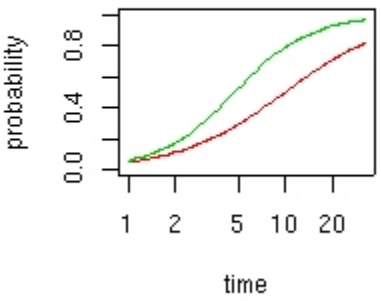
A methodology for inducing a chronology of the Pä li Canon

Paul Kingsbury
University of Pennsylvania

1. Introduction.

The use of statistics to describe language change has already enjoyed notable success in its relatively brief life. While the use of lexicostatistics and glottochronology met with considerable skepticism, more recent studies have returned more unambiguous results. Foremost among these studies have been the ones exploring the so-called Constant Rate Effect (Kroch 1989 and others). This hypothesis seeks to explain the S-shaped curve which best describes how any particular aspect of a grammatical system changes. In this type of curve, change starts slowly, then accelerates, then slows again, as seen in the lines in figure 1. It was previously thought that this curve reflected the underlying mechanism of linguistic change, that the rate of language change itself changed across time. The Constant Rate Effect does not challenge the description, but instead changes the explanation behind the curve. Instead of the rate of change changing, it is the frequency of the contexts in which a change is found which changes. As a context becomes more frequent, a linguistic change found in or dependent upon that context will also become more frequent.

Figure 1



It is also the case that sometime the same change occurs in multiple contexts. Previously it was thought that the rate of change was independent in each context. The Constant Rate Effect shows that this view is false and that the rate of change is constant across all contexts--it is actually the frequency of the contexts which varies. This is one part of the "constant" in the Constant Rate Effect.

The other important part of the CRE which remains constant is the rate of change itself. While it is undeniable that the curves show an apparent change in steepness, under the proper mathematical viewpoint this change disappears. The mathematical construct necessary here is borrowed from biology and called the *logistic*, described by the equation

$$p = \frac{e^{k + st}}{1 + e^{k + st}}$$

where p is the rate (probability) of attestation, t is the historical time, s is the (constant) slope of the curve, and e and k are constants.[1] One the right-hand side of the equation, the only factor changing is t. When t is small, at the beginning of the linguistic change, the denominator of the fraction will dominate and p will be small, reflecting a low rate of attestation. When t is very large, at the end of the linguistic change, the difference between the numerator and the denominator will be very small and p will be close to 1--meaning a nearly 100% probability of attestation. In between the two extremes, the ratio between the numerator and denominator changes very quickly, reflecting the high "steepness" of the middle of the curve. In this way the logistic equation describes the sigmoid curve.

The methodology for demonstrating the CRE depends on having a relatively good-sized corpus from which many examples of some change can be extracted, both in the positive (the change having taken place) and the negative (the change having not taken place). The relationship between these numbers gives the

[1] E is the mathematical constant, the base of the natural logarithm and approximately 2.718. K describes where the curve intersects the y-axis. As such, it is necessary for graphing the curve but otherwise uninteresting to the linguist.

probability of attestation for that change in the text in question. In order to determine a good time dimension, it is necessary that the texts of the corpus extend across some reasonable length of time, usually on the order of two or three centuries. With these methodological constraints in place, the CRE has been demonstrated for Old English (Kroch 1989), Yiddish (Santorini 1993), Ancient Greek (Taylor 1994), and others.

Difficulties arise, however, when one or more of the criteria for the CRE methodology is absent or faulty. Such is the case for Pä li. The Pä li literature includes nearly 4 million words in the canonical literature alone, spanning four centuries at the very least.[2] Pä li was used at a time of great linguistic change in India, as the elaborate grammatical systems of Old Indic (*eg* Vedic and Sanskrit) were collapsing and the seeds of the vastly typologically different Modern Indic (*eg* Hindi) systems were just germinating. For example, between Old and Modern Indic the verb system lost the medial voice altogether, lost several modes such as the subjunctive and optative, and reconstructed the non-present system from single inflected forms to periphrastic constructions. The noun system lost the dual number and basically the entire case system, moving from a nominative-accusative system to an ergative. Any of these changes are countable and therefore could serve as the basis of a diachronic study. What is missing is the clear division into historical strata. That is, while there are a lot of texts spread across a lengthy period of time, the correlation between any given text and any particular period is usually unknown. This study will attempt to show how a rough chronology can be induced based on a known historical change, focussing on the shift from multiple formations for finite past tenses ("preterites" or "aorists") to only a single formation.

2. Grammatical Background.

Old Indic, as attested in Vedic and Sanskrit, had a wealth of finite past tenses, including a perfect, an aorist, and an imperfect, each based upon a different stem formation. By the time of even the earliest Middle Indic, however, the perfect was largely gone except for a few fossilized remnants, and the imperfect had been folded into the aorists. The aorists throughout this time were the most productive system, and the system itself showed multiple formations. These are traditionally divided into four types. Of these, two are characterized by the addition of a morpheme -s-, while the other two are not. The root and thematic aorists, collectively known as asigmatic aorists, are generally considered highly archaic and unproductive. In contrast, the sigmatic aorists, formed by either -s- or -is- or phonological variants thereon, are highly productive. This can be seen partly by the fact that every verb which attests an aorist attests one of the sigmatic forms, while the same is not true for the root and thematic formations. Also, looking at later Middle Indic shows that the nonsigmatic aorists have disappeared altogether. Therefore, there is a clear competition between the sigmatic and nonsigmatic formations, a competition which the sigmatic eventually wins. Since the two types of formations are semantically identical, they appear in (mostly) the same contexts, and thus a simple two-way contrast can be counted: does a past-tense sentence attest a sigmatic or a nonsigmatic aorist? For any given text the proportion of sigmatic to nonsigmatic aorists can be determined and this proportion used as the basis of the induced chronology.

3. Methodology.

As mentioned before, the Pä li Canon comprises some four million words. There are a number of traditional divisions of these texts, either by style into nine types or into the Tipiöaka ("Three Baskets") and their subdivisions, reflecting a more semantic or functional division. The latter is more commonly used. The Pä li Canon was transmitted orally for several centuries and experienced considerable modification during that time, as divisions and even individual texts were expanded and abbreviated. Therefore, this study has taken the traditional divisions and further refined them into units which have a reasonable chance of stemming from a single historical period. This results in somewhat over 10 000 texts to be dealt with, ranging in size from many two-liners[3] to one text of over 20 000 words.[4] Of these texts, slightly over 5000 attest at least one aorist of any formation, providing a good source of data for this study.

---

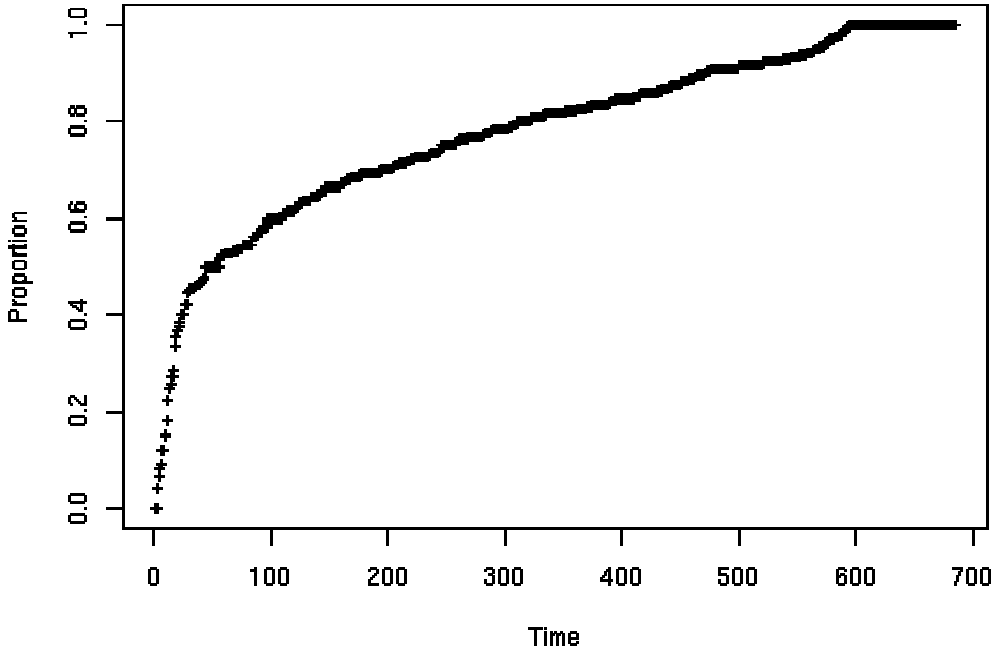
[2]Including later works such as commentaries and histories vastly increases both the size of the corpus and its chronological range, but these works are omitted from the current study.

[3]Many of the metrical texts are poems of only a single verse, and the traditional scholarship recognizes that many of

Unfortunately, most of these texts cannot be used because they do not attest aorists robustly enough. That is, while the proportion between (eg) 15 sigmatic and 5 nonsigmatic aorists is significant, the difference between one sigmatic and no nonsigmatics is not trustworthy. For example, consider a text consisting solely of two lines of verse, attesting only a single sigmatic aorist. Was that form used because it was the only formation available at the time the poem was written, or because it fit the metre while the corresponding nonsigmatic did not? Such a short text does not allow that question to be answered with any confidence. Therefore, any text which attested fewer than 10 aorists was discarded from the study.[5] This left 685 texts,attesting a total of 27145 tokens or an average of 40 aorists per text.[6] For each text the number of sigmatic and nonsigmatic aorists was counted, and the proportion of sigmatic aorists to total aorists was calculated. The average proportion was somewhat over 0.76, indicating that about three out of every four aorists in the set of texts were sigmatic. This is perfectly consistent with the idea that the sigmatic aorists are the productive and winning formation in Pä li.

A working assumption was made that this calculated proportion was a perfect correlate to historical age, and that any text which showed a low proportion was necessarily earlier than any text with a higher proportion. This is at best only approximately true but allows the rest of the analysis to proceed. The texts are ordered by proportion and plotted as Figure 2.

Figure 2:
Texts plotted by rising proportion



these poems were written by distinct authors.

[4]The Yamaka, an exhaustive treatise on dualism.

[5]Why 10? Why not?

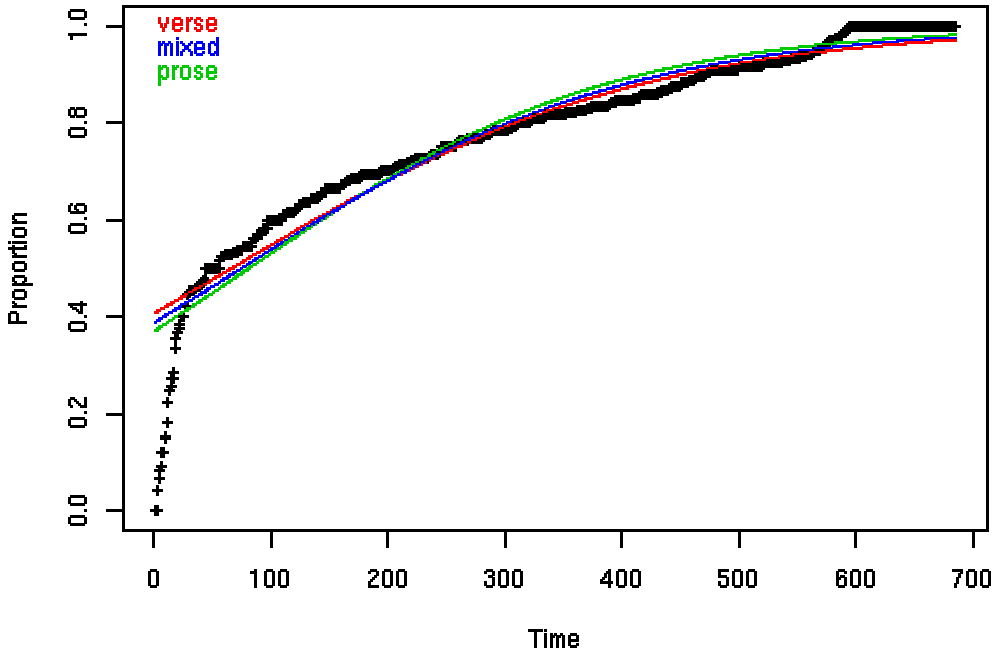
[6]This number is slightly skewed by two texts which attested over 1000 aorists each, but since the average number of aorists per text is unimportant to the study as a whole, this is acceptable. Removing these two texts from the calculation still leaves an average of almost 31 aorists per text.

The first thing that is remarkable about this plot is that it does not look much like the expected sigmoid. Specifically, the beginning of the curve rises quickly, rather than the middle as would be expected. This indicates that the time dimension as plotted is inaccurate, and some portions need to be stretched and others compressed. In terms of the chronology of the texts, the assumption had been made with this plot that texts were written at a constant rate (say, one per year). This is clearly false. It is therefore necessary to adjust the assumed date of composition to better fit the expect curve.

Before this is done, however, it is useful to consider the effect of context. Part of the Constant Rate Effect is the assumption that the context of a particular grammatic feature governs the rate of change. A useful division of contexts is the difference between metrical text (verse) and straight prose. Both these types are well attested in Pä li and in the selected texts of this study. Somewhat over half of the texts used are just prose and about 13% are verse. The remaining third are texts where the verse and the prose is intermixed. It is expected that these three types will show different curves, reflecting different usage of the innovative sigmatic aorists in the different contexts.

In order to determine an improved chronology, the constants from the logistic equation must be determined. This is done relatively easily by assuming the plotted curve is a "noisy" or partially inaccurate reflection of the true curve, which can be found with a logistic regression, a standard statistical function[7]. Because of the expected difference between verse and prose, three separate logistic regressions are run, resulting in the

Figure 3:
with regression curves

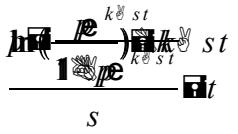


three curves seen in Figure 3.

As expected, the three types of text show different regression curves, indicating that there is a real difference in the usage of sigmatic aorists. (If there had been no difference, the smooth curves above would have been identical.) The difference is not as robust as might have been expected, though. As

[7]My thanks to Sasha Popescul for his help with this part of the statistics.

expected, furthermore, the `mixed' curve is squarely in between the `prose' and `verse' curves, reflecting correctly that mixed texts pattern partly like prose texts and partly like metrical texts.
With the curves in place and the coefficients of the logistic equation known, it is now possible to reverse the process and solve for the time factor of the equation. That is, the logistic equation from above has an equivalent equation call the logit, from which it is arithmetically trivial to solve for *t*:



<--> <-->

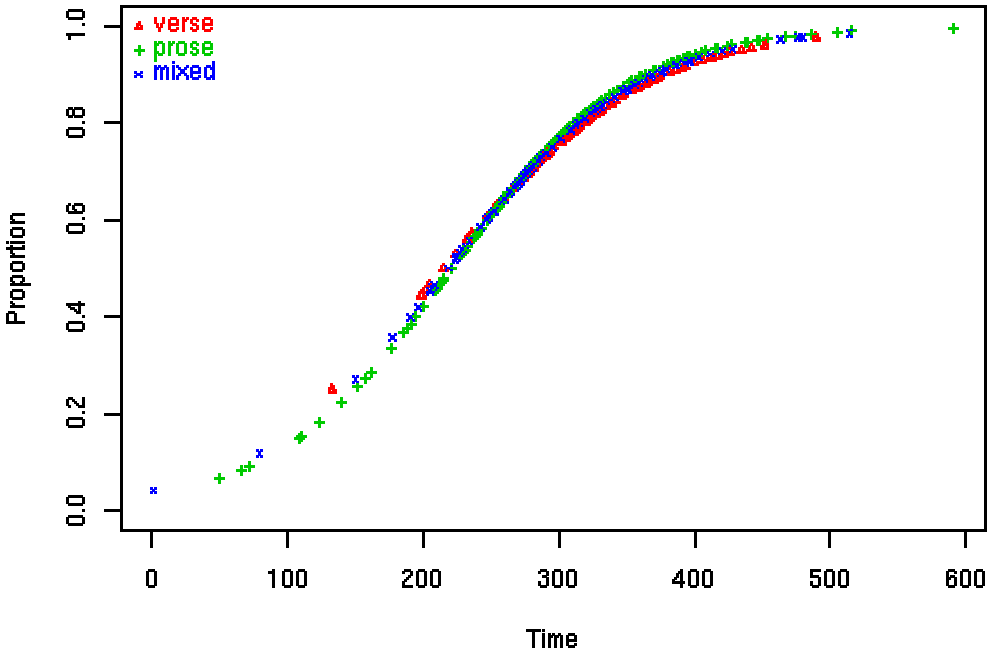
logit logistic

For each text, the proportion of sigmatic aorists to total aorists is fed into this equation. The resulting time values differ quite dramatically from the originals, for two reasons. First, many proportion values were either one or zero, indicating the total absence of either nonsigmatic or sigmatic aorists, respectively. Either of these values produces a mathematical impossibility, for 1/(1-1) is undefinable, often taken as infinite, and ln(0/1) is also undefinable, sometimes taken to be negative infinity. Thus, the ends of the graph are pushed out to infinite values. While this might seem to be a mistake or a byproduct of the statistics, it actually reflects the linguistic reality quite well. There is always a small chance that a nonsigmatic aorist could be used, even at a very late date, indicating that the end point of the grammatical change never actually arrives.[8] The inverse case is harder to understand intuitively; it suggests that there is always a possibility of producing a sigmatic aorist even before the innovation begins.[9] Fortunately there are not many of these cases--only 3, one from each of the text types. The remaining (sensical) time numbers are scaled and recentered to conform to the old time numbers, and plotted into Figure 4.

[8]In terms of a learning paradigm of linguistic change, there is always a chance that a child learner will hear a nonsigmatic aorist, incorporate it into her own language, and pass it on to the next generation. This chance becomes vanishingly small, but never completely disappears.

[9]This means effectively that it is (slightly) possible to create a sigmatic aorist, even before that becomes a grammatic formation in the language!

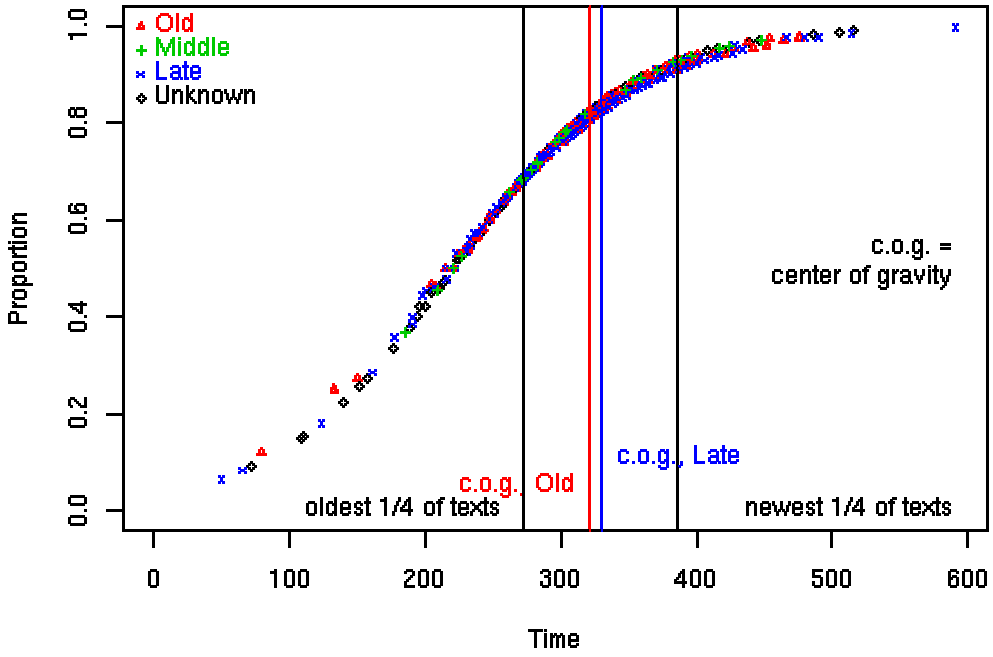
Figure 4:
with new time scale



We now have the classic sigmoid curve. Note, however, how the graph is densest somewhat to the right and above the halfway point--timestamps 200 to 400, approximately, or with proportion values between 0.6 and 0.9. This is to be expected. The movement towards the sigmatic aorist had been ongoing by the time of the Pali Canon, having begun even back in Old Indic. What is happening in Pä li is the logical conclusion of this change. The differences between the three types of text are emphasized as well.

The open question at this point is how this relates to other received wisdom about the chronology of the texts under consideration. There has been a fair amount of research (eg, Winternitz 1956, Pande 1974, Law 1974, Warder 1980, Norman 1983, von Hinüber 1996) concerning the development of the Pä li Canon, although these works are hindered either by their lack of methodological rigor or their narrow scope. These scholars have offered chronologies based on other aspects of Pali linguistics, textual criticism, theology, and internal evidence such as quotations and citations within the Canon. The coordination of these criteria is no small matter. Nevertheless, they do offer a guide to the possible chronology of the texts. On the basis of these studies I have partitioned the Canon into three strata, "old" "middle" and "late", with an additional layer of "mixed" for those texts which are probable combinations of material from multiple strata. Details on this stratification can be found in Kingsbury (2002). Naturally, there is a sizable residue of texts which cannot be assigned to any stratum and are labelled as "unknown." The texts under consideration in this study fall into the five strata in roughly the expected proportions. It is the "old" "late" and "unknown" texts which are the most interesting.

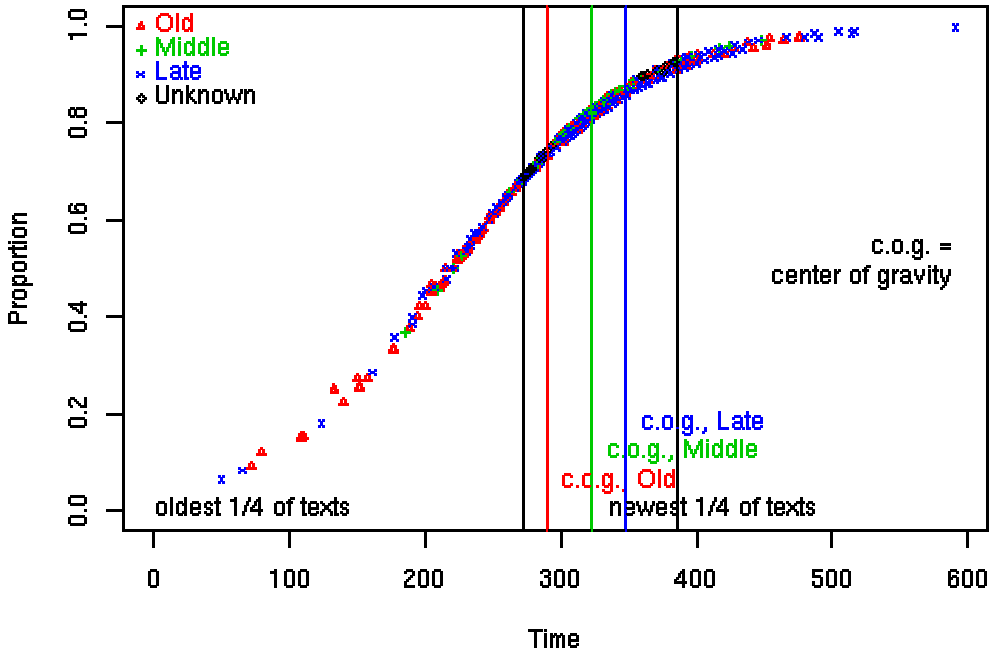
Figure 5:
strata according to secondary literature



As seen in Figure 5, the "unknown" texts dominate the field, comprising 206 of the 695 texts under consideration. The vertical lines indicate the "centers of gravity" for all texts and the texts classified as "old" or "late". Note that the "late" texts are appearing, on average, slightly later than the "old" texts, seen by the fact that the median of the "old" texts is to the left of the median of the "late" texts.[10] This difference can be exploited for reclassifying the "unknown" texts. The idea is simple: once again, the assumption is made that the proportion of sigmatic aorists is an indicator of age. The "unknown" texts with a clearly low proportion, then, can be reclassified as belonging to the "old" stratum and similarly with high proportions and the "late" stratum. The only problem is the definition of "clearly," and that is easily solved by selecting only the outying quarters of texts as defined by the induced timestamps. These divisions are indicated in Figure 5 by the black vertical lines. It might seem more intuitive to use thirds rather than quarters, since we are trying to divide the texts into three strata rather than four. This would be the wrong approach, however, since it is not clear that texts placed near the (eg) one-third mark should necessarily belong to one stratum or the other. There is a margin of error, in other words, and using the one-third and two-thirds divisions would fall within that margin. Dividing by quarters, on the other hand, avoids this margin of error and the hypothesized stratification is more trustworthy. The same applies for texts in the middle--only the middlemost quarter may be trusted. With these caveats in place, however, most of the "unknown" texts can be reclassified, resulting in the distribution seen in Figure 6.

[10]The "middle" texts fall in place between the "old" and "late", but for the sake of clarity that line is omitted.

Figure 6:
Updated chronology



This process allows 141 of the 180 "unknown" texts to be assigned to a stratum. It is important to remember that this is only a tentative chronology for these texts. The single criterion of the aorists, while more compelling than the more subjective tests, cannot be taken as the last word and must be combined with other judgements. The texts above are thus presented as a list of good candidates for further examination.

4. Accuracy

In order to make any estimates about the accuracy of this methodology a little sleight-of-hand needs to take place. That is, in an ideal world the proper stratification for each of the "unknown" texts would be known and thus that stratification could be compared to the one reached by this method. But of course, it is the very fact that the texts are of unknown date that necessitates this method in the first place! Instead, a variant of `held-out estimation' can be used. In this strategy, a number of texts which have been assigned to some stratum are temporarily reassigned to the "unknown" stratum and the procedure run as before. These `held-out' texts will then be assigned to a stratum just like any other "unknown" stratum text, and this assignment can be compared to the `correct' stratification previously given. Repeating this several times, with different texts held out for testing, allows for a good estimation of the accuracy of the procedure on the truly unknown texts.

Results, while not staggering, are encouraging. On average, texts are assigned by to the same stratum by both this methodology and by the subjective criteria 24% of the time. Since choosing randomly would produce 20% accuracy, this is not terribly impressive. However, two of the strata used by the subjective method are "mixed" and "unknown", which this method does not attempt to assign texts to. It is thus incorrect to grade this method on those texts. Using only the "old" "middle" and "late" strata raises the accuracy to about 58%, compared to a 33% (1 in 3) baseline of random selection. It must be remembered, also, that the target stratification is not without error itself. Many of the texts which were supposedly misclassified by this methodology have chronologies which are suspicious at best. For example, the texts

in the Niddesas are all assigned to the "old" stratum on the basis of their being commentaries on the Suttanipä ta, a text which is unanimously agreed to be very old. However, being a commentary on an old text does not require that the Niddesas themselves be old, and indeed many of the Niddesas were restratified into the "middle" or "late" strata. Similarly, the texts in the Suttavibhaì ga, being combinations of old rules and later explanations, are a difficult target and should realistically be assigned to a "mixed" category rather than "old." On the converse, texts which clearly belong to one stratum or another are usually correctly assigned to that stratum. For example, the texts of the Suttanipä ta are almost always assigned to the "old" stratum, rarely the "middle" stratum, and never to the "late," while the Apadä nas are usually assigned to the "late" stratum, with the remainder falling in the "middle" stratum. These 'misclassified' texts, then, could be interpreted not as errors but rather as anomalies worthy of closer attention.

5. Conclusion.

A little simple mathematics and a single assumption allows for a large number of previously unclassified texts to be assigned to a chronological stratum, by taking a small difference and exaggerating it. In theory, it is possible to use the same methodology to change all of the stratification, but it must be remembered that this study used only one grammatical change out of many, and that the chronology used was based only a probabilistic distribution. That is, it is probable that a text with a low proportion of sigmatic aorists comes from an earlier stratum, but is not necessary for that to be the case. A certain degree of variation is expected at all points in the chronology. While the effects of this variation cannot be compensated for in a study using a single linguistic variable, by combining multiple independent variables should lessen the effect of variation in any single variable. Since, however, language is in constant change, there is a wealth of appropriate variables to measure and induce chronologies from. The comparison and combination of these independent chronologies will result in a single chronology for the Pä li Canon which is both wide in scope and methodologically rigorous.

6. References

Bloch, Jules. 1877/1965. Indo-Aryan: From the Vedas to Modern Times. English edition, trans. Alfred Master. Paris: Librairie d'Amerique et d'Orient.

Fahs, Achim. 1985. Grammatik des Pä li. Leipzig: VEB Verlag Enzyklopädie.

Geiger, Wilhelm. 1956. Pä li Literature and Language. Second Edition, trans Batakrishna Ghosh. Calcutta: University of Calcutta Press.

Hinüber, Oskar von. 1977. Notes on the e-preterite in Middle Indo-Aryan. *Münchner Studien zur Sprachwissenschaft* 36: 39-48.

Hinüber, Oskar von. 1994. Selected Papers on Pä li Studies. Oxford: Pali Text Society.

Hinüber, Oskar von. 1996. A Handbook of Pä li Literature. Berlin: Walter de Gruyter & Co.

Kingsbury, Paul. 2002. The Chronology of the Pä li Canon: The case of the aorists. Unpublished PhD dissertation, Department of Linguistics, University of Pennsylvania.

Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1: 199-244.

Law, B.C. 1974. History of Pä li Literature. Varanasi: Bhartiya Publishing House.

Norman, K.R. 1983. Pä li Literature. Wiesbaden: Otto Harrassowitz.

Pande, Govind Chandra. 1974. Studies in the Origins of Buddhism. Second Edition. Delhi: Motilal Banarsidass.

Santorini, Beatrice. 1993. Phrase structure change in Yiddish. *Language Variation and Change* 5: 257-283.

Taylor, Ann. 1994. The change from SOV to SVO in Ancient Greek. *Language Variation and Change* 6: 1-37.

Warder, A.K. 1980. Indian Buddhism. Second edition. Delhi: Motilal Banarsidass.

Winternitz. 1972. A History of Indian Literature. Vol II, second edition. New Delhi: Oriental Books Reprint Corporation.