Enumerating Words and Compositions

Jair Taylor

3-9-2012

Consider the following problem: How many words can be made from the word "Mississippi"? By this I mean I'd like to know the number of words that can be made using the letters m, i, s, p so that the number of times each letter is used is at most the number of times it appears in the word (disregarding capitalization). To answer this question, consider how we might choose such a word. First, we decide how many of each letters we will use. For example, let's say we choose m, 3 s's, 2 p's and and no i's. How many words can we make using all of these letters and no others? Well, an initial guess might be $6! = 40320$, since we have 6 total letters that may be put in any order. However, we have overcounted: some permutations of the letters correspond to the same word. For example, if we have the word "mpsssp", if we switch the two p's, the word remains the same. Similarly, we may permute the s's in any way and still the word will be the same. So in our count of 6!, we have counted the word "mpssspii" 3!2! times. Each word will be counted this many times, so the true count will be 6!/(3!2!). In general, if we have letters $c_1, ..., c_,$, and the number of copies of c_i we have is n_i, then the number of words using all (and only) these letters will be the so called *multinomial*

$$\frac{n!}{c_1!c_2!...c_m!}.$$

Now, consider the expression

$$(1+t)(1+t+\frac{t^2}{2!})(1+t+\frac{t^2}{2!}+\frac{t^3}{3!}+\frac{t^4}{4!})(1+t+\frac{t^2}{2!}).$$

We have chosen the degree of each polynomial factor to be the number of times each letter appears in the word "Mississippi": 1 m, 2 i's, 4 s's and 2 p's. If you were to expand this polynomial without combining the like terms, each term would correspond to choosing a term in each expression; for example, you might get the expression

$$t \cdot \frac{t^3}{3!} \cdot \frac{t^2}{2!} \cdot 1 = \frac{t^6}{3!2!}.$$

This looks similar to our expression 6!/(3!2!). We can think of each term as correspond letters we may use, with the chosen power of x in the kth factor corresponding to the kth letter; for this one corresponds to m, 3 s's, 2 p's and and no i's. The power of t in the resultant product gives the total number of letters we've chosen. But we're not complete: we need that 6! in the numerator. To make it appear, we use the Laplace transform. Recall that the Laplace transform $Lf(s)$ of a function $f(t)$ is defined to be

$$Lf(s) = \int_0^\infty e^{-ts} f(t)\,\mathrm{dt}.$$

The reason the Laplace is useful for our purposes is because it has the property that if $f(t) = t^n$ then its Laplace transform is $Lf(s) = \frac{n!}{s^{n+1}}$. This is close to what we want - it makes the factorial appear. Having the s appear in the denominator with power $n+1$ is a little inconvenient, so we use instead a modified Laplace transform, which I denote by l, defined by $lf(s) = \frac{1}{s} Lf(\frac{1}{s})$, i.e.:

$$lf(s) = \frac{1}{s}\int_0^\infty e^{-\frac{t}{s}} f(t)\,\mathrm{dt}.$$

Then we get the useful property:

$$\text{If } f(t) = t^n, \text{ then } (lf)(t) = n! \cdot t^n.$$

In other words, it allows us to change an *exponential* generating function

$$f(t) = \sum_n \frac{a_n}{n!} t^n$$

into the ordinary generating function

$$lf(s) = \sum_n a_n\, t^n.$$

Applying this to the expression above, we get the expression $\frac{6!}{3!2!} s^6$. Thus, summing all of these terms together, we see that that the coefficient of s^n is the number of words of length n. If we don't care about the lengths and only want the total number of words, we may set $s = 1$. Putting this all together, we get the following.

Theorem 1. Let $\{c_1, c_2, ...\}$ be a sequence of letters, and $\{n_k\}$ a sequence of nonnegative integers, and let a_n be the number of words made from these letters of length n such that the number of times the letter c_k appears is at most n_k. Then we have

$$\begin{aligned} \sum_n a_n s^n &= l\left(\prod_{k=1}^m \sum_{i=1}^{n_k} \frac{t^i}{i!}\right) \\ &= s^{-1}\int_0^\infty e^{-t/s} \prod_{k=1}^m \sum_{i=1}^{n_k} \frac{t^i}{i!}\,\mathrm{dt}. \end{aligned}$$

In particular, the total number of such words is

$$\sum_n a_n = \int_0^\infty e^{-t} \prod_{k=1}^m \sum_{i=1}^{n_k} \frac{t^i}{i!}\,\mathrm{dt}.$$

For example, the number of words made from “Mississippi” is then

$$\int_0^\infty e^{-t}(1+t)(1+t+\frac{t^2}{2!})(1+t+\frac{t^2}{2!}+\frac{t^3}{3!}+\frac{t^4}{4!})(1+t+\frac{t^2}{2!}+\frac{t^3}{3!}+\frac{t^4}{4!})\mathrm{dt} = 107899.$$

There are a number of variations on this theme. Recall that a *composition* of a number n in an ordered tuple $(a_1, a_2, ..., a_m)$ of positive integers whose sum is n. It is essentially the same as a word - the only difference being that we think of the characters being numbers and we are interested in their sum. We can count these just as easily, by introducing a new variable u which keeps track of the value or “weight” of each character. Since each factor in the expressions above represent a letter, we may replace the t in these factors by some $u^k t$, where k is the weight of the character. We immediately get the following.

Corollary 2. Let $\{n_k\}$ be a sequence of nonnegative integers, and let $a_{n,m}$ be the number of compositions of n in m parts so that the number of times any k appears is at most n_k. Then

$$\sum_{n,m} a_{n,m} u^n \, s^m \quad = \quad l\left(\prod_{k=1}^{\infty}\sum_{i=1}^{n_k}\frac{(u^k t)^i}{i!}\right)$$

(In these notes, the Laplace is always taken with respect to t.)

By setting $u = 1$, we ignore the weights on the symbols $1, 2, ...$ and we regain our original formula; if we set $s = 1$ then the coefficient of u^n gives the number of compositions of n with this restriction, regardless of the number of parts.

Let’s give an application of the above. Because our polynomials are the Taylor series of the exponential, it is easy to send the number of letters to infinity by replacing a polynomial by e^x. For example, one might ask a question like “How many compositions of a number n are there only using letters 1 and 2?” To answer this, we set our weights to be 1 and 2 and - ignoring for the moment questions of convergence - send the number of terms to infinity in the above expression, getting

$$\sum_{i=1}^{\infty}\frac{(ut)^i}{i!} \quad = \quad e^{ut}$$

$$\sum_{i=1}^{\infty}\frac{(u^2 t)^i}{i!} \quad = \quad e^{u^2 t}$$

so we only need to compute

$$\int_0^\infty e^{-t} e^{ut+u^2 t}\,\mathrm{dt} = \frac{1}{1-x-x^2}$$

which you might recognize as the generating function for the Fibonacci numbers:

$$\frac{1}{1-x-x^2} = 1 + x + 2x^2 + 3x^3 + 5x^4 + ...$$

How about the following problem: How many compositions of n are there that have at most one part which is 1? We can do the same trick as above; but to eliminate the possibility of having two or more ones, we limit the first factor to only two terms: $1+ux$. The 1 represents the choice of no 1's, and the ux represents choosing exactly one 1 to be in our composition. Then the factor for the 2's is e^{u^2t}, as above, and in general the factor corresponding to the possible part k is e^{u^kt}. We get

$$(1+ux)e^{u^2t}e^{u^3t}... = (1+ux)e^{t(u^2+...)} = (1+ux)e^{\frac{t}{1-u}-t-ut}$$

and so

$$\begin{aligned}\int_0^\infty e^{-t}(1+ux)e^{\frac{t}{1-u}-t-ut}\,\mathrm{dt} &= \frac{(2u^3-2u^2-u+1)}{(u^4+2u^3-u^2-2u+1)} \\ &= 1+u+u^2+3u^3+4u^4+...\end{aligned}$$

is the desired generating function. Since it is rational, we can use standard techniques to find a recurrence relation for the sequence of coefficients and so compute them very efficiently. I have verified these numbers via a brute force calculation, and I've submitted the sequence to The On-Line Encyclopedia of Integer Sequences A206268. Of course, I could have just as easily found generating functions for other similar questions.

Now, we will generalize the idea substantially. We would like to consider the problem of counting words, or compositions, with various restrictions - in particular, to count the number of words created from our multiset that have some given word as a subword. To return to our running example, we might ask how many words can be formed from the letters in "xylophone" that contain "ox" as a subword? Note that when I say *subword*, I mean that that the letters of the word have to appear in the right order, and consecutively. For example, "miss" is such a word, but "sis" is not.

To answer this question, we need the following tool - sequence of polynomials with a certain nice property. To find them, I first guessed that they might exist and then used Sage to compute their coefficients if they did - then from these values I was able to guess a simple formula.

Definition. We write

$$q_i(t) = \sum_{k=0}^{i-1}(-1)^{i-1-k}\frac{1}{(k+1)!}\binom{i-1}{k}t^{k+1} \text{ for } i \geq 1$$

and $q_0(t) = 1$. The first few such polynomials are $q_0(t) = 1$, $q_1(t) = t$, $q_2(t) = -t + \frac{t^2}{2}$.

Lemma. Then for any $n \in \mathbb{N}$ we have

$$\int_0^\infty e^{-t}\, q_i(t)\, \frac{t^n}{n!}\,\mathrm{dt} = \binom{n+1}{i}. \tag{1}$$

Proof. For $i = 0$ we get $\int_0^\infty e^{-t}\frac{t^n}{n!}\mathrm{dt} = 1 = \binom{n+1}{0}$. For $i \geq 1$ we have

$$\begin{aligned}
\int_0^\infty e^{-t}\, q_i(t)\, \frac{t^n}{n!} &= \int_0^\infty \sum_{k=0}^{i-1} (-1)^{i-1-k} \frac{1}{n!(k+1)!}\binom{i-1}{k} t^{n+k+1}\,\mathrm{dt} \\
&= \sum_{k=0}^{i-1} (-1)^{i-1-k} \frac{(n+k+1)!}{n!(k+1)!}\binom{i-1}{k} \\
&= \sum_{k=0}^{i-1} (-1)^{i-1-k} \binom{n+k+1}{n}\binom{i-1}{k}
\end{aligned}$$

and so we've reduced to showing

$$\sum_{k=0}^{i} (-1)^{i-k} \binom{n+k+1}{n}\binom{i}{k} = \binom{n+1}{j+1}.$$

This identity can be easily established using Wilf-Zeilberger theory or other means.□

Definition. Given a word nonempty word W, a *composition* is an ordered list of nonempty words such that, when concatenated in order, produce W. We take the convention that the empty word has exactly one composition, namely the composition with no parts.

For example, ("Miss", "is", "ippi") is a composition of "Mississippi". This is exactly analogous to the usual definition of composition of a number; it is easy to see that the number of compositions of a word is equal to the number of (numerical) compositions of the length of that word.

Definition. Let M be a finite multiset, and let A be a set of compositions of words on M. Call the compositions in A "allowed" compositions. For a given word W, let $n_{W,k}$ be the number of allowed compositions of W in exactly k parts, and denote by len W the length of W (i.e., the number of letters). We say that a specific word is allowed if the composition of that word with one part is an allowed composition. Define

$$p_{M,A}(t,s) = \sum_{W \in A, k \geq 0} n_{W,k} s^{\mathrm{len}\ W} q_k(t/s).$$

Then we have:

Theorem 3. Let a_n be the number of allowed words of length n. Then

$$\sum_n a_n s^n = lp_{M,A}(s).$$

Note that we allow the Laplace transformation of a function $f(s,t)$, where the integration is respect to t. The transform lf is then only a function of s, and is defined by

$$lf(s) = \frac{1}{s}\int_0^\infty e^{-\frac{t}{s}} f(s,t)\,\mathrm{dt}.$$

Proof. We have

$$\begin{aligned} l(q_k(\frac{t}{s})) &= \frac{1}{s}\int_0^\infty e^{-\frac{t}{s}} q_k(\frac{t}{s})\mathrm{dt} \\ &= \frac{1}{s}\int_0^\infty e^{-u} q_k(u) s\mathrm{du} \\ &= \int_0^\infty e^{-u}(u^0) q_k(u) s\mathrm{du} \\ &= \binom{0+1}{k} \end{aligned}$$

which is 0 for $k > 1$, and 1 when $k = 1$ or $k = 0$.

Thus we have

$$\begin{aligned} l\left(\sum_{W\in A, k\geq 0} n_{W,k} s^{\mathrm{len}\ W} q_k(t/s)\right) &= \sum_{W\in A} s^{\mathrm{len}\ W} n_{W,0} + s^{\mathrm{len}\ W} n_{W,0} q_1(t/s) \\ &= n_{\emptyset,0} + \sum_{W\in A} n_{W,1} s^{\mathrm{len}\ W} (\emptyset \text{ denotes the empty word}) \end{aligned}$$

since there is only one word that has any allowed compositions with 0 parts, namely the empty word - $n_{\emptyset,0}$ is 0 or 1 depending on whether the empty word is allowed. Since $n_{W,1} = 1$ for every word but the empty word, we get the desired expression.□

You might well wonder, at this point, what the point of defining these polynomials in such a strange way is. Since we need to find the statistics $n_{W,k}$ in order to calculate them, why on earth would we ever use them to find the number of words when we already had access to them in the first place?! The reason is that the polynomials $p_{M,A}$ play well together - we can easily string several requirements on our set of allowed words together to get a new polynomial. The only caveat is that the associated multisets should be disjoint. For example, we could ask: how many words can be made from the multiset "aaaabbb" that contain both of the subwords "aaa" and "bb"? The following theorem tells us that to do so it is only necessary to compute the polynomials p_{M_1,A_1}, p_{M_2,A_2}, where $M_1 = \{a, a, a, a\}$, $M_2 = \{b, b, b\}$

and the sets of allowed words A_1, A_2 are those containing “aaa” and “bb”, respectively, and then multiply them.

Definition. If ψ is a composition of a word W on an alphabet M (possibly a multiset, although the multiplicities are irrelevant here.) For $M' \subseteq M$ and a word W on M, let W' be the word on M' induced from W, i.e., W' is W with all letters not in M' removed. Then let ψ' be the composition of W' created as follows: for each part of ψ, divide that part into pieces which are the substrings whose letters are in M', separating these substrings when there is a letter between them that is not in M’. Then remove any empty parts. We call ψ' the composition of W *induced* by M'.

This is a long-winded definition of a simple idea. For example, if $M = \{a, a, b, b\}$ and $M' = \{a, a\}$, then the induced composition of (“abba”, “aab”, “b”) is (“a”, “a”, “aa”). This is implemented as the function “inducedcomp” in my code.

Definition. Let (M_1, A_1) and (M_2, A_2) be two disjoint multisets and two sets of allowed words. Let $M = M_1 \cup M_2$ and let A be the set of compositions ψ of words on M so that the compositions of ψ induced by M_1 and M_2 are in A_1 and A_2, respectively. We write $(M_1, A_1) * (M_2, A_2) = (M, A)$. This is implemented as the function “combinedcomps”.

The following is the most important part of the theory. Unfortunately, as of right now I cannot prove this, so I won’t call it a theorem. However, I can provide numerical evidence - see the Sage worksheet paired with this paper, and I feel very strongly that it is true. I hope to prove it soon.

Conjecture 4. Let M_1 and M_2 be disjoint, with sets of allowed compositions (of words) A_1, A_2 respectively. Then

$$p_{(M_1,A_1)*(M_2,A_2),A} = p_{M_1,A_1}(t, s) \cdot p_{M_2,A_2}(t, s).$$

Of course, inductively, this extends to any number of multisets $M_1, ..., M_n$ and associated sets of allowed words $A_1, ..., A_n$.

Proof. To come!

Corollary. Let M_n be the multiset consisting of a single letter repeated n times, $n \geq 0$, and let A be the set of all words on M - i.e., every word is allowed. Then

$$p_{M_n,A}(s, t) = \sum_{i=1}^{n} \frac{t^n}{n!}.$$

Notice that this polynomial is not a function of s.

Proof. To come!

Now, we can return to the problem of counting words on a multiset that have a given subword. We can apply Theorem 4 directly: Let M_1 be some multiset, W be a word on M_1, and

A_1 be the set of compositions of words on M that have at least one of their parts containing the subword. Then if M_2 is a multiset disjoint from M_1, and taking A_2 to be the set of all compositions on M_1 (i.e., putting no restrictions), $(M_1, A_1) * (M_2, A_2)$ will count for us the number of words on $M_1 \cup M_2$ containing W. To see this, recall that the allowed words are defined to be those whose composition in one part is allowed. The composition in one part of a word is allowed in $M_1 \cup M_2$ if and only if its induced composition is allowed, which means that, removing the letters of M_2 and dividing the remaining letters into parts corresponding to their division by letters of M_2, one of the parts will contain W; and this happens exactly when our word contains W.

Recall the problem posed above: How many words can be made from "xylophone" containing "ox"? To answer this, we divide up our multiset into the two smaller multisets and use Sage to compute the appropriate polynomials. We find $p_{M_1,A_1}(t, 1) = t + t^2$ where $M_1 =$ "xoo" and the allowed words are those containing "ox". Note that we are throwing away the extra information of the lengths of the allowed words by setting $s = 1$. Then we let M_2 be the remaining letters: "ylphne", and allow any words; we don't need Sage to compute this polynomial, because the last corollary implies that it should be $(1 + t)^6$ since there are 6 letters, none repeated. So we get that the number of words is

$$\int_0^\infty e^{-t} \left(t + t^2\right) (1 + t)^6 \, \mathrm{dt} = 95901.$$

There are a number of applications of these ideas that I have not mentioned. For example, all of the above we can apply to compositions of a number - we might ask, for example, how many compositions of n have the substring "$1212''$? This might turn out to be too hard for the theory to handle - we need to put bounds on the number of 1's and 2's. For example, we can ask: how many compositions of n have the substring "$1212''$, with fewer than five 1's and fewer than five 2's? We can also find variants of Corollary 2, by adding more variables with different weights. For example, putting a second variable v with different weights is adding a second dimension to our compositions - this will allow us to count the number of walks on the integer lattice from the origin to a given point. We might also ask for the number of walks that have a certain shape somewhere in the walk.