Sampling techniques in metalexicographic research[*]

Agnieszka Anuszka Bukowska
Adam Mickiewicz University, Poznan, Poland

Browsing through International Journal of Lexicography archives and other metalexicographic work one could easily notice that sampling techniques are generally neglected by metalexicographers, rarely described exhaustively by the authors themselves and almost never discussed, even though numerous researchers sample in order to make generalizations about the whole dictionary text, usually too large to be studied in its entirety. Not rarely samples consisting of one stretch only, usually selected judgmentally, are used to draw inferences about the whole dictionary text and serve as a basis for statistical analysis, which produces results of uncontrolled reliability. This study aims both at exposing the pitfalls of currently used sampling techniques and at proposing probability sampling instead.

Two basic probability sampling schemes were examined: simple random and stratified selection of pages. Censuses based on three dictionaries, three characteristics examined in each one, confirmed my concerns regarding one-stretch sampling. Simple random selection of pages produced, as expected, far more satisfying results in virtually all the cases. This can be, however, bettered by stratification in case of entry-based characteristics in larger dictionaries. Page-based characteristic, mean number of entries per page in this study, did not benefit from stratification. The smallest of my dictionaries presented a range of problems mostly connected with stratified sampling. Furthermore, empirical evaluation of sampling techniques proposed in Coleman – Ogilvie (2009) demonstrated that randomization within strata is also crucial.

1. Background

Browsing through International Journal of Lexicography archives and other metalexicographic work one could easily notice that sampling techniques are generally neglected by metalexicographers, rarely described exhaustively by the authors themselves and almost never discussed, even though numerous researchers sample in order to make generalizations about the whole dictionary text, usually too large to be studied in whole. A lot of energy is put into analyzing the samples, but very little thought seems to be given to the mechanisms of sample selection themselves. Not rarely samples consisting of one stretch only, usually selected judgmentally, are used to draw inferences about the whole dictionary text and serve as a basis for statistical analysis, which produces results of uncontrolled reliability. Such a lack of good practice is even less justifiable in view of the fact that dictionaries are fairly good sampling objects offering numerous possibilities of randomization and easy access to each and every element of their structure at virtually no cost.

This study aims both at exposing the pitfalls of currently used sampling techniques and at proposing probability sampling instead, i.e. techniques where each dictionary entry stands a chance of being included with a probability that can be determined. What makes these techniques different from predominantly intuitive approaches adopted by numerous researches is its grounding in probability theory, which makes it possible to control the reliability of the results.

Two basic schemes will be examined: simple random sampling, which in our case means

[*] **Fout! Alleen hoofddocument.**I would like to thank Robert Lew for inspiration, all the help and comments on this paper as well as Michał Jankowski and Tadeusz Piotrowski for making the SGML code extracts available for research purposes.

simply taking a random selection of pages from the whole dictionary; and stratified sampling, which consists in dividing the dictionary into non-overlapping parts called strata – e.g. letters or parts edited by different editors – and selecting a simple random sample from each one of them. Because pages are the only elements numbered in a paper dictionary, and the researcher may be interested in parameters counted on an entry basis, the pages drawn will sometimes have to be treated as clusters of entries. Therefore two additional sampling schemes will have to be considered: cluster sampling and stratified-cluster sampling. Based on these samples, estimators will be constructed. Those are functions of the sample that are supposed to yield some knowledge about dictionary parameters. Good estimators should be unbiased (meaning that there should be no difference between the estimator's expected value and the true value of the parameter), consistent and efficient. To assess efficiency, I will use confidence intervals (CIs) which with 1-α probability contain the true value of the parameter in question.

2. Current sampling practice

Most of the samples in current metalexicographic research are judgmental one-stretch samples based on what metalexicographers intuitively consider reliable and representative, usually without having tested this representativeness in any way. There is a myth that letters in the middle of the alphabet are best suitable to serve as a sample (see e.g. Miyoshi 2007:31) because lexicographers must have settled to regular modus operandi by the time they reach them. In other cases sample selection is not justified at all (e.g. Cormier 2008).

If one-stretch sampling were to yield satisfactory results, the characteristics studied would have to be evenly distributed throughout the whole dictionary which is almost never true due to changing or inconsistent lexicographic policies (de Schryver 2005, Coleman – Ogilvie 2009:2), differences in individual editorial practices in multi-editor works (Ogilvie 2008), alphabet fatigue (Zgusta 1971:352) An excellent example of inconsistencies and therefore a compelling argument against one-stretch sampling was given by de Schryver (2005). But even if lexicographers were perfectly consistent, one-stretch sampling is still very tricky as differences between dictionary parts may be due to the inherent properties of the lexicon of a given language.

Very few studies have employed techniques more elaborate than one-stretch sampling. Yet, even if multiple stretches are used, the sample selection procedure remains undocumented, even in the works of such prominent authors as Rundell (2006) or Bogaards (2008).

Systematic sampling where a starting point is selected and then every x-th page is sampled is occasionally found (e.g. in Cormier – Fernandez 2005). This method, while having an intuitive advantage of ensuring balanced coverage of the whole alphabet offers only limited potential for randomization and it must be borne in mind that '[t]he theory of probability (...) and current theories of statistical inference have little to say regarding the behavior of non-random samples, and therefore little to say regarding the confidence with which we can draw inferences from them' (Freeman 1963: 166).

Examples of techniques other than systematic sampling are scarce. Worth mentioning are two studies by Xu, both using a similar sampling technique i.e. random sampling with stratification (including post-hoc stratification) according to part of speech, word frequency and markedness of vocabulary (Xu 2008) and word frequency and part of speech (Xu 2005), and Sarah Ogilvie's 2009 study of the treatment of loanwords. Her complex design resembles stratified sampling, ensures good coverage of the alphabet and thus avoids bias towards a given donor language. Nonetheless the complexity of the design, including a series of conditional probabilities as a result of 'alternating between 'number of pages' and 'page number'' (Sarah Ogilvie, personal communication), makes it difficult to construct a theoretical model in order to check whether unbiased estimation is attainable in this case.

To the best of my knowledge only one paper to discuss sampling methodology appeared in print so far: Coleman – Ogilvie (2009). It stresses the importance of covering the whole alphabet and advocates stratification by letters and by editor in multi-editor works. Based on a census of Hotten's 1859 dictionary, the researchers empirically evaluate sampling the first 1000 and the first 10% entries of the entire dictionary as well as the first 50 entries and the first 10% of entries under each letter postulating the use of the later two as appropriate. However, these methods are not random, they exhibit a likely bias towards the beginning of each letter and additionally the third one due to differences in letter size will over-represent 'smaller' and under-represent 'bigger' letters. Unfortunately, no proposals are given to balance this over- and under-representation by constructing an appropriate estimator formula.

3. The study

As already mentioned before, the current study will propose and empirically evaluate sampling techniques that would allow to easily construct unbiased or at least asymptotically unbiased[1] estimators. I will also examine which techniques are most efficient i.e. which produce a possibly narrow confidence interval for the parameter studied.

I assume that a paper dictionary will be sampled and the discussion that follows is most directly relevant to paper dictionary sampling. This does not mean that the result will not be applicable to electronic dictionary sampling but because there is no page numbering, the designs will have to be modified. All our samples will be selected using a random number generator, with equal probabilities and without replacement. As pages are the only elements numbered in a paper dictionary, it is pages that will be drawn. Parameters characterizing pages (e.g. the number of entries per page) may be of interest, but more frequently researchers will be interested in parameters counted on an entry basis (e.g. mean number of examples per entry). In such cases, pages will be considered clusters of entries, which has its consequences for estimator formulas. Readers interested in mathematical details shall consult Barnett (1974) or Deming (1950). Additionally, I assume that cost (i.e. time) of the procedure of drawing the sample is negligible regardless of the method. Sample size (10%) and α-level (0.05) will be kept constant in all random methods for illustrative purposes.

All the samplings are supposed to be doable manually, but because of the large number of samples examined and censuses performed I am using electronic SGML-tagged versions of three existing paper dictionaries: The New Kościuszko Foundation Dictionary (NKFD) English-Polish, Webster's Revised Unabridged Dictionary (Webster), and New English-Polish Dictionary (PiotrSal). The former two are relatively large whereas PiotrSal is a small

[1] Estimators with a known bias that approaches zero when sample size increases.

dictionary. In the NKFD and PiotrSal files pagination tags were added manually, while Webster has already been provided with pagination. As these versions may differ slightly from their printed equivalents, the results do not apply directly to the aforementioned dictionaries. This shall not, however, affect the results concerning sampling techniques in any way.

The characteristics examined have to be easily searchable automatically, thus dependent on tagging. I will estimate the total number of entries, as it is often used as an auxiliary statistic and it will serve as an example of a page-based parameter. Apart from that, a number of entry-based parameters will be examined. These include 'obsolete' labeling and per-entry rate of quotations in Webster, per-entry rate of equivalent disambiguators and 'formal' labeling in NKFD, mean number of equivalents per entry and 'US' labeling in PiotrSal. While some of them might be claimed to be at least partially dependent on inherent characteristics of the lexicon, others rely solely on lexicographers' modus operandi e.g. quotation provision.

4. Results and discussion

For all of the above mentioned characteristics censuses have been performed and within-letter means have been calculated and compared with the overall dictionary mean. None of the dictionaries exhibits heavy concurrent over- and under-treatment in terms of mean number of entries per page but the distributions are far from uniform. Entry-based characteristics display more glaring inconsistencies. As will be shown, all of them can be balanced using randomization.

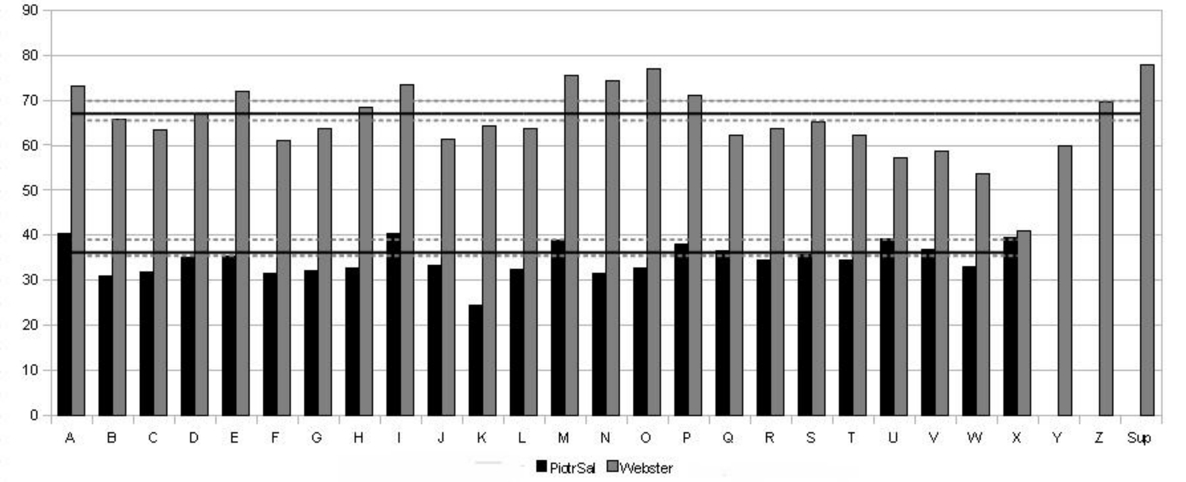

Figure 1. Mean number of entries per page in Webster and PiotrSal

First, let me consider mean number of entries per page. Figure 1 presents the distribution of mean number of entries per page throughout the alphabet (bars), the true means (black continuous lines) and simple random sampling CIs (gray dashed lines) for Webster and PiotrSal.

	true mean	min mean		max mean		max distance		min distance		simple random (		CI length	stratified CI		CI length
Webster	67,06	41,00	X	77,81	Sup	26,06	X	0,02	D	65,53	69,87	4,34	65,68	70,72	5,04
NKFD	42,58	35,66	R	49,67	Z	7,09	Z	0,00	P	39,65	42,98	3,33	41,05	44,42	3,37
PiotrSal	36,26	24,50	K	36,86	V	11,76	K	0,24	Q	35,52	39,11	3,59	33,51	34,88	1,37

Table 1. Mean number of entries per page - a summary

Table 1 summarizes details for the estimation of mean number of entries per page in these dictionaries and in NKFD. One may see that in both Webster and PiotrSal inaccurate choice of one-stretch sample might result in under- or overestimation on the order of a third of the true mean (letters X and K in Webster and PiotrSal respectively). In NKFD the maximum distance between the true mean and within-letter means is not that large, nonetheless randomization helped to achieve better results and narrow down the scope of results. In all three cases the true value of the parameter is contained in the CI. The CI length ranges between three and four entries, which I personally would consider satisfactory. A closer look at Table 1 reveals that in neither NKFD nor Webster did stratification manage to produce more efficient estimates: the CIs for stratified sampling are slightly wider. Stratification in PiotrSal proved problematic as, even though the CI is substantially narrower than in simple random sampling, it does not include the true mean (therefore those cells are shaded gray in Table 1). Here I would like to add a few comments regarding the assumptions: stratification was aimed to be proportional, however in a dictionary as small as PiotrSal the effects of rounding were no longer negligible as in larger dictionaries: e.g. letter F in PiotrSal covers 24 pages, L only 15. When taking a 10% sample both were represented by two pages. Therefore the allocation cannot be considered proportional any longer. Mind that calculations based on the assumption of proportionality and on identical data yielded a 40.00 – 41.34 CI which translated into heavy bias.

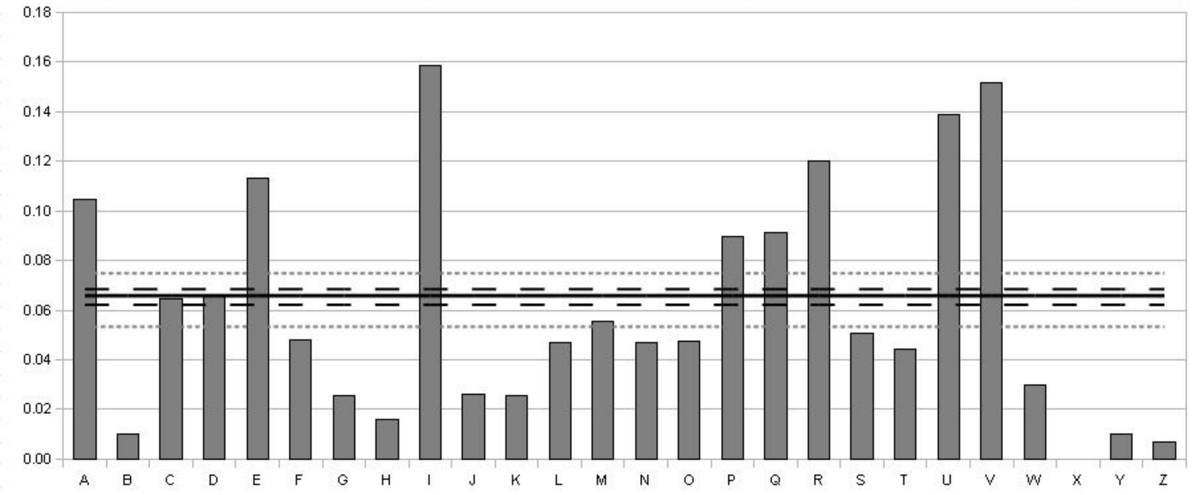

Figure 2. NKFD - “formal” labeling

Mean number of entries per page, even though not without inconsistencies, presented fairly uniform distributions when compared with entry-based characteristics. In Figure 2 one can see the distribution of ‘formal’ labels in NKFD; the least uniform characteristic in my data. Bars represent within-letter means, the gray dashed line represents the 95% confidence interval for simple random selection of pages (cluster sampling), the black fine dashed line the 95%

confidence interval for stratified selection of pages (stratified cluster sampling). Other figures presented herein will follow the same convention. Firstly, only three within-letter means (for C, D and M) fall within the CI for simple random selection of pages which is the wider one in this case. When we take the stratified CI into account this is satisfied only for letters C and D. Secondly, both CIs contain the true mean, as expected. Thirdly, the stratified CI is considerably narrower than the simple random CI (0.0063 vs 0.0215 as seen in Table 2, which translates into an increase in precision of slightly more than 340%). As I will show, this is true of any entry-based characteristic in large dictionaries.

	true mean	**min mean**		**max mean**		**max distance**		**min distance**		**simple random CI**		**CI length**	**stratified CI**		**CI length**
Webster – "obsolete" labeling	0,1485	0,0058	Sup	0,2338	U	0,1427	Sup	0,0004	Q	0,1358	0,1616	0,0259	0,1449	0,1513	0,0063
Webster – quotation provision	0,3309	0,0163	X	0,6164	W	0,3147	X	0,0029	E	0,3092	0,3771	0,0679	0,3204	0,3405	0,0201
NKFD – equivalent disambiguators	0,6699	0,2683	X	1,0448	R	0,3749	R	0,0054	T	0,6015	0,7218	0,1203	0,6517	0,6850	0,0332
NKFD – "formal" labeling	0,0658	0,0000	X	0,1583	I	0,0924	I	0,0005	D	0,0534	0,0749	0,0215	0,0622	0,0684	0,0063
PiotrSal – equivalents	2,3765	1,6847	U	2,9447	F	0,6918	U	0,0062	O	2,1404	2,3855	0,2451	2,1712	2,5560	0,3848
PiotrSal – "US" labeling	0,0239	0,0074	I	0,0408	K	0,0169	K	0,0000	T	0,0171	0,0306	0,0135	0,0095	0,0357	0,0262

Table 2. Entry-based characteristics - a summary

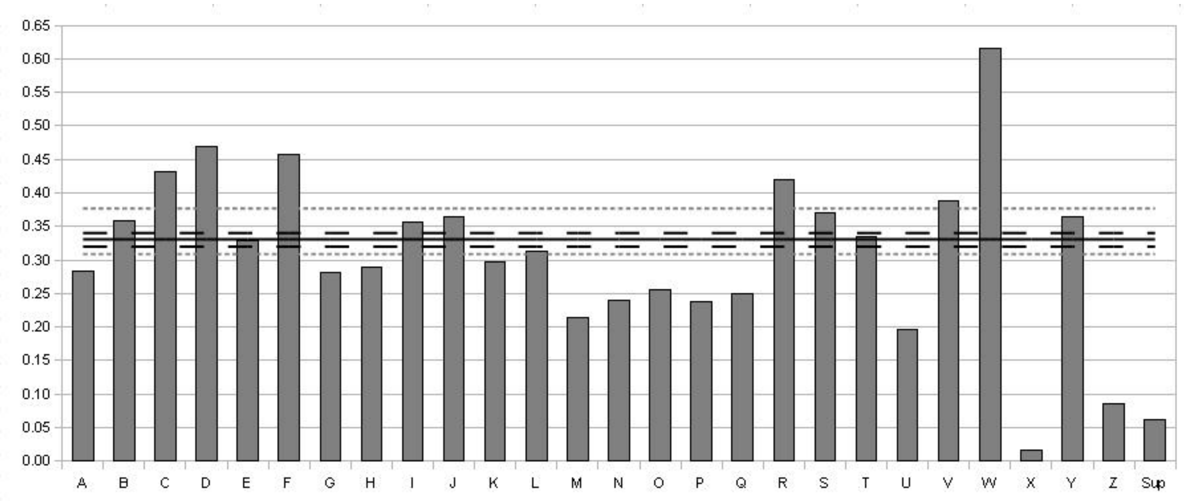

Figure 3. Webster - quotation provision

'Formal' labeling displayed most glaring inconsistencies, but as already stated above other characteristics are very unevenly distributed as well. Quotation provision in Webster presents an interesting instance as it exhibits a substantial drop in the middle of the alphabet i.e. in a place supposedly ideal for sampling. I claim that this characteristic is predominantly dependent on lexicographers' modus operandi, therefore the distribution presented in Figure 3 presents an excellent example against the myth that lexicographers settle to regular work mode by the time they reach this part of the alphabet. As in the previous case, randomization managed to cope with the variation in within-letter means. The simple random CI is 0.0679 and encompasses only seven within-letter estimates. Here again, stratification translated into considerable increase in precision (again over 340%) and the stratified CI encompasses only two within-letter means, those for E and T.

An examination of other entry-based characteristics in large dictionaries i.e. in Webster and NKFD yielded very similar results. Details can be seen in Table 2. When it comes to ‘obsolete’ labeling in Webster, it turned out that two letters contain almost no ‘obsolete’ labels: X and the Supplement, the latter should not surprise. There are also stretches with considerable over-representation of ‘obsolete’ labels: most glaring in U and Y, but prominent also in D, F and W. Here, as the distribution is a little bit more uniform, quite a lot of letters (11) fall within the simple random CI, but when we consider the much narrower stratified CI, this is true only for four letters (O, Q, S, T). In this particular case stratified CI is very narrow (0.0063, see Table 2) which translates into well over 400% increase in efficiency when compared to simple random selection of pages.

The situation is very similar in the case of equivalent disambiguators in NKFD. As seen in Table 2, there are letters that over- or under-represent the dictionary content considerably. In R the maximum distance between within-letter and true means is attained but W and U follow suit when it comes to over-representation. M, N, O, Q and especially X, Y and Z fall considerably below the true mean. In this case, stratification also translated into an increase in efficiency, this time slightly over 360%. When we take the stratified CI into consideration it turns out that few one-stretch samples can compete with this estimate (B, C, P and T).

This very lucid picture, speaking in favor of stratified sampling, gets a little blurred when we consider data from PiotrSal. This small dictionary presents a number of problems that might well be characteristic of a dictionary of this size. We have already seen that estimation of mean number of entries per page was not accurate. When dealing with entry-based characteristics I did not encounter this problem but in this case stratification did not generate better results than simple random selection of pages. As the reader may see in Table 2, stratified CI was one and a half times longer then simple random CI in the case of mean number of equivalents per entry and nearly two times longer in the case of ‘US’ labeling. This would not be much of a problem itself but ‘US’ labeling simple random estimate itself generates a very wide CI which covers 47.6% of the entire range of within-letter means (0.0074 in I to 0.0408 in K), which I doubt would satisfy any researcher. It is probably caused by both relatively small sample size, very uneven distribution (see Figure 4) and low frequency of labeling (only 31 labels in the sample).

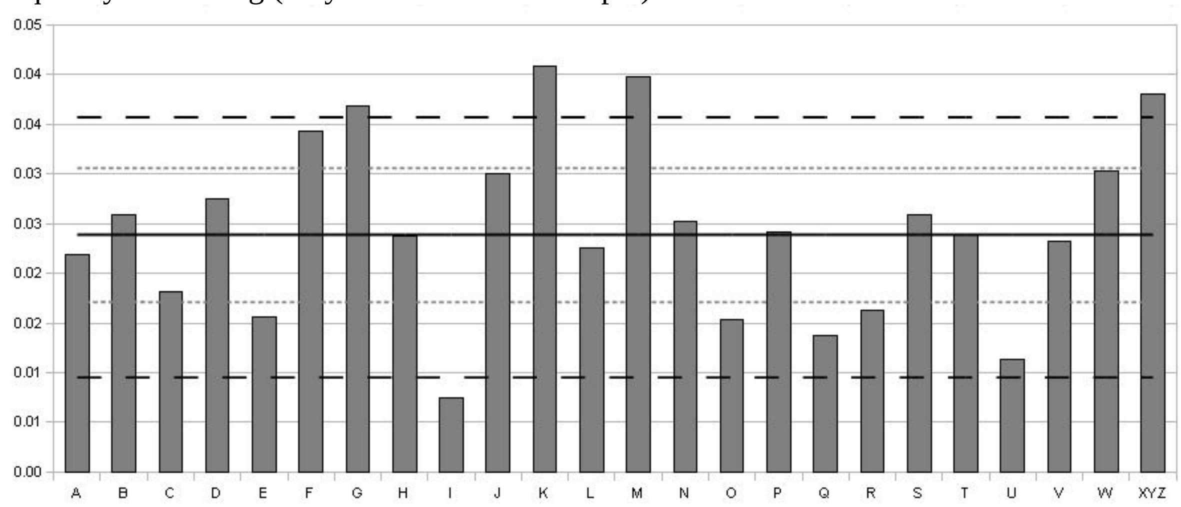

Figure 4. PiotrSal - “US” labeling

Despite its clearly unsatisfactory character, this estimate still, at least to me, presents an advantage over a one-stretch sample. Namely, it does issue a warning, clearly calling for more data. Point estimates derived from one-stretch samples can never do this. Moreover, as seen in Table 2 and Figure 4, both CIs are neatly symmetrical around the true mean which means that point estimation proved quite accurate, unlike many one-stretch samples.

Those who examined Table 2 in greater detail must have noticed that in each case there is a letter that seems to estimate the dictionary total almost perfectly. True as it is, there is one major problem with these estimates: unpredictability. D, Q and T recur in the set of best one-stretch estimates in various dictionaries but I would be rather inclined to say this is due to chance, at least I have no evidence and no intuition as to why it should not be due to chance.

I suppose many researches would be inclined to use stratified sampling in their research. Therefore I would like to address one more issue: failure to randomize within strata. Coleman and Ogilvie (2009: 10f) investigated taking the first 10% and the first 50 entries under each letter and advocated using the latter method. As already observed, neither of these methods is random, therefore I decided to address this issue empirically. As my default sample size for random sampling is also 10%, it can be compared directly with the first 10% under each letter. However, 10% in my dictionaries is always more than 50 entries under each letter. Because I want to evaluate the effect of the methods of sample selection and not that of sample size apart from taking the first 50 entries under each letter, I will also take the first x entries with such an x that the total sample size be the same as in the case of random sampling (which is of course 10% of the whole dictionary text). For the ‘first 50’ and ‘first x’ methods I will estimate the overall mean using both arithmetic and weighted means as to investigate the bias resulting from disproportional representation of various letters.

I have already raised my concern that allocating the same number of entries to each letter regardless of their original size will lead to over-representation of ‘smaller’ letters and under-representation of ‘bigger’ letters. Intuitively, the latter seems more serious as bigger letters such as e.g. C or S seem more likely to exhibit more variation than smaller ones and therefore it would be advisable to allocate more entries to those letters. In fact, the so called Neymann allocation (cf. Barnett 1974: 94ff and Deming 1950: 226ff), which has been demonstrated to be optimal, consists in allocating sample size proportionally to within-stratum variation. It appears that the Coleman-Ogilvie method is doing exactly the reverse. Using weighted mean will obviously not eliminate the loss in precision resulting from non-optimal allocation; it will, however, eliminate the bias resulting from disproportional representation of different strata. What remains is the bias towards the beginning of each letter which is obviously unknown in general.

Summary data for the Coleman-Ogilvie sampling can be found in Table 3. Estimates that fall outside the CI for stratified *random* sampling (as this was chosen as a natural point of reference) have been shaded gray. A cursory glance at Table 3 reveals that the majority of estimates were inaccurate. As I will show in a moment the picture is even bleaker than it might seem now.

	true mean	stratified CI		First 50	First 50 weighted	first x	first x weighted	First 10%
Webster – “obsolete” labeling	0,1485	0,1449	0,1513	0,1274	0,1347	0,1450	0,1520	0,1354
Webster – quotation provision	0,3309	0,3204	0,3405	0,2430	0,2352	0,3008	0,2983	0,3008
NKFD – equivalent disambiguators	0,6699	0,6517	0,6850	0,6592	0,6948	0,6760	0,6333	0,5902
NKFD – “formal” labeling	0,0658	0,0622	0,0684	0,0354	0,0357	0,0511	0,0488	0,0551
PiotrSal – equivalents	2,3765	2,1712	2,5560	2,3692	2,4108	2,3213	2,3655	2,3766
PiotrSal – “US” labeling	0,0239	0,0095	0,0357	0,0242	0,0259	0,0268	0,0302	0,0312

Table 3. Coleman - Ogilvie (2009) sampling revisited

In some cases (‘obsolete’ labeling in Webster or mean number of equivalents per entry in PiotrSal) stratification alone managed to provide remarkably better estimates than single-stretch sampling. With the former, all but one estimate are still outside the stratified CI but the distances from the true mean are not particularly large.

With quotation provision in Webster, the bias towards the beginning of the letter results in considerable under-estimation of the mean number of quotations per entry. A quick glance at Figure 3 will make us realize that despite stratification the use of the ‘first 50’ technique results in an estimate very close to that resulting from choosing the letter P i.e. one of the most serious under-estimates resulting from inaccurate choice of a one-stretch sample. Increase in sample size does help but still we are dealing with considerable under-estimation, this time erring in the region of the letter K. All those estimates fall outside the confidence interval for any random technique.

Mean number of equivalent disambiguators in NKFD also shows that the methods proposed by Coleman and Ogilvie (2009) proved no doubt more accurate than single-stretch sampling. In this particular case ‘first x’ unweighted mean turned out to be almost exactly the same as the true mean (0.6760 and 0.6699 respectively). It is interesting to note what happens if the two biases overlap: paradoxically the elimination of one source of bias (i.e. disproportional representation of different letters) resulted in a deterioration of estimates.

Coleman – Ogilvie method sometimes yields unacceptable results: in the case of ‘formal’ labeling it resulted in considerable underestimation. Here the difference between the best of these estimates and the true value is 0.107, and the estimator value in this case is almost identical with the within-letter mean in M. Table 3 also shows that these estimates fall outside the confidence interval for stratified random sampling. Obviously, one must bear in mind that ‘formal’ labeling exhibits a great deal of variation and many of the one-stretch samples would yield graver errors in estimation.

Finally, let me discuss PiotrSal. With mean number of equivalents per entry any sampling technique consisting of selecting some initial entries yielded almost ideal results regardless of sample size, allocation and estimator formula. It remains open to discussion whether this could be interpreted as a result of the relative uniformity of the distribution.

'US' labeling estimation in PiotrSal presents a very interesting instance of sample size increase having a detrimental effect on estimation. What is particularly interesting in this case is that each successive method that potentially should have been better than the previous ones results in less and less accurate estimates. We can see it first with the elimination of bias resulting from uneven allocation, then in sample size increase, and finally in changing allocation to proportional. In this case all these methods provided estimates within the confidence interval for stratified sampling, which proved to be particularly broad for this characteristic.
I would dare to draw only one conclusion based on the data presented above: Coleman – Ogilvie (2009) sampling presents a major improvement on single-stretch sampling. Beyond that it is impossible to make any generalizations. In some instances it proved accurate, as in estimating the mean number of equivalents per entry in PiotrSal; in others these methods yielded considerable but completely *unpredictable* bias.

5. Conclusions

The present research has aimed at exposing the pitfalls of one-stretch sampling commonly encountered in metalexicographic research and at examining random sampling techniques i.e. simple random and stratified selection of pages.

The censuses performed revealed that the distributions were all far from uniform and very few within-letter means came close to the true value of the parameter. Therefore one-stretch sampling presents a considerable threat to reliability of inferences drawn.

Simple random selection of pages produced, as expected, far more satisfying results in virtually all the cases. This can be, however, bettered by stratification in case of entry-based characteristics in larger dictionaries. Page-based characteristic, mean number of entries per page in this study, did not benefit from stratification. PiotrSal, a small dictionary presented a range of problems mostly connected with stratified sampling. Therefore my recommendation as for today would be to prefer simple random selection of pages in smaller dictionaries unless stratification is desired for other reasons.

Empirical evaluation of sampling techniques proposed in Coleman – Ogilvie (2009) demonstrated that randomization within strata is also crucial.

There are various limitations to the present study. First of all, it deals with estimating parameters in one dictionary only. Obviously, a researcher might be interested in comparing samples from several dictionaries. As already noted by Coleman and Ogilvie (2009: 5) the comparator text should encompass the same ranges in all the dictionaries being compared. Straightforward as it may seem, two questions remain unanswered: the treatment of differences in alphabetization and the choice of the dictionary to be randomized when the dictionaries differ in size considerably.

Second of all, this study concerns paper dictionaries. When sampling an electronic dictionary, depending on the interface, it might well be possible to take a simple random sample in the case of entry-based characteristics. At the other end of the continuum, no headword list might be available. In such a case an external list of words e.g. taken out of a corpus will be needed. There will be cases, however, when this will not suffice, in particular in the case of specialized lexicography e.g. slang or dialect dictionaries where suitable corpora are not available.

My characteristics have all been very easily quantifiable, others obviously might not. Some might argue that when the interest is mostly qualitative and not quantitative, one can allow for less rigorous sampling scheme. I would take issue with this view. Even though not expressible in terms of means or other statistics, the picture would still be heavily biased.

Bibliography

Dictionaries

New Kościuszko Foundation dictionary English-Polish. New York: The Kościuszko Foundation – Kraków: Universitas, 2003.

A Dictionary of Modern Slang, Cant, and Vulgar Words. London: John Camden Hotten, 1859.

New English-Polish, Polish-English dictionary: Part 1, English-Polish. Warsaw: Spotkania, 1999.

Webster's Revised Unabridged Dictionary. Springfield: G & C. Merriam Co., 1913.

Other sources

Barnett, V. (1974). *Elements of sampling theory*. London: The English Universities Press Ltd.

Bogaards, P. (2008). 'Frequency in Learners' Dictionaries'. In Bernal, E.; DeCesaris, J. (eds.). In *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, 1231-1236.

Coleman, J.; Ogilvie, S. (2009). 'Forensic dictionary analysis: Principles and practice'. In *International Journal of Lexicography* 22 (1). 1-22.

Cormier, M. (2008). 'Usage labels in the Royal Dictionary (1699) by Abel Boyer'. In *International Journal of Lexicography* 21 (2). 153-171.

Cormier, M.; Fernandez, H. (2005). 'From the Great French Dictionary (1688) of Guy Miège to the Royal Dictionary (1699) of Abel Boyer: Tracing inspiration'. In *International Journal of Lexicography* 18 (4). 479-507.

de Schryver, G-M. (2005). 'Concurrent over- and under-treatment in dictionaries: The Woordeboek van die Afrikaanse Taal as a case in point'. In *International Journal of Lexicography* 18 (1). 47-75.

Deming, W. E. (1950). *Some theory of sampling*. New York: John Wiley & Sons – London: Chapman & Hall.

Freeman, H. (1963). *Introduction to statistical inference*. Rearing, MA: Addison-Wesley Publishing Company.

Miyoshi, K. (2007). *Johnson's and Webster's verbal examples: With special reference to exemplifying usage in dictionary entries*. Tübingen: Niemeyer.

Ogilvie, S. (2008). 'Rethinking Burchfield and world Englishes'. In *International Journal of Lexicography* 21 (1). 23-59.

random.org [on-line]. http://www.random.org/sequences [Access date: Nov. 2009]

Rundell, M. (2006). 'More than one way to skin a cat: Why full-sentence definitions have not been universally adopted'. In Corino, E.; Marello, C.; Onesti, Ch. (eds.). *Atti del XII Congresso di Lessicografia*, Torino, 6-9 settembre 2006. Allessandria: Edizioni dell'Orso. 323-337.

Xu, H. (2005). 'Treatment of deictic expressions in example sentences in English learners' dictionaries'. In *International Journal of Lexicography* 18 (3). 289-311.

Xu, H. (2008). 'Exemplification policy in English learners' dictionaries'. In *International Journal of Lexicography* 21 (4). 395-417.

Zgusta, L. (1971). *Manual of Lexicography*. Prague: Academia.