THE JOSEPH GREENBERG PROBLEM: COMBINATORICS AND COMPARATIVE LINGUISTICS

ALEXANDER YONG

1. INTRODUCTION

In 1957, the eminent linguist Joseph H. Greenberg (1915–2001) proposed the method of *Mass comparison* (also known as *multilateral comparison*) for determining genetic relatedness between languages [Gr57a]. I first learned about his work though the PBS/BBC documentary "In search of the first language"; a transcript (quoted below) is found at http://www.pbs.org/wgbh/nova/transcripts/2120glang.html.

Allowing Greenberg himself to summarize his idea:

> "I usually had preliminary notebooks in which I took those elements of a language, which, on the whole, we know are the most stable over time. These are things like the personal pronouns, particularly first and second person, names for the parts of the human body... I would look at a very large number of languages in regard to these matters, and I did find that they fell into quite obvious groupings."

This "presorting" technique is controversial, see, e.g., [Ri92, Gr93]. Its use [Gr87] to see that 650 languages of North and South America fall into three families (Eskimo-Aleut, Na-Dene and Amerind) is hotly debated. For example, James Matisoff argued:

> "Eyeballing data is prescientific, or nonscientific. There are so many ways you can be led astray, because very often, words look as if they have some connection, and they have no historical connection whatsoever."

Now, a succinct and combinatorial argument for Mass comparison is in [Wi13a]:

> "He also criticized the prevalent view that comprehensive comparisons of two languages at a time (which commonly take years to carry out) could establish language families of any size. He pointed out that, even for 8 languages, there are already $4,140$ ways to classify them into distinct families, while for 25 languages there are $4,749,027,089,305,918,018$ ways."

The two numbers offered are from [Gr57b] (he gives no other such enumerations). Indeed, the exactness of his enumeration for 25 languages makes a strong visual impression. However, in the interest of accuracy, we point out that in fact it is *slightly* erroneous: the correct enumeration is $4,638,590,332,229,999,353$.

Combinatorialists will recognize that the numbers Greenberg wanted to be *Bell numbers* $B(n)$. Nowadays, a quick lookup at the On-Line Encyclopedia of Integer sequences (http://oeis.org) detects the discrepancy. However, since his count may be of some historical interest, we elaborate upon the correction, and how one comes to notice it.

Date: October 10, 2013.

Actually, a main purpose for this elaboration is pedagogical. The author introduces the "Joseph Greenberg problem" to introductory classes in combinatorics: *Compute Greenberg's stated numbers.* Thus, the discussion might be of interest to the combinatorics instructor, or to the reader who is not already versed on the topic.

While Greenberg used Bell numbers to support Mass comparison, the same combinatorics does not support the plausibility of the resulting Americas classification. The probability, under the uniform distribution, of a classification of 650 languages having less than 100 families is near zero, but almost 100% for those in the range $[120, 150]$. Similarly, for 1000 languages, which seems like the upper bound for described languages of the Americas, the range is $[170, 210]$. These ranges bound, and are close to the more generally accepted viewpoint that there are between 150 and 180 families (a list of families is found in, e.g., [Wi13b]). They therefore suggest a theoretical interpretation of the observed range as, roughly, coming from the uniform distribution.

Also, most random classifications with this number of families and languages have a moderate number (9 to 19) of *language isolates*. This is somewhat consistent with the number of isolates/unclassified languages in the actual consensus classification (we seem to underestimate the number of isolates/unclassified languages by a factor of 5 to 10).

We contribute this combinatorial analysis to the above debate within comparative linguistics.

2. Multiset counting, Stirling and Bell numbers, and generating series

2.1. **The 8 languages case.** Greenberg's calculation of $4,140$ for the number of ways to classify 8 languages into families is correct. I find it helpful for the student to first determine the number using multiset counting.

Suppose that the (Native American) languages to be classified are

Alutiiq, Eyak, Cup'ik, Naukan, Mahican, Inupiaq, Tlinqit, Kalallisut

There are 22 partitions of 8. Each partition describes language family sizes. For example, in the case $5 + 2 + 1$, classifications correspond to to arrangements of the multiset A, A, A, A, A, B, B, C. Hence, the arrangement A, B, A, A, C, A, B, A encodes

$A \leftrightarrow \{$Alutiiq, Cup'ik, Naukan, Inupiaq, Kalallisut$\}$, $B \leftrightarrow \{$Eyak, Tlinqit$\}$, $C \leftrightarrow \{$Mahican$\}$

Thus there are $\binom{8}{5\ 2\ 1} = \frac{8!}{5!2!1!}$ such arrangements.

For the partition $4 + 2 + 2$, we rearrange A, A, A, A, B, B, C, C and both

$$A, B, B, A, C, C, A, A \quad \text{and} \quad A, C, C, A, B, B, A, A$$

encode the same language groupings. As there are two repeated parts one corrects for this overcount by dividing by $\frac{1}{2!}$ so that the number of distinct groupings is $\frac{1}{2!}\binom{8}{4\ 2\ 2}$.

Similarly, for the partition $2+2+2+1+1$ the count is $\frac{1}{3!}\frac{1}{2!}\binom{8}{2\ 2\ 2\ 1\ 1}$. It is not too laborious to compute all 22 numbers of this kind, add them all up, and thus recover Greenberg's stated number. Perhaps Greenberg did somesuch calculation as a check.

2.2. **The 25 languages case.** The method just used for the 8 language case becomes unpalatable for 25 languages because now there are $1,958$ partitions. It is logical to discuss now standard combinatorics, found in textbooks such as [Br10].

Suppose $\{a_n\}_{n\in\mathbb{Z}_{\geq 0}}$ is a sequence of answers to a counting problem (such as the Joseph Greenberg problem for n languages). The uninitiated might find it suprising that one can ever get traction on a problem by considering the infinitely many such problems, and then rephrasing the question as the coefficient of $\frac{x^n}{n!}$ in the **exponential generating series**

$$\sum_{n=0}^{\infty} a_n \frac{x^n}{n!} = a_0 + a_1\frac{x^1}{1!} + a_2\frac{x^2}{2!} + a_2\frac{x^3}{3!} + \cdots .$$

Let $S(n,k)$ be the number of ways to split n languages into exactly k language families; this is known as the **Stirling number**. One has the exponential generating series identity

$$\sum_{k=0}^{\infty} k!S(n,k)\frac{x^n}{n!} = (e^x-1)^k. \tag{1}$$

This can be obtained by the **product rule** for exponential generating series: Given

$$f(x) = \sum_{k=0}^{\infty} f_k \frac{x^k}{k!}, \text{ and } g(x) = \sum_{\ell=0}^{\infty} g_\ell \frac{x^\ell}{\ell!},$$

the coefficient of $\frac{x^n}{n!}$ in $h(x) = f(x)g(x)$ is $\sum_{k+\ell=n}\binom{n}{k} f_k g_\ell$. This is interpreted as counting two (ordered) boxes worth of combinatorial objects:

(I) the first of the type enumerated by $f(x)$, and
(II) the second of g's type;
(III) with the labels distributed to the boxes in all possible ways.

Note that $e^x - 1$ is the series for *unordered* collections of languages. Thus, $(e^x-1)^k$ is the series for classifications into k *ordered* families – a $k!$ factor overcount of $S(n,k)$.

Expanding the right hand side of (1) by the binomial theorem, one deduces

$$S(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i\binom{k}{i}(k-i)^n, \tag{2}$$

where $S(n,0) = 0$. One can lament that while this expression is explicit and nonrecursive, it is not manifestly nonnegative or even rational, even though it computes $S(n,k) \in \mathbb{Z}_{\geq 0}$.

Greenberg's problem for n languages is computed by the **Bell number** defined by $B(n) = \sum_{k=0}^{n} S(n,k)$. By (2) we have a non-recursive expression for $B(n)$. However, computing $B(25)$ by hand this way is impractical.

Actually, Greenberg's footnote on page 43 of [Gr57b] shows he knew the recurrence

$$B(n+1) = \sum_{k=0}^{n}\binom{n}{k}B(k), \tag{3}$$

presumably used in his computation.[1] A combinatorial proof of (3) follows by decomposing all classifications of $n+1$ languages by the number of other languages in that family. Still, executing the calculation for $B(25)$ would have been a task.

[1]Greenberg cites [Or42] that cites [Ep39] (which has values of $B(n)$ for $n \leq 20$). By 1962, [Le62] gave values for $n \leq 74$ and cite [Gu50] who had values for $n \leq 50$. Hence $B(25)$ was known at least to experts, if not widely available, by the time of [Gr57b]. Anyway, it seems likely he just computed it himself. Later, in [Gr87] he discusses the Bell numbers again. He writes that, using a computer program, $B(20) \approx 5.172\times 10^{10}$ (the correct value is $\approx 5.172 \times 10^{13}$); he does not restate $B(25)$ but refers the reader to [Gr57b].

Now, since $(e^x-1)^k/k!$ is the generating series for $S(n,k)$, summing over all disjoint cases k of the number of families, $\sum_{k=0}^{\infty}\frac{(e^x-1)^k}{k!}=e^{e^x-1}$ is the generating series for $B(n)$.

Often, this is where the classroom or textbook analysis stops, with the computation treated as moot. However, carrying it out, using, e.g., `Maple`, the instructor can make clear the speed one calculates the answer. For example, in `Maple` one computes as follows:

```
> 25!*coeftayl(exp(exp(x)-1),x^25);
                                4638590332229999353
```

In the interpretation for the Joseph Greenberg problem, this is a small surprise.

3. Back-of-the-envelope calculations

An advantage of generating series is that they are adaptable to related enumerations. While Greenberg used $B(25)$ to support Mass comparison, similar numerics do not seem to be consistent with a key consequence, his work on languages of the Americas [Gr87].

3.1. **A simple plausibility test and interpreting the consensus range.** Greenberg [Gr87] classified 650 indigenous languages of North and South America into three families[2]. This has been criticized for, e.g., not providing sufficient statistical evidence for claimed commonalities between languages. Specific to this situation is the use of a disputed method (Mass comparison) *and* a large disagreement (two orders of magnitude) among linguists as to the number of families. Side-stepping well-established lines of debate, imagine a restart using combinatorics. Roughly, how many language families should there be?

Consider each classification in an unbiased way. The generating series for language families with between a and b families is $\sum_{k=a}^{b}\frac{(e^x-1)^k}{k!}$. Using this, the probability a random classification on 650 languages has at most 3 families is $0.238\times 10^{-843}\%$. (This is comparable to the probability of randomly finding a prechosen atom from the observable universe correctly, ten times in succession.) This hardly disproves Greenberg's classification – but it does quantify how improbable it is from the baseline.

The probability that the number of families are in the ranges $[50,110]$, $[111,120]$, $[121,130]$, $[131,140]$ and $[141,150]$ are 0.0000565%, 0.56%, 37.1%, 58.8% and 3.5% (rounded), respectively. Hence almost all of the density is in the range $[121,150]$. Thus, we would naïvely guess (taking into account margin of error) that there are a few hundred families. This is decent agreement with today's consensus of 150 to 180 families [Wi13a, Wi13b]. For $1,000$ languages, which seems to me to be the upper end of the number of described indigenous languages of the Americas, the model predicts the range $[170,210]$.

These bounds on the consensus range suggest that the current "observed" range of $[150,180]$ families is consistent with the uniform distribution. That is, the naïve model gives a theoretical interpretation of the observed range.

3.2. **Language isolates.** A *language isolate* is a language family consisting of only one language. Euskara, the ancestral language of the Basque people, is one such example.

[2]The 650 figure is found, e.g., in [ChRo05, pg.105]

If all (modern) human languages originate from a single source, then one should consider, as Greenberg does, language classifications with no isolates. Indeed, in his Americas classification, Greenberg placed many generally regarded isolates (Yuchi, Chitimacha, Tunica, among others) into his Amerind superfamily.

Thus, suppose we consider such classifications for eight languages. The number of possibilities is the coefficient of $\frac{x^8}{8!}$ in e^{e^x-x-1} (which is 715). While this is not small, it says that the probability of a random language grouping having no language isolate is $\frac{715}{4140}$ or approximately 17%. For 25 languages, the probability is about 8.75%, whereas for 200 languages it is about 1.93%. In the case of $n = 500$ (about the number of languages analyzed in [Gr63]), it is 0.927%. For 650 languages (roughly the number studied in [Gr87]) it is 0.747%. In other words, most random language family configurations have a language isolate. This is not supportive of Greenberg's hypothesis.

By the product rule, the number of classifications on n languages with f families and i isolates is the coefficient of $\frac{x^n}{n!}$ in $\frac{x^i}{i!}\frac{(e^x-x-1)^{f-i}}{(f-i)!}$. For $n = 650$, if $f = 150$, the expected number of isolates is about 9 and nearly 5% have more than 14 isolates in the right tail. If instead $f = 180$ (the upper range of the consensus number of families), the expected number of isolates is roughly 19.5 with nearly 5% having more than 26 isolates in the right tail. Thus, the couple of dozen isolates/unclassified languages identified in the actual classification seems somewhat high (or our estimate seems somewhat low). This would predict that a number of current isolates/unclassified languages should amalgamate, reducing the total number of families down towards 150. We refrain from a "just-so" argument for why the number of isolates seems higher than the model predicts.

3.3. **What about Africa?** About $1,500$ languages ([ChRo05, pg.104]) of Africa were classified by Greenberg [Gr63], using Mass comparison, into 4 language families. Our naïve tests also reject this conclusion, estimating instead a few hundred language families.

Yet Greenberg's classification is generally regarded as a success! However, in [Sa09, pg 561] this success is qualified: "for the majority of Africa's best documented languages". In *loc. cit.* it is noted that there is a documentation problem for less popular languages, and the total number of languages in Africa varies substantially: from 2058 by one count, to 1441 in another. Moreover, the review [Di08] argues that there are 19 language families, given present knowledge. Also, D. Ringe suggested to us that Africa is special because of $2,000$ years of Bantu expansion that wiped out many language families. In any case, the situation is more complicated than first supposed. At this time, we merely conclude that langauge dispersal in Africa does not follow the same model as that for the Americas.

Concluding, the topic provides a source of classroom discussion/debate, and has room for further analysis and experimentation.

ACKNOWLEDGEMENTS

We thank Donald Ringe for a helpful and encouraging email correspondence (and his appearance in the documentary), from which I learned much of the comparative linguistics background I've presented (any errors are my own). We also l thank Eugene Lerman, Oliver Pechenik (especially for suggesting a criticism of [Gr87]), Anh Yong, and the Fall 2013 Math 580 class at UIUC. The author was supported by an NSF grant.

REFERENCES

[Br10] R. Brualdi, *Introductory Combinatorics*, Pearson Prentice Hall, Upper Sadle River, New Jersey, 2010.

[ChRo05] S. Chapman and C. Routledge (ed.), *Key thinkers in linguistics and the philosophy of language*, Oxford University Press, 2005.

[Di08] G. Dimmendaal, *Language ecology and linguistic diversity on the African continent*, Language amd Linguistics Compass, 2/5(2008), 840–858.

[Ep39] L. F. Epstein, *A function related to the series for* $\exp(\exp\, x)$, J. Mathematics and Physics, vol. 18(1939), 153–173.

[Gr57a] J. H. Greenberg, *The nature and uses of linguistic typologies,* International Journal of American Linguistics (1957), 23(2), 68-77.

[Gr57b] ———, *Genetic relationship between languages*, in "Essays in Linguistics", Chicago: University of Chicago Press, 1957, chapter 3, 35–45.

[Gr63] ———, *The Languages of Africa*, International journal of American linguistics, (1963)29, 1, part 2.

[Gr87] ———, *Language in the Americas*, Stanford University Press, 1987.

[Gr93] ———, *Observations concerning Ringe's 'Calculating the factor of chance in language comparison'*, Proceedings of the American Philosophical Society, (1993), 137.1 (1): 79–90.

[Gu50] H. Gupta, *Tables of Distribution*, East Punjab University, Research Bulletin, v. 2, 1950, p.44.

[Le62] J. Levine and R. E. Dalton, *Minimum Periods, Modulo* p*, of First-Order Bell Exponential Integers*, Math. Computation, Vol. 16, No 80(1962), 416–423.

[Or42] O. Ore, *Theory of equivalence relations*, Duke Math. J. **9**(1942), 573–627.

[Ri92] D. Ringe, *On calculating the factor of chance in language comparison*, Transactions of the American Philosophical Society, Vol 82, Part 1, 1992.

[Sa09] B. Sands, *Africa's Linguistic Diversity*, Language and Linguistics Compass 3/2(2009), 559–580.

[Wi13a] Wikipedia contributors, *Joseph Greenberg*, Wikipedia, http://en.wikipedia.org/wiki/Joseph_Greenberg (accessed September 19, 2013).

[Wi13b] Wikipedia contributors, *Indigenous languages of the Americas*, Wikipedia, http://en.wikipedia.org/wiki/Indigenous_languages_of_the_Americas (accessed September 25, 2013).

DEPT. OF MATHEMATICS, UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN, URBANA, IL 61801

E-mail address: ayong@uiuc.edu