Sampling in Historical Lexicographic Research

Agnieszka Anuszka Bukowska
Adam Mickiewicz University

1. Introduction

The vast majority of research into historical lexicography involves analysis of dictionary text. Some of the texts subject to analysis are short enough to be analyzed in their entirety, but in most cases sampling is needed. Modern sampling techniques developed by generations of statisticians are meant to ensure sample representativeness. However, metalexicographers seem to notoriously fail to apply statistically sound sampling methods in selecting dictionary text, even though many of them are relatively convenient to use.

In this article I will show that in the majority of cases sampling schemes that are actually used for dictionary text research do not conform with statistical sampling theory. Statistically sound alternatives will be presented and evaluated empirically. These include simple random selection of pages (SRS) and stratified selection of pages (SS), i.e., dividing a dictionary into groups called strata and selecting an SRS out of each stratum.

2. Literature review

2.1. Sampling in previous research

Thirty papers on historical lexicography with their publication dates spanning almost 70 years (1939–2008) have been gathered for the purpose of this analysis. The overview is by no means exhaustive or systematic, as it was extremely difficult, if not impossible, to gather examples of sampling in a systematic manner for reasons such as lack of a reliable corpus of metalexicographic texts and lack of plausible keywords that would produce the whole population of texts that employ sampling.

The current literature review is based on:

- searching the International Journal of Lexicography 1988–2010 archives for *sampl**, *random** and *letter*;
- browsing systematically through the Lexikos 1999–2009 archives, Proceedings of the Euralex International Congress (1990–2010), BudaLEX ’88 Proceedings, ZuriLEX ’86 Proceedings and LEXETER ’83 Proceedings, as well as Lexicographical archives;
- a variety of additional sources that crossed my path (mostly through references in other papers);

Therefore, the review I present below shall *not* be treated as an exhaustive and representative meta-study of sampling techniques and I will not use any methods of statistical inference in this review.

What I have found in the papers examined were predominantly single-stretch samples, i.e., samples where the lexicographer judgmentally selects a stretch, usually a single letter section, and examines it as a whole without taking other items into account. Such a sampling scheme accounts for 18 instances. Next, there are five instances of multiple-stretch sampling where more than one stretch is selected in the same manner. In four instances only was the sample stratified (i.e. Rodríguez-Álvarez and Rodríguez-Gil (2006), Ogilvie (2008), Martínez Egido (2002), Coleman (2004–2009)). Please note, however, that those stratified samples are not necessarily random and my previous research has shown that randomization within strata is also crucial (Bukowska 2010). Two instances of systematic sampling, e.g., every 10th or every 20th page, have been identified (Cormier 2008, Cormier and Fernandez 2005). In one instance it was not possible to identify the sampling scheme.

 Selected Proceedings of the 2012 Symposium on New Approaches in English Historical Lexis (HEL-LEX 3), ed. R. W. McConchie et al., 27-34. Somerville, MA: Cascadilla Proceedings Project.

It can easily be seen that non-random, single stretch sampling dominates. Statistical theory can tell us nothing about the behavior of non-random samples and I will further show that single stretch sampling presents a serious threat to research reliability.

When it comes to justification for sample scheme selection, usually nothing is provided at all. In those rare cases when researchers actually do justify their choice, most of them mention that such a sample is convenient in terms of size (e.g., Reed 1962). Interestingly, this is often accompanied by a failure to provide any information on the actual sample size. My impression is that such practice reflects a general tendency to pay more attention to sample size than to the method of sample selection.

Other reasons for sample scheme selection mentioned in some papers, most obviously Miyoshi (2007), include authority and a desire for a given piece of research to be directly comparable with other research in the field. Valid as this last point may be, such a strategy results in a vicious circle of unreliable sampling.

Finally, some authors, e.g., Miyoshi (2007), claim that it is possible to judgmentally select a representative sample as lexicographers mature in their job until they reach a given letter in the alphabet. This gives rise to the myth that letters in the middle of the alphabetic range are best suitable for sampling for this reason.

2.2. Papers dealing with dictionary sampling

Even though sampling is frequently used in dictionary text research, few papers dealing explicitly with sampling methodology have appeared in print. To the best of my knowledge, apart from my previous contribution (Bukowska 2010), only one purely metalexicographic paper has addressed dictionary sampling so far, namely Coleman and Ogilvie (2009). Their research is based on a census of Hotten’s 1859 dictionary, and includes an empirical evaluation of the following sampling schemes: 1a) initial 1000 entries; 1b) initial 10% of the entire dictionary; 2a) initial 50 entries under each letter; and 2b) initial 10% of entries under each letter. The authors concluded that 2a and 2b are most suitable for dictionary sampling, depending on dictionary size, i.e., they recommend covering the whole alphabetic range and stratifying by letter and by editor in multi-editor works. Note, however, that, while presenting a major improvement upon single-stretch sampling, none of these methods is random and my previous research (Bukowska 2010) has shown that randomization within strata is also crucial.

Surprisingly, the most valuable, in my opinion at least, contribution on dictionary sampling comes not from metalexicographers, but from researchers estimating the size of the human lexicon, most notably from Thorndike (1924) and Nation (1993). As their arguments are basically convergent, I will discuss both papers together. They advocate the use of random or systematic space-based sampling, meaning that pages (rather than entries) must be selected for practical reasons. At the same time, they warn that some space-based techniques may introduce a bias towards high-frequency entries due to inappropriate treatment of entries spanning many pages. As we will see later, my data tend to confirm this empirically: in Schmidt’s dictionaries there are entries spanning several pages which present a considerable sampling problem.

3. The study

3.1. Dictionaries used

For reasons specified in the following section the study itself involves digital, SGML- or XML-tagged versions of existing paper dictionaries. These include:

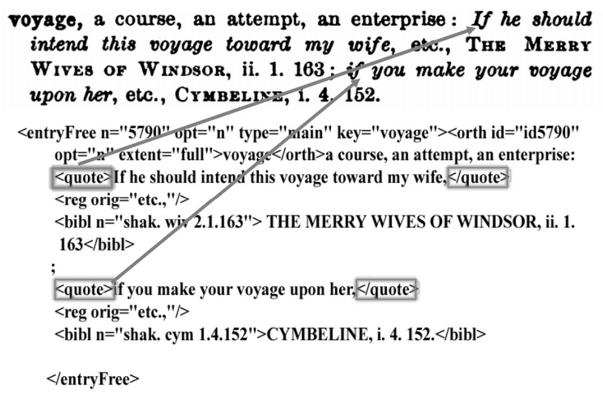
- Webster’s Revised Unabridged Dictionary — 1762 pages, approx. 118,100 entries
- Alexander Schmidt — Shakespeare Lexicon and Quotation Dictionary — 1410 pages, approx. 20,100 entries
- Charles T. Onions — A Shakespeare Glossary — 259 pages, approx. 9,800 entries
- Alexander Dyce — A General Glossary to Shakespeare’s Works — 880 pages (in two volumes), approx. 6,100 entries

3.2. Preliminaries

In the present contribution I assume that a single paper dictionary will be sampled. Although entries are usually what is of interest, pages have been drawn, as they are the only easily randomizable elements in a paper dictionary. The random.org service has been used for drawing page numbers. Sample size has been kept constant at 10%, the accepted α-level being 0.05.

The general research methodology consisted of collecting dictionary statistics based on a census and comparing these values with statistics derived from the sampling schemes under examination. For this reason, it was not possible to perform the count manually and digitized versions of the above mentioned dictionaries have been used. The following dictionary characteristics will be examined:

- mean number of entries per page;
- mean number of quotations provided per entry;
- mean number of bibliographic references provided per entry;



voyage, a course, an attempt, an enterprise : ***If he should intend this voyage toward my wife,*** etc., THE MERRY WIVES OF WINDSOR, ii. 1. 163 ; ***if you make your voyage upon her,*** etc., CYMBELINE, i. 4. 152.

<entryFree n="5790" opt="n" type="main" key="voyage"><orth id="id5790" opt="n" extent="full">voyage</orth>a course, an attempt, an enterprise:
<quote>If he should intend this voyage toward my wife,</quote>
<reg orig="etc.,"/>
<bibl n="shak. wiv 2.1.163"> THE MERRY WIVES OF WINDSOR, ii. 1. 163</bibl>
;
<quote>if you make your voyage upon her,</quote>
<reg orig="etc.,"/>
<bibl n="shak. cym 1.4.152">CYMBELINE, i. 4. 152.</bibl>

</entryFree>

Figure 1: Exemplification of quotation provision in Dyce's dictionary

For the purpose of clarity let me now exemplify quotation provision with an entry for *voyage* from Dyce's dictionary. The entry in question can be seen in Figure 1 in two variants: as it appears in print and in the corresponding XML code. As can be seen in this example, the entry for *voyage* features two quotations: one from *The Merry Wives of Windsor* and the other from *Cymbeline.* They are both clearly reflected in the corresponding XML code (see the arrows connecting relevant code elements to the surface-structure in the printed dictionary). For research purposes I used self-developed PERL scripts to obtain dictionary statistics on the basis of XML code, which means that the XML tags were what was actually counted. This implies that an assumption was made about the XML-tagging being error-free and corresponding exactly to the printed dictionary text. I believe that even though there might be slight discrepancies between the true, printed dictionary text and its post-digitized version, they are insignificant for my research goals.

3.3. Mean number of entries per page

Let me start the discussion with a simple, page-based characteristic, i.e., the mean number of entries per page. This statistic might not be of direct interest in itself, but it is often used as an auxiliary. Moreover, this is one of the very few characteristics where the primary sampling unit corresponds to the actual object of our interest and clustering is not needed.

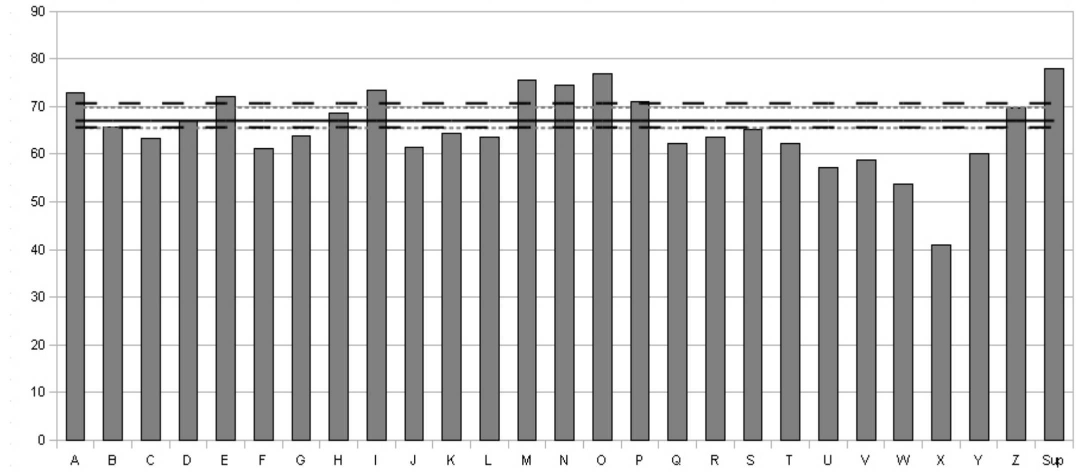


Figure 2: Mean number of entries per page in Webster

Before discussing Figure 2 in greater detail, let me first introduce a statistical concept that plays a crucial role in this piece of research, the confidence internal (CI). This is an interval centered around the *estimated* mean and which with 1-α probability, i.e., 95% in our case, contains the true value of the parameter in question. The narrower the CI the better, as this means that the position of the true mean (based on a census) can be estimated with greater accuracy. Moreover, the convention used in Figure 2 will be used further throughout the whole study, and therefore it will briefly be explained. The gray bars represent true within-letter means based on a census. They can be treated as examples of potential single-stretch (one-letter) samples. In smaller dictionaries, some of the letters have been pooled and presented together in a single aggregate bar. The black solid line stands for the overall dictionary mean based on a census, i.e., the true value of the characteristic in question. The two grey dotted lines mark out a 95% confidence interval for the mean derived from simple random sampling (SRS). The two black dashed lines represent the 95% CI for stratified sampling (SS). Of course, this piece of research presents a rather artificial situation in which the true mean is known and its position can actually be compared to the CI. In real life situations only the CI is obtained as a result of sampling and, supported by statistical theory, we are forced to believe that the true mean lies somewhere within the CI.

Table 1: Mean number of entries per page in Webster

Letter	A	B	C	D	E	F	G	H	I	J
Within-letter mean	73.01	65.67	63.36	67.04	72.00	61.07	63.73	68.50	73.35	61.36
Letter	K	L	M	N	O	P	Q	R	S	T
Within-letter mean	64.36	63.69	75.58	74.39	76.93	71.13	62.27	63.57	65.22	62.29
Letter	U	V	W	X	Y	Z	Sup			
Within-letter mean	57.26	58.65	53.80	41.00	60.00	69.60	77.81			
	Overall dictionary mean		67.06	SRS CI	65.53–69.87		SS CI	65.68–70.72		

As can be seen in Figure 2, the dictionary exhibits quite a lot of variation in within-letter means, ranging from 41 entries per page in X to 76.93 in O and 77.81 in the Supplement, with an overall dictionary mean of 67.06. Detailed statistics for all the letters can be found in Table 1. Despite this variation, both simple random and stratified sampling generate precise and accurate CIs of 4.34 entries for SRS and 5.04 for SS, which also means that no increase in precision has been observed as a result of employing stratification. As can be seen in the graph, the CI contains the true mean.

Results for the three Shakespearean dictionaries have been presented in Table 2. Letters provided in parentheses are the ones with the lowest and highest within-letter means.

Table 2: Mean number of entries per page in the three Shakespearean dictionaries

Dictionary	True mean	Range of within-letter means	SRS mean estimate	SRS CI	SS mean estimate	SS adjusted CI
Onions	38.09	27.52 (R)–46.52 (UV)	38.00	34.67–41.33	37.41	34.75–40.06
Dyce	6.96	6.09 (B)–10.17 (U)	6.93	6.27–7.59	6.70	6.05–7.35
Schmidt	14.26	9.82 (T)–25.00 (U)	13.44	11.44–14.96	15.06	13.66–16.49

First of all, it needs to be stressed that all three Shakespearean dictionaries are rather small. For this reason, the exact page count matters and the effects of rounding come into play. In order to get precise estimates in SRS, it was necessary to account for fractions of pages at the beginning and end of each letter. In stratified sampling, on the other hand, it was necessary to use weighted mean estimates, as the proportional allocation assumption no longer holds. Putting it simply, if a dictionary is large and the same fraction is being examined under each letter, an arithmetic mean of all the elements drawn could be used as an estimator. It is called proportional allocation in statistical literature. Obviously the proportionality assumption does not hold exactly, but in a large dictionary deviations from proportionality should not influence the general validity of the results obtained. Weighted estimates have been used for deriving estimates in Webster and they proved to be nearly identical to the estimates based on the unweighted means presented in this paper. In a small dictionary, however, the falsity of the assumption mentioned above comes into play. Let us imagine that a given letter only comprises 14 pages, which means that, with a sampling ratio of 10% as in the present study, in theory 1.4 should come into a sample. As it is impossible to estimate within-stratum variance for a single page, at least 2 pages must be sampled in practice. Proportionally, 2.0 presents a significant deviation from 1.4. Readers interested in statistical details and relevant formulas should consult literature on statistics such as Cochran (1953), Barnett (1974) or Deming (1950).

Second of all, as can be clearly seen in Table 2, in those dictionaries the relative precision attained is lower than in the case of Webster, i.e., the CIs are relatively wide when compared to the range of all the within-letter means. For example, SRS for Webster's dictionary yielded an estimate of the mean number of entries per page with a precision of 4.34 entries, whereas for Onions this value is not only higher in absolute terms (6.66 entries) but, above all, in relative terms as on average there are almost twice as many entries on a Webster page as on an Onions page. The same applies to all the Shakespearean dictionaries. This is due to the fact that the absolute sample size decreased as a result of decreasing dictionary size. The theory of statistical inference was initially based on the assumption that the sampled population is infinite. Thus, it is often not relative but absolute sample size that matters from the point of view of estimate precision. Striking as it may seem, a sample of 500 out of a million might yield as precise a result as a sample of 500 out of a few thousands (Cochran 1953: 17). Therefore it would be interesting to compare estimates based on the same absolute sample size for all the dictionaries and compare their precision. In my opinion, chances are that this problem would be mitigated.

3.4. Quotation provision and bibliographic references

Let us now turn to entry-based characteristics. In this case, cluster sampling will be employed, as we are no longer interested in pages, but in entries. However, for practical reasons, pages, i.e., clusters of entries will be drawn, which makes mathematical calculations more complex.

Table 3: Mean number of quotations per entry in Webster's dictionary

Letter	A	B	C	D	E	F	G	H	I	J
Within-letter mean	0.16	0.20	0.23	0.25	0.19	0.23	0.16	0.16	0.22	0.20
Letter	K	L	M	N	O	P	Q	R	S	T
Within-letter mean	0.14	0.16	0.12	0.14	0.14	0.13	0.13	0.22	0.19	0.18
Letter	U	V	W	X	Y	Z	Sup			
Within-letter mean	0.14	0.20	0.26	0.02	0.19	0.06	0.06			
	Overall dictionary mean	0.33	SRS CI	0.31–0.38	SS CI	0.32–0.34				

Let us first discuss the mean number of quotations provided per entry in Webster's dictionary. In Figure 3 and Table 3 we can see that in this case random sampling again proved more reliable than single-stretch sampling. Moreover, this example presents a convincing argument against the opinion that letters in the middle of the alphabetic range are most suitable for sampling, as there is a noticeable drop in this area.

In general, the results are similar to those presented in section 3.3., but this time stratification yields far more precise estimates than SRS (CI length of 0.07 vs 0.02).

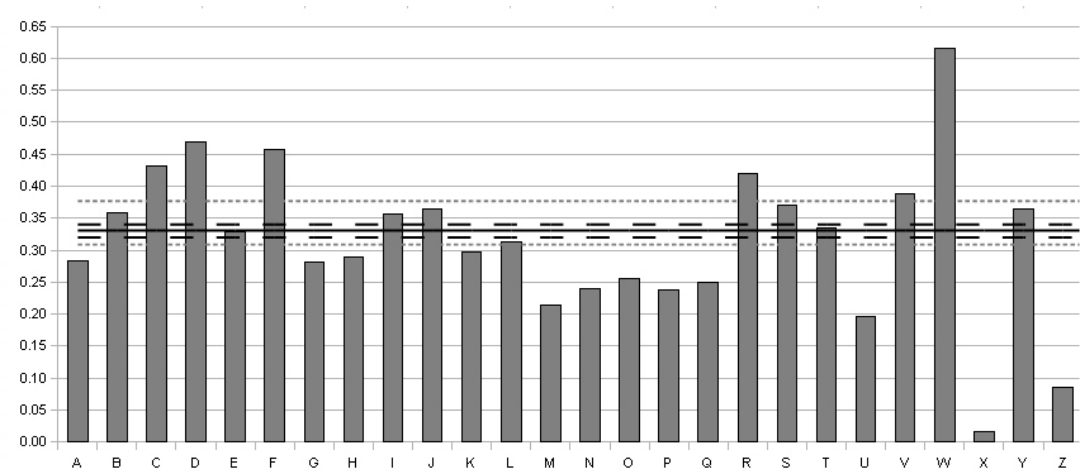


Figure 3: Mean number of quotations per entry in Webster's dictionary

In the Shakespearean dictionaries, both weighted estimators and page adjustment have been used because of small dictionary size.

Table 4: Quotation provision in the Shakespearean dictionaries

Dictionary	True mean	Range of within-letter means	SRS mean estimate	SRS CI	SS mean estimate	SS CI
Onions	1.86	1.49 (A)–2.46(T)	1.71	1.46–1.96	1.74	1.62–1.86
Dyce	2.08	1.39 (U)–2.35 (R)	2.20	2.02–2.37	2.09	2.01–2.16
Schmidt	5.16	2.66 (U)–8.24 (T)	4.42	3.13–5.71	5.28	4.94–5.62

Table 5: Bibliographic references in the three Shakespearean dictionaries

Dictionary	True mean	Range of within-letter means	SRS mean estimate	SRS CI	SS mean estimate	SS CI
Onions	3.15	2.37 (UV)–3.67 (T)	3.15	3.08–3.22	3.25	3.10–3.40
Dyce	2.35	1.63 (U)–2.60 (N)	2.32	2.15–2.48	2.36	2.28–2.43
Schmidt	11.06	4.88 (U)–17.47 (WXYZ)	11.03	6.36–15.70	12.25	10.36–14.13

In the case of quotations in Onions, both CIs contain the true mean. Unfortunately, the SRS CI is so broad that few within-letter means fall outside (F, G, H, KL, R, S, T—all of them exceeding the upper confidence limit). Stratification helped to improve upon this result with only 6 out of 18 single-stretch samples (B, C, E, M, NO, WXYZ) falling into the CI. Even though these results might not seem satisfactory to many researchers, it must be borne in mind that this dictionary only comprises 259 pages; thus the sample, expressed as a fraction of the dictionary, is small in absolute terms, which is reflected in the level of confidence with which inferences can be drawn. When it comes to bibliographic references, little difference in precision between the two estimation methods has been observed, with the CIs containing the true mean in question and being of acceptable length in both cases.

In Dyce, in the case of SRS the situation looks similar (only 9 out of 22 single-stretch estimates falling outside the SRS CI). Stratification helped to improve upon this precision level, leaving the majority of single-stretch estimates outside the CI. Similarly to the case of Onions, when estimating the mean number of bibliographic references provided per entry, both simple random and stratified sampling resulted in similar estimates, both of acceptable quality.

Schmidt, the largest of the Shakespearean dictionaries, thus expected to be the easiest to sample, proved to be the most problematic both in terms of quotations and bibliographic references. For simple random sampling, both the CIs proved extremely long (covering 46% and 74% of the single-stretch estimate ranges respectively), which I doubt would satisfy any researcher. Stratification improved on these results, yielding an acceptable, but still long, CI. A closer look at this dictionary and the data revealed that Schmidt is only seemingly a large dictionary, the number of entries there being still relatively small. What makes it large is the length of entries. When providing quotations and bibliographic references for some of his entries, Schmidt sometimes seemingly got carried away and gave a staggering 500 references, while other entries have just one or two. This variance is clearly reflected in CI length. Additionally, such mega-entries can present a source of bias, especially in stratified sampling, because when a page with a mega-entry is drawn, it automatically adds the following pages to the sample, boosting the mean and weight of this particular letter. More research on counterbalancing this problem is needed.

4. Summary and discussion

From the discussion in section 3, we can infer that random sampling techniques provide a reliable tool for dictionary text research. Out of the twelve trials presented above, in no case did the method fail (with $\alpha = 0.05$ we would expect a failure in 5% of the cases). Some researchers might argue though that there are still many single-stretch samples that provide accurate estimates of the true parameter values. True as this is, there still remains a large degree of unpredictability as to where these good estimates might lie. As we can see in tables 2, 4 and 5 above there are some letters that notoriously provide bad estimates, some of them being rather predictable (e.g., a large number of entries per page should not surprise in letter U). We might be able to guess which parts of the alphabet are likely to exhibit abnormal behavior but the reverse proves a daunting task. For this very reason, more reliable random sampling is needed.

In this paper two basic schemes have been presented: simple random and stratified sampling. The main rationale behind stratification includes a desire to improve upon precision or to make between-strata comparisons. As the latter depends on specific research questions, I will only focus on the former. As we have seen above, stratification might indeed increase the precision of estimates. This happens only in the case of entry-based characteristics and especially in large dictionaries such as Webster. One needs to bear in mind, however, that this means additional effort, uncertainty as to which estimator should be used (arythmetic or weighted), and less robustness in general.

Obviously, this study fails to cover all the sampling-related needs of metalexicographers. The next step would most likely involve comparison of two, or even more, dictionaries. The general rules outlined above would hold as well, but new problems emerge: differences in size, type of alphabetization, treatment of homographs/division into entries and a low overlap ratio between the dictionaries in question. A pilot study on the three Shakespearean dictionaries revealed that the latter two issues present a considerable problem. Despite being based on the same corpus, Dyce and Onions display only about 30 or 60 percent overlap (depending on the direction of comparison), which additionally depends heavily on the way the division into entries and senses is treated, the latter being a possible source of considerable bias. More research into these issues is needed.

References

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

Bukowska, Agnieszka (2010). Sampling techniques in metalexicographic research. In Anne Dykstra & Tanneke Schoonheim (eds.), Proceedings of the 14th EURALEX International Congress. Leeuwarden/Ljouwert: Afûk, 1258–1269.

Cochran, William Gemmel (1953). Sampling techniques. New York: J. Wiley & Sons; London: Chapman & Hall.

Coleman, Julie (2004–2009). A History of Cant and Slang Dictionaries. Oxford: Oxford University Press.

Coleman, Julie & Ogilvie, Sarah (2009). Forensic dictionary analysis: Principles and practice. International Journal of Lexicography, 22, 1–22.

Cormier, Monique. 2008. Usage labels in the Royal Dictionary (1699) by Abel Boyer. International Journal of Lexicography, 21, 153–171.

Cormier, Monique & Fernandez , Heberto (2005). From the Great French Dictionary (1688) of Guy Miège to the Royal Dictionary (1699) of Abel Boyer: Tracing inspiration. International Journal of Lexicography 18, 479–507.

Deming, William Edwards (1950). Some theory of sampling. New York: John Wiley & Sons; London: Chapman & Hall.

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

Martínez Egido, José Joaquín (2002). La obra lexicográfica de Lorenzo Franciosini: "Vocabulario italiano-español, español-italiano" (1620) [Lexicographic works of Lorenzo Franciosini: "Vocabulario italiano-español, español-italiano" (1620)]. Alicante: Fundación Biblioteca Virtual.

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

Nation, Paul (1993). Using dictionaries to estimate vocabulary size: essential, but rarely followed, procedures. Language Testing, 10, 1, 27–40.

Ogilvie, Sarah (2008). Rethinking Burchfield and world Englishes. International Journal of Lexicography, 21, 23–59.

Onions, Charles Talbut, ed. (1911). Shakespeare Glossary. Oxford: Clarendon Press.

Porter, Noah, ed. (1913). Webster's Revised Unabridged Dictionary. G & C. Merriam.

Reed, Joseph W. (1962). Noah Webster's debt to Samuel Johnson. American Speech, 37, 2, 95–105.

Rodríguez-Álvarez, Alicia & Rodríguez-Gil, María Esther (2006). John Entick's and Ann Fisher's dictionaries: An eighteenth-century case of (cons)piracy? International Journal of Lexicography, 19, 287–319.

Schmidt, Alexander, ed. (1902). Shakespeare Lexicon. Berlin: Georg Reimer.

Thorndike, Edward (1924). The vocabularies of school pupils. In J. Carelton Bell (Ed.), Contributions to education. New York: World Book, 69–76.

Selected Proceedings of the 2012 Symposium on New Approaches in English Historical Lexis (HEL-LEX 3)

edited by R. W. McConchie, Teo Juvonen, Mark Kaunisto, Minna Nevala, and Jukka Tyrkkö

Cascadilla Proceedings Project Somerville, MA 2013

Copyright information

ISBN 978-1-57473-455-3 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Bukowska, Agnieszka Anuszka. 2013. Sampling in Historical Lexicographic Research. In *Selected Proceedings of the 2012 Symposium on New Approaches in English Historical Lexis (HEL-LEX 3)*, ed. R. W. McConchie et al., 27-34. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2833.