leech, 2006), for languages which exist only in spoken form, representativeness is much easier to define: the corpus can be said to be representative if the frequencies of items in the corpus faithfully reproduce the frequencies of the same items in speech. Thus, for the model to yield reliable results, we need a representative corpus. In practice that means that the corpus should contain texts of various genres (interviews, dialogues, folklore etc.), texts should cover a wide range of topics (including topics connected to the traditional culture and way of life as the vocabulary of these areas is especially likely to retain native items), they should be produced by speakers of different age, sex, background, etc. Failure to represent certain genres or topics in the corpus leads to certain items or classes of items being overseen by the researcher. For example, although our corpus covers a wide range of topics and genres, there were no occurrences of the words *tɨ* 'lungs' and *lɨ* 'spine', the only two words in the dialect that retain the phoneme */ɨ/*. The reason for that was, of course, not their overall low frequency in speech, but lack of texts recorded in situations where use of those words would be appropriate.

7 Further work

In order to verify the model presented here, it will be necessary to look at the data from other languages with similar status. As there exists a handful of manually annotated corpora for various indigenous languages of Russia which have undergone the same influence for roughly the same period as Beserman, the task of analyzing two or three more languages with comparable data seems realistic. Of course, it would be more productive to analyze larger corpora, but this is more of an obstacle here because such languages usually don't have corpora whose size would significantly exceed one or, at best, several hundred thousand tokens.

Apart from other languages in similar circumstances it would be helpful to see if the model works for languages that are engaged in language contact but not endangered (specifically, languages whose own word-formation mechanisms are still active), e. g. literary Udmurt.

If the data from other comparable language corpora indeed verifies the model, a possible further step would be to come up with a diachronic model that would describe the process whereby the native vocabulary is being gradually replaced with loanwords in a language whose own word-formation system has ceased to function.

References

Joan Bybee. 2010. *Language, usage and cognition.* Cambridge University Press, New York.

Robert M. W. Dixon. 1977. Where have all the adjectives gone? *Studies in Language* 1.1:19–80.

Ariadna I. Kuznetsova et al. 2013. *Slovar' besermjanskogo dialekta udmurtskogo jazyka* [Dictionary of the Beserman dialect of Udmurt]. Tezaurus, Moscow, Russia.

Wassily Hoeffding. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58 (301):13–30.

Geoffrey Leech. 2006. New resources, or just better old ones? The Holy Grail of representativeness. *Language and Computers*, 59.1:133–149.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21:121–137.

Sarah G. Thomason. 2001. *Language contact*. Edinburgh University Press, Edinburgh, UK.

Walt Wolfram. 2002. Language death and dying. *The handbook of language variation and change*, 764–787. Blackwell Publishing, Oxford, UK.

George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.