How much of an ancient language is invisible?[1]

GEOFFREY SAMPSON

Sussex University

Abstract

Statistical techniques have been developed which allow estimates of numbers of 'missing' types to be inferred from the frequency spectrum of types in an observed sample. These techniques can be used to discover how complete a picture of the vocabulary (or other resources) of an ancient language is offered by its finite body of extant documents. As an example, I apply the techniques to the vocabulary of the *Book of Odes*, one of the earliest monuments of Chinese literature.

Keywords: vocabulary size, frequency spectrum, Old Chinese

1. A not unanswerable question

For the student of any ancient language, an interesting question is: how fully does the finite extant corpus reflect the resources available to users of the language when it was alive? In principle the question applies to any aspect of language structure, but the obvious case is vocabulary. How many words did the language contain which have not come down to us, and how common were they?

These questions might sound completely unaswerable. But statistical techniques have been developed which enable estimates to be inferred from the frequency spectrum of the extant material: that is, from the numbers of forms occurring at various frequencies within that material.

This paper illustrates the potential of such techniques by applying them to one of the earliest monuments of Chinese literature, the *Shi Jing* 詩經, in English commonly called the 'Book of Odes'. This is an anthology of 305 poems composed over roughly the period 1000–600 B.C. The Chinese of that period is nowadays called 'Old Chinese' (an earlier term was 'Archaic Chinese'). I shall refer to the anthology as the *Odes*, in italics, so that I can use 'Odes' in roman to refer to the individual poems. As an example of the use of the statistical techniques in question, I hope the paper may be of interest to readers who are not necessarily students of this particular language.

2. Real and shadow texts

The relevant statistical literature is normally couched in terms of observing a 'sample' drawn from a 'population', and making inferences about the number and frequency of various 'species' in the population from the frequency spectrum of their incidence in the sample. Rather than looking at individuals of different biological species, we shall be looking at the incidence of different written Chinese graphs ('characters'), so it will be more natural to use the terms '(graph-)token' and '(graph-)type' in place of 'individual'

and ‘species’.

The ‘population’ concept also requires glossing. It is not reasonable to assume that graph frequencies in the *Odes* are thoroughly representative of usage in all written Old Chinese. Thus, the *Odes* contains many words denoting emotions, which will surely have occurred more rarely if at all in financial accounts. If we make inferences about graph frequencies from the *Odes* to a larger ‘population’, that population must consist of material comparable in genre to the *Odes*. (The idea of unseen material representing the same genre as the *Odes* is not purely hypothetical. It is known that, at an early date, the *Odes* comprised 311 rather than 305 poems; we have the titles and places in sequence of the missing six, but their text is lost – though it is by no means out of the question that some of them might emerge from present-day archaeological work. But even the full set of 311 poems can only have been a small subset of the range of poems which *could* have been written.) I shall refer to the ‘population’ we are concerned with as including real Odes and ‘shadow Odes’. We shall use statistics to make inferences from real to shadow Odes.

As for ‘genre’: suppose that, at some future date, English were represented only by a corpus of twentieth-century novels. English vocabulary contains many words from the chemistry domain: *hydroxyl*, *pipette*, *ketone*, *titration*, *lanthanum*, … If it turned out that all these words were missing from the novels corpus, we know that this would not be an accident: they are not ‘novelistic’ words. Statistics drawn from the novels corpus could not be expected to hint at their existence. On the other hand, if it turned out that *ponderous*, *axe*, *yachting*, *glibly*, *titivate*, … were missing, these would be accidental gaps – they are words as suitable to occur in novels as many others, and statistics from the novels corpus might reveal something about the extent of such gaps.

In the case of the *Odes*, genre in the sense of subject domain may not be a salient issue. The real *Odes* cover a wide range of topics, from individuals’ emotions and

relationships, through the natural world and activities of the farming year, to political events and warfare. It is not self-evident that there were subject domains in the life of the time which were not represented in the *Odes*. Specialized areas of English vocabulary like that of chemistry are a product of the very refined division of labour which has evolved in modern societies, whereas three thousand years ago it is unlikely that division of labour had proceeded so far in any society, including that of China.[2] The relation of representativeness between real and shadow *Odes* is important in another way, though. Chinese script is logographic, so that in general a word is represented predictably by one graph. But sometimes, even in the received text of the *Odes*, the choice of graph for spoken word is less predictable. For instance, 享 and 饗 seem to be used interchangeably for *haŋ 'feast'.[3] Very often, a word which has a graph of its own will alternatively be written using a graph whose primary use is to write some other (near-)homophone, for instance 難 *nân 'difficult' is also used in the *Odes* to write *nâi~nâiʔ 'ample', properly 儺, and *nanʔ~nrânʔ 'respectful', properly 戁. And Martin Kern (2005) tells us that recent archaeological finds have shown that, before script standardization late in the third century B.C., versions of the *Odes* were circulating in which word-to-graph mappings deviated from the subsequent standard far more, including the use of many graphs that did not survive, in any use, into the standard script. Thus in one manuscript *diuk 'fine, good', standardly written 淑, is instead written with a graph compounded from 弔 above 口. It is not clear from Kern's discussion whether this kind of variation represents separate written traditions each relatively consistent in itself, perhaps associated with different regions or scriptoria, one of which eventually became the standard, or whether alternatively most writings were available to most scribes, so that choices were eventually standardized on a word-by-word basis. This is an issue on which, if it cannot be resolved more directly, the statistical techniques to be discussed might shed light.

Many linguists might prefer to study vocabulary as spoken words, rather than as the graphs used to write them. However, for Old Chinese, to attempt that would raise an enormous number of debatable issues, which would get in the way of expounding the main topic of this paper. Apart from unpredictability in mappings from words to graphs, mentioned above, there are also very many unpredictabilities in the opposite direction: a single graph will often be used to stand for a range of semantically and etymologically unrelated (near-)homophones, e.g. 夷 stands for any of a set of words, all pronounced *ləi and meaning respectively 'barbarian', 'level', 'peaceful', 'easy', 'custom', and distinguishing between polysemy and accidental homophony is often difficult. Also, there are a number of alliterative disyllables, e.g. *dzûi-ŋûi 崔嵬 'craggy', which are each written with two graphs but which in linguistic terms should probably be seen as single words. For the purposes of this paper, vocabulary items will be written rather than spoken forms, and alliterative disyllables will each be counted as two separate tokens.[4]

3. Frequency spectrum and software

The received text (the 'Mao version') of the *Odes* comprises 29,720 graph-tokens, representing 2836 distinct graph-types.[5] The highest-frequency graphs (with number of occurrences, reconstructed pronunciation, and rough English gloss) are:

之	1023	*tə	object pronoun and genitive marker
不	635	*pə	not
我	592	*ŋâiʔ	me, my, us, our
其	539	*gə	his, her, its, their ~ *kə this, that
有	537	*wəʔ	have, there is/are

At the other end of the frequency spectrum, there are 802 hapax legomena (types instantiated once only). A few of these are:

飭	*lhək	set, place, arrange
焚	*bən	burn
遁	*lûnʔ ~ lûns	withdraw
穧	*dzîh ~ tsîh	sheaf, bundle
鑠	*hjauk	beautiful, fine

To aid readers who wish to check for possible calculation errors, I have put online copies of the data and software used for this paper, where these were created by me. The frequency spectrum of the *Odes* is at <www.grsampson.net/SOfofs.txt>; this file was derived using <www.grsampson.net/SGetOfofs.pl> from the Project Gutenberg edition of the *Odes* at <http://www.gutenberg.org/files/23873/23873-0.txt>. The joint probability of unseen graph-types (see sec. 4 below) was estimated from SOfofs.txt using <www.grsampson.net/D_SGT.c> (this program was produced for broader purposes, and only the first line of its output is relevant to the present paper). Numbers of 'missing types' (see sec. 5) were estimated from SOfofs.txt using the zipfR package discussed by Evert & Baroni (2007).

4. Probability of unseen graphs

The first question we can use this frequency spectrum to seek to answer is what the joint probability is of those graphs which existed in the language but never occur in the (real) *Odes*. If a graph-token were chosen at random from the shadow *Odes*, what is the probability that it would be a token of some graph-type or other not found in the real *Odes*?

To estimate this probability I use one among a family of frequency-estimation techniques that owes its origin to work by Alan Turing and his assistant I.J. Good in the course of the 1940s' cipher-breaking effort at Bletchley Park, Buckinghamshire, which played an important part both in winning the Second World War and in developing the first digital computers. This approach was first discussed in print in Good (1953), and more recently in Church, Gale, & Kruskal (1991). It is based on a theorem (which I shall not state, let alone prove, here) that relates type frequencies in a hypothetical perfectly-representative sample to type frequencies and frequency spectrum in an observed sample. For discussion of why one might prefer the Good–Turing approach to alternative estimation techniques, and for a definition and empirical assessment of the 'Simple Good-Turing' variant of the approach used here, see Gale & Sampson (1995).

Applied to the *Odes* frequency spectrum, Simple Good–Turing estimates the joint probability of graphs not found in the (real) *Odes* as 0·02699. Since the average length of the real Odes is about a hundred graph-tokens and six Odes are missing, if we set the size of the shadow Odes at 600 tokens then the technique estimates that 16·2 of them (i.e. 0·02699 × 600) will be 'new' graphs not found in the real *Odes*. (Of course, '0·2 of a graph' is a meaningless concept, but here and below I include one place of decimals to emphasize that figures are estimates – statistical 'best guesses' – rather than pieces of exact knowledge.) If we considered a set of shadow Odes equal in total size to the real Odes – 29,720 tokens – then 802·1 of them would be 'new' graphs.[6] (This shadow size is purely hypothetical. We know of a few early poems not included in the *Odes*, but Legge (1871: Prolegomena, pp. 3–4) offered evidence to suggest that the *Odes* contains the great majority of all poems circulating in China at the time.)

5. Size of unseen vocabulary

This does not tell us how many different graph-types will be represented by these

tokens. At one extreme, all eight hundred or so graph-tokens in the larger hypothetical shadow *Odes* could be tokens of the same 'new' type; at the other extreme, they might all be different from one another, i.e. each one might be a hapax legomenon within shadow and real *Odes* together. Common sense tells us that the reality is likely to be closer to the latter than the former extreme. (A graph-type occurring eight hundred times in a shadow *Odes* equal in size to the real *Odes* would be more frequent there than all but the highest-frequency graph in the real *Odes*, making it exceedingly improbable that it did not occur even once in the real *Odes*.) But we should like to produce an estimate of numbers of new types more precise than this.

This question has been examined by Harald Baayen (2001). Again I refer readers to Baayen's own writing for details of his statistical reasoning, which draws on mathematical techniques that I am not competent to expound.

Different mathematical functions have been advocated in the literature discussed by Baayen as suitable to generate good approximations, given appropriate values of free variables, to the somewhat irregular frequency spectra found in real corpora of natural language. For the spectrum of the *Book of Odes* Baayen finds that a close fit is given by Sichel's generalized inverse Gauss–Poisson model (Sichel 1971). See Figure 1, in which black bars show observed frequencies of the fifteen lowest frequencies, and grey bars the frequencies predicted using this model. The model predicts (independently of 'shadow Odes' size, i.e. as an estimate of the entire vocabulary available in the *Odes* genre) a total vocabulary of 3791·5 graph types, against 2836 observed in the real *Odes*. Thus there would be about 956 unseen graph-types: rather more, but not massively more, than the number of hapax legomena in the real *Odes*.

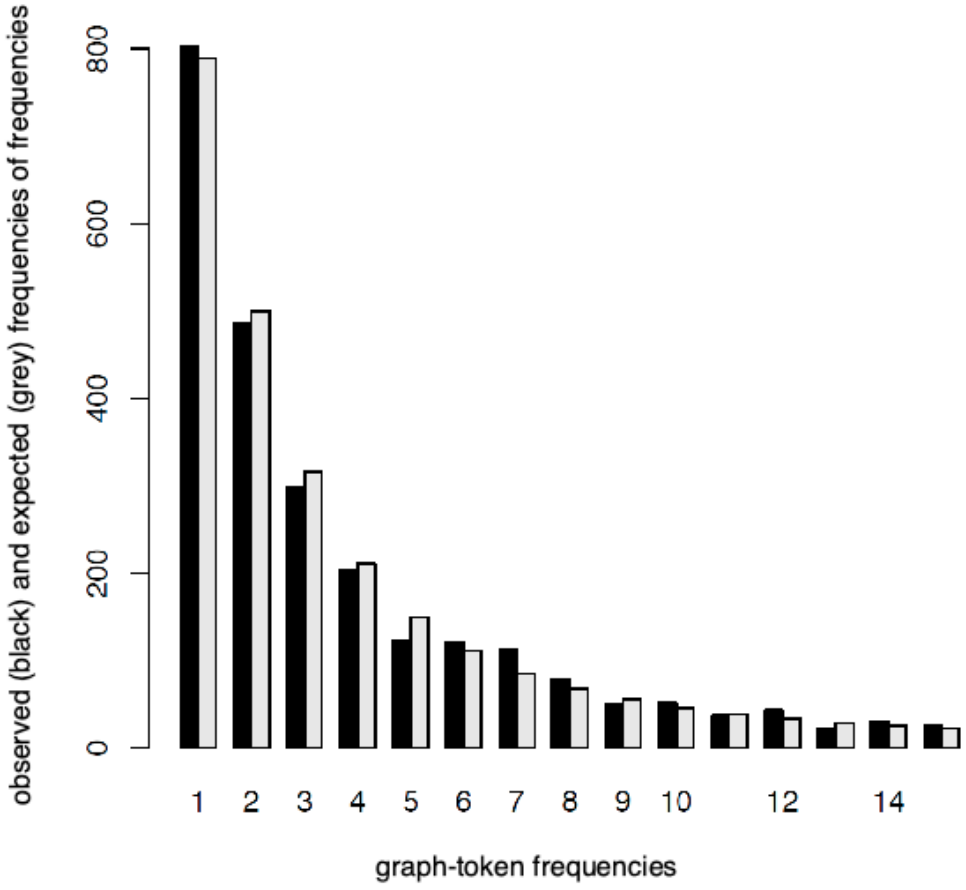

Figure 1

6. Idealized mathematical assumptions

When techniques of statistical inference are applied, it very often happens that the mathematical reasoning which justifies the techniques makes assumptions about the data possessing ideal properties which real-life data lack. But, often, it turns out empirically that ways in which reality deviates from the ideal do not damage the value of the techniques. This situation arises with the techniques discussed above, which assume that the choice of successive words in a text is affected exclusively by the probabilities of the individual words, and is independent of the identities of preceding and following words. That assumption is obviously false. Word choice is heavily constrained by

grammatical and rhetorical context. But (perhaps surprisingly) this failure to obey one of the assumptions of the techniques does not necessarily cause them to deliver misleading results. For instance, Baayen (pp. 67–9, 163) discusses experiments, using the prose of *Alice in Wonderland*, which show that the grammatical constraints of English do not in practice bias the techniques he uses there for estimating vocabulary size.

Nevertheless, we must always be alive to the possibility that new kinds of data may involve new and less innocent deviations from statistical assumptions. *Alice in Wonderland* is English prose; the *Odes* are Old Chinese poetry. As such, the latter contain many repeated phrases and lines, comparable to refrains in Western poetry – something that is not at all usual in prose. Furthermore, it is characteristic of Old Chinese grammar that adjectives in predicative position are often reduplicated, e.g. *laʔ məiʔ siau siau 予尾消消 'my tail is ragged' (Ode 155), whereas reduplication rarely occurs in European languages. It could be that frequent repetitions of these kinds undermine the statistical techniques (and it is not obvious how one might test for that).[7]

Having said that, though, there are enough cases where the techniques are robust in the face of deviations from ideal assumptions that it is reasonable to suppose, with due caution, that the results obtained here may be valuable.

7. Conclusions

In conclusion, then, what should we make of the figures obtained above – should we see them as large or as small? This is a matter of judgement rather than a question which can itself be settled statistically. But, if the number of graph-types missing from the *Odes* is on the order of a thousand, I would see that as a fairly small gap. Bernhard Karlgren's *Grammata Serica Recensa* (1957) listed all the graph-types known, at the time when that book was compiled, to have been used in the pre-Han period (i.e. before

about 200 B.C.). It contains about 6600 entries for graph-types which are distinct in the standard *kai shu* script style, ignoring entries which are included in order to illustrate earlier forms of the graphs.[8] This is more than twice the number of types in the *Odes*, so the types which are 'invisible' in the *Odes* could easily all be graph-types which are known to us from other sources. Nothing guarantees that that is so (and my guess would be that the vocabulary must have included at least a few words of which we are ignorant), but we have no reason to assume that there were more than a few unknown words.

The question whether the non-standard *Odes* editions discussed by Kern represented distinct script traditions, or different choices by scribes aware of alternative writings for individual words, is a continuum rather than sharp either–or. The statistical results above suggest to my mind that the reality lies nearer the former than the latter end of the continuum. If most scribes were aware of most alternative writings, the 'new' graphs in the shadow *Odes* ought to include many non-standard graphs for *Odes* words, of the kind discussed by Kern, in addition to (non-standard as well as standard) graphs for words which happened not to be used in the *Odes*, and one might think that these would total more than a thousand types.

We must always bear in mind that, if Old Chinese included domains of written vocabulary that were inherently unsuitable for use in poetry, comparable to chemical terminology in modern English, *Odes* statistics can tell us nothing about them. With that proviso, and with respect to this particular ancient language, I suggest that the likeliest answer to the question of my title is 'Not very much at all'.

REFERENCES

Baayen, R. Harald. 2001. *Word frequency distributions*. Dordrecht: Kluwer.

Church, Kenneth, William Gale, & J.B. Kruskal. 1991. The Good–Turing theorem. Appendix A of Kenneth Church & William Gale, A comparison of the enhanced Good–Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language* 5, 19–54.

Evert, Stefan & Marco Baroni. 2007. zipfR: word frequency distributions in R. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstration Sessions, Prague*, 29–32, and online at <www.stefan-evert.de/PUB/EvertBaroni2007.pdf>.

Gale, William & Geoffrey Sampson. 1995. Good–Turing frequency estimation without tears. *Journal of Quantitative Linguistics* 2, 217–237. Reprinted in Geoffrey Sampson, *Empirical Linguistics*, London and New York: Continuum, 2001, and online at <www.grsampson.net/AGtf.html>.

Good, I.J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika* 40, 237–264.

Karlgren, Bernhard. 1957. *Grammata Serica recensa*. Stockholm: Museum of Far Eastern Antiquities.

Kern, Martin. 2005. The *Odes* in excavated manuscripts. In Martin Kern (ed.), *Text and ritual in early China*. Seattle: University of Washington Press.

Legge, James. 1871. *The Chinese classics … vol. IV, part I, containing the first part of the She-King …* London: Henry Frowde.

Lu Chih-wei. 1960. The status of the word in Chinese linguistics. In P. Ratchnevsky (ed.), *Beiträge zum Problem des Wortes im Chinesischen*. Berlin: Akademie-Verlag, 34–47.

Pulleyblank, Edwin. 1995. *Outline of Classical Chinese grammar*. Vancouver: University of British Columbia Press.

Schuessler, Axel. 2009. *Minimal Old Chinese and Later Han Chinese: a companion to* Grammata Serica recensa. Honolulu: University of Hawai'i Press.

Sichel, Herbert. 1971. On a family of discrete distributions particularly suited to represent long-tailed frequency data. In Nico Laubscher (ed.), *Proceedings of the Third Symposium on Mathematical Statistics*, Pretoria: Council for Scientific and Industrial Research, 51–97.

FOOTNOTES

1 In carrying out the work reported here I have profited greatly from correspondence with Harald Baayen of the Eberhard Karls Universität, Tübingen, to whom I offer my warmest thanks. I alone am responsible for the finished paper.

2 Furthermore, even if some manual trade had evolved a vocabulary as specialized as modern English chemical vocabulary, because literacy was limited the specialist words might not have been assigned written graphs, in which case for our purposes they did not exist.

3 In this paper, Chinese forms labelled with asterisks will be Old Chinese pronunciations as reconstructed by Axel Schuessler (2009); those in italics will be Modern Standard Chinese 'reading pronunciations'. Schuessler calls his reconstruction 'Minimal Old Chinese' in order to underline the point that the language may well have contained further phonetic contrasts which cannot be recovered from the evidence available.

4 Later Classical Chinese contained a number of non-alliterative disyllables for exotic fauna and flora, probably borrowed from other languages, e.g. *shānhú* 珊瑚 'coral', *fènghuáng* 鳳凰 'phoenix', but these appear not yet to have entered the language by the *Odes* period (Pulleyblank 1995: 9). The modern language also has a very large number of compounds of native roots, written as sequences of graphs, which are commonly described as 'words', because they are translation-equivalents of words in European languages and (probably in consequence) are written solid in the official *pinyin* romanization system; if they are counted as words, the majority of modern vocabulary is polysyllabic. Whether in terms of language structure these forms are usefully seen as 'words' as distinct from set phrases is a longstanding discussion topic within Chinese linguistics (e.g. Lu

1960), but in any case such compounds scarcely occur in the *Odes*, and I mention them only to avoid possible confusion on the part of readers more familiar with the modern than the early language.

5 These counts refer just to the bodies of the poems themselves, excluding traditional titles and prefaces (which post-date the poems).

6 The coincidence between 802 hapax legomena in the real *Odes* and 802·1 new graphs in the shadow Odes is just that: a chance coincidence.

7 A further property of the *Odes* which is unlike modern prose and which might perhaps lead to statistical bias is that the received text of the *Odes* seems corrupt in places, possibly in consequence of the text having been reconstructed (from memory, it was traditionally believed) after the Burning of the Books in 213 B.C. For instance, the first eight graphs of Ode 107 make no sense in the context of the rest of this poem: they were perhaps copied into that Ode by mistake from Ode 203, where the identical wording fits its context.

8 Rather than tediously carrying out an exact count, I averaged counts of relevant entries in every fiftieth double-page spread and multiplied to give an estimate for the entire dictionary.