Agnieszka Anuszka Bukowska

Sampling techniques in metalexicographic research

Praca magisterska napisana
w Instytucie Filologii Angielskiej
Uniwersytetu im. Adama Mickiewicza
pod kierunkiem prof. Roberta Lwa

Poznań, 16 May 2010

OŚWIADCZENIE

Ja, niżej podpisany/a

Agnieszka Anuszka Bukowska

student/ka Wydziału Neofilologii

Uniwersytetu im. Adama Mickiewicza w Poznaniu

oświadczam,

że przedkładaną pracę dyplomową

pt. Sampling techniques in metalexicographic research

napisałem/am samodzielnie.

Oznacza to, że przy pisaniu pracy, poza niezbędnymi konsultacjami, nie korzystałem/am z pomocy innych osób, a w szczególności nie zlecałem/am opracowania rozprawy lub jej istotnych części innym osobom, ani nie odpisywałem/am tej rozprawy lub jej istotnych części od innych osób.

Jednocześnie przyjmuję do wiadomości, że gdyby powyższe oświadczenie okazało się nieprawdziwe, decyzja o wydaniu mi dyplomu zostanie cofnięta.

Poznań, 16 maja 2010

(miejscowość, data)

(czytelny podpis)

Table of contents

List of figures

List of abbreviations

CI	confidence interval
CS	cluster sampling
LDOCE	Longman Dictionary of Contemporary English
MED	Macmillan English Dictionary
NKFD	New Kościuszko Foundation Dictionary
OALD	Oxford Advanced Learner's Dictionary
PiotrSal	New English-Polish Dictionary
PSU	primary sampling unit
SCS	stratified cluster sampling
SRS	simple random sample/sampling
SS	stratified sampling
SSU	secondary sampling unit
SU	sampling unit
Webster	Webster's Revised Unabridged Dictionary

Abstract

A careful examination of lexicographic papers reveals that sampling techniques are generally neglected by metalexicographers. Authors rarely document, still less discuss, the sampling schemes used. This is surprising in view of the fact that sampling is actually something many researchers do when they wish 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. Additionally, systematic sampling was evaluated empirically. Censuses based on three dictionaries, three characteristics examined in each one, confirmed my concerns regarding single-stretch sampling. Simple random selection of pages and systematic sampling produced, as expected, far more satisfying results in virtually all cases. This can be, however, bettered by stratification in case of entry-based characteristics in larger dictionaries. Mean number of entries per page, which constitutes a page-based characteristic 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 and Ogilvie (2009) demonstrated that randomization within strata is also crucial.

Introduction

Browsing through International Journal of Lexicography archives and other metalexicographic work it is not hard to notice that sampling techniques are generally neglected by metalexicographers, rarely described exhaustively by the authors themselves and almost never discussed even though numerous researchers engage in sampling 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 in a judgmental way, are used to make inferences about the whole dictionary text and serve as a basis for statistical analysis, thus providing results of uncontrolled reliability. As Freeman puts it "The 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). Such a lack of good practice is even less justifiable in view of the fact that dictionaries are a fairly good sampling object offering numerous possibilities of randomization and easy access to each and every element of their structure at virtually no cost.

In this thesis I will present various probability sampling techniques adapted for metalexicographic use and suitable for making controlled inferences about dictionary text.

Chapter 1: Sampling – basic terms and definitions

1.1. Basic terms and notation used

This section gives a short overview of terms and definitions connected with sampling. Although most of them will be used in Chapter 5, where I propose and evaluate random sampling techniques, these terms need to be introduced here as they are also used while reviewing current sampling practice (Chapter 2). Those basic terms will also be of use for readers wishing to consult statistics literature. This introduction will be made in general terms but nonetheless I will attempt to give some analogies to dictionary sampling. Summary tables of notation used in the thesis will also be presented.

1.1.1. Basic definitions

- Sampling – Sampling can be viewed as part of statistical practice, the process of selecting individual observations which are supposed to yield some knowledge about the total. Sampling schemes may be subdivided into two following groups:
    - Probability sampling – in probability sampling schemes, which are in the very center of the current research, every element of the population about which inferences are to be drawn has a greater than zero chance of being included in the sample, and this probability can be determined. The elements to be included in the sample are chosen at random e.g. by using a random number generator, by throwing a dice, etc.

    - Non-probability sampling – in contrast to probability sampling, in this case not all elements have a chance of being selected, or the probability of them being selected cannot be determined. This includes *convenience sampling*, e.g. choosing a school nearest the researcher's home to obtain a sample of students, and judgmental *sampling* where the choice is made based on what the researcher considers to be representative and suitable for the study. In metalexicographic research, judgmental samples based on a single dictionary stretch (one letter usually) are commonly encountered. See sections 2.1.1. and 4.1 for more details.

True random sampling is more than merely selecting units to be studied: "[s]ampling is the science and the art of controlling and measuring the reliability of useful statistical information through the theory of probability" (Deming 1950: 2).

- Sampling frame – a list of all the sampling units that constitute a universe from which a sample is obtained or even more generally "any device by which the N sampling units are identifiable one by one" (Deming 1950: 76). In dictionary sampling it can be e.g. the list of all pages, the list of all entries etc. Because random number generators are the most convenient tool to draw a sample with, I will aim at obtaining a numbered list to serve as a sampling frame.
- Sampling unit (or SU) – As already mentioned above, sampling units are elements in the sampling frame. "Sometimes the sampling units may be the individual members of the study population. Often this is not so and the sampling frame is a coarser subdivision of the study population, with each unit containing a distinct set of population members. (Barnett 1974: 8)". See section 3.2 for more information on selecting a SU in dictionary research.
    - Primary sampling unit – henceforth PSU – in multi-stage sampling those are units drawn at the first stage of the sampling procedure.
    - Secondary sampling unit – henceforth SSU – in multi-stage sampling those are units drawn at the second stage of the sampling procedure.
    - Tertiary sampling unit – defined analogically to PSU and SSU
- Population – "A statistical population is to be thought of as a set of values" (Wood et al. 1986: 48). This is a set of not only the values actually observed but also those potentially observable. It must be borne in mind that population does not equal sampling frame. Let us consider a straightforward lexicographic example: we are in-

terested in exemplification rate in a given paper dictionary. The population in such a case is a set of entries with examples attributed to them. Nonetheless, pages are likely to be drawn if a paper dictionary is sampled, so the set of all dictionary pages will constitute our sampling frame.

- Estimator – it is a function of the sample used to estimate an unknown population parameter. Good estimators have the three following characteristics: they are unbiased, consistent and efficient (see below).
- Bias – in formal terms this is the difference between the estimator's expected value and the true value of the estimated parameter. Obviously, if the expected value of the estimator equals the true parameter value, the estimator is unbiased. Preferably, unbiased estimators are to be used, but nonetheless there are instances where using a slightly biased estimator is justified e.g. because the unbiased one is far less efficient and the bias is known and negligible in larger samples[1]. Using non-probability sampling schemes is always a source of uncontrolled bias.
- Consistency – an estimator (or a sequence of estimators to be precise) is consistent if it converges in probability to the population parameter i.e. if for all $\varepsilon>0$ (no matter how small) $\lim_{n\to\infty} P(|(t_n-\Theta)|<\varepsilon)=0$ where P stands for probability, t_n is the sequence of estimators, n – sample size and Θ - the population parameter. In simple words, if increasing the sample size increases the probability of the estimators being closer and closer to the true population parameter, then the estimator is consistent.
- Efficiency – an estimator is efficient if it has a possibly low mean square error (MSE = variance + bias squared). If the estimator is unbiased, the smaller its variance the more efficient it is. Smaller variance generally means that more precise inferences about population parameters may be obtained i.e. confidence intervals are narrower. In this research various sampling methods will be evaluated in terms of estimator variance i.e. in terms of precision reached for a given sample size. Later in this thesis I will use the term precision, which in my opinion is more intuitive, interchangeably with efficiency.
- Confidence interval (CI) – an interval that with a $(1 - \alpha)$ probability includes the true population parameter. Its length is directly proportional to the population standard deviation and inversely proportional to sample size. Obviously, one wants the CI

[1] i.e. the estimator is asymptotically unbiased – the bias approaches zero when sample size increases.

to be as narrow as possible. Saying that a dictionary's mean number of entries per page is between 30 and 40 is less informative than saying it is between 34 and 36. Consequently, for a constant sample size, the smaller the variation, the more precise the estimate.

- Simple random sampling (without replacement) – henceforth SRS – This is the simplest possible scheme, just imagine N balls in a bowl from which one wants to draw n balls, each ball appearing only once. A straightforward lexicographic example would be drawing n words from a numbered list of N words. In this thesis I will use the term *simple random selection of pages* which, even though pages are drawn as in simple random sampling is not identical to SRS. This is a broader term since entries will very often be of interest and therefore pages will be treated as clusters of entries. Formally speaking, such a scheme is called cluster sampling (see below).
- Systematic sampling – imagine that we are in possession of a complete list of population members. In metalexicography a list of pages in a paper dictionary and a list of entries are almost always available. Taking a systematic sample consists of choosing a starting point and working progressively through the list in some regular manner. In lexicographic research it would be e.g. taking every 20th page of a given dictionary or every 20th entry. Please note that this scheme is not an instance of probability sampling with the possible exception of the starting point (if randomly chosen), but even in such a case the rest of the sample is chosen in a deterministic way.
- Stratified sampling – (SS) – Imagine that the population of interest is divided into non-overlapping groups called strata. Stratified sampling consists of treating each stratum as a sub-population and choosing a sample (usually a SRS) independently from each stratum. In metalexicography this could be e.g. randomly choosing 10% of entries under each letter.
- Cluster sampling – (CS) – This method is in a sense the reverse of stratified sampling. The population of interest is again subdivided into non-overlapping groups, but this time only some of these groups are selected and examined in whole.
- Multi-stage sampling – is a modification of the CS scheme. Imagine that one does not examine the selected cluster in whole but again sub-samples them (usually using

SRS). This would be two-stage cluster sampling. One can however proceed with sub-sampling to obtain multi-stage cluster samples.

- Stratified cluster sampling – henceforth SCS – This method is a hybrid of SS and CS. It is basically similar to SS but within each sub-population a CS and not a SRS is taken.

Please refer to Appendix 1 for a summary table of the notation used throughout this thesis.

Chapter 2: Current sampling practice

Most of the samples in current metalexicographic research are judgmental single-stretch samples based on what metalexicographers consider reliable and representative, usually without having tested this representativeness in any way. Were the dictionaries compiled in a perfectly consistent way, such techniques would be more justifiable. This is however rarely the case. As Coleman and Ogilvie put it: "Few dictionaries are consistent in the application of lexicographic policies, but this need not be presented as a flaw: good lexicographers learn from experience, remain flexible in their practice, and adapt their policies to the needs of each entry."(Coleman and Ogilvie 2009: 2). An excellent example of inconsistencies and therefore a convincing argument against single-stretch sampling is given by De Schryver (2005). But even if the lexicographers were perfectly consistent, single-stretch sampling would still be very tricky as the variance between different dictionary parts may be due to the inherent properties of the lexicon of a given language.

In the sections to follow I give a short overview of current sampling practices with particular focus not on the most commonly found practices but on research displaying more sophisticated sampling techniques.

2.1.1. Single-stretch sampling

A sampling method consisting in selecting a single stretch of the dictionary text and examining it in whole is, intuitively speaking, the one used in the majority of cases. Usually this stretch consists simply of one letter of the alphabet. Various justifications are given for such sample selection: e.g. Miyoshi (2007) samples letter L because it was

used in previous research in the field and because of its convenient size. There is also a myth among metalexicographers that letters in the middle of the alphabet are best suited to serve as a sample because lexicographers must have settled to a regular work mode by the time they reach this part of the alphabet. However, there are researchers who simply decide to start with letter A (e.g. Roberts 2007: 283) and do not justify their choice at all (see also Cormier 2008). Statistical formalists would discredit the method on the grounds of its non-random character which, formally speaking, makes any use of inferential statistics impossible. There are many cases in science where mathematical assumptions are only roughly satisfied or even neglected for practical reasons but single-stretch dictionary sampling, in order to give satisfactory results, would require the assumption that the characteristics studied are uniformly distributed throughout the whole dictionary which is almost never true for several reasons including: changing or inconsistent lexicographic policies (De Schryver (2005), Coleman and Ogilvie (2009)), changes in editorial staff (Ogilvie (2008)), dictionary fatigue (Zgusta (1971) as quoted in De Schryver (2005 :60)) and finally characteristics inherent to the lexicon structure of a given language.

2.1.2. Systematic sampling

From a purely theoretical point of view this sampling method is no different from single-stretch sampling as described above provided that both the letter in single-stretch sampling and the first page in systematic sampling are chosen at random. This is because systematic sampling is just cluster sampling with the number of selected clusters m=1 (Barnett 1974: 121). However, this method can yield better estimates as it provides good coverage of the whole alphabet.

This method, though not as popular as single-stretch sampling, is also widely used. In Cormier and Fernandez (2005) every 20th page, starting at page 5 selected at random, is sampled. Unfortunately, this sample was used only in part of their research. They also constructed a second "control" single-stretch sample (to let – to lighten) consisting of 100 entries. The authors claim any other sampling method would be unmanageable as the dictionary text was not scannable.

2.1.3. Stratified sampling

Apart from single-stretch and systematic sampling I found three instances of stratified sampling: two studies by Xu and one by Sarah Ogilvie. In Xu (2005) and (2008) random stratified sampling according to word frequency and part of speech can be found. As the author states, "[t]he selected entry words were further balanced within word-classes" (Xu 2005: 293), so the sampling scheme also involved post-hoc stratification.

Stratified sampling can also be found in Ogilvie (2008). A complex scheme is used in order to ensure good coverage of the alphabet and avoid 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, p.c.), makes it difficult to construct a theoretical model in order to check whether unbiased estimation is attainable in this case.

2.1.4. Sampling in Coleman and Ogilvie (2009)

To the best of my knowledge only one paper to discuss sampling methodology has appeared in print so far: Coleman and 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 four sampling schemes: taking 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. The researchers advocate the use of the later two as appropriate, making the choice dependent on dictionary size. "These results also demonstrate the importance of matching sample size to purpose: as the samples are chopped into ever smaller pieces their reliability decreases. In a bigger dictionary 10% of entries under each letter would be a more reliable sample than the first 50 entries, but for a small dictionary a 10% sample gives unreliable results (Coleman and Ogilvie 2009: 9f)". The researchers also advocate grouping letters together in a small dictionary before stratifying, which seems commonsensical. 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. Unfortu-

nately, no proposals are given to balance this over- and under-representation by constructing an appropriate estimator formula. I will evaluate those methods empirically in 4.3. .

Chapter 3: The study: preliminaries

3.1. Preliminary assumptions

Generally it will be assumed that a paper dictionary is to be sampled and the discussion that follows in Chapter 5 concerns mainly paper dictionary sampling. This does not mean, however, that the result are not applicable to electronic dictionary sampling but because of lack of page numbering the designs will have to be modified. Because of the large number of samples analyzed, I performed automatic search and count using self-developed Perl scripts but I assume that all the samplings proposed are doable manually as well. Additionally, the following assumptions concerning the sampling procedure will be made:

- Cost (i.e. time) of the procedure of drawing the sample is negligible regardless of the method thus the cost of the whole research is directly proportional to the sample size.
- Alpha level is kept constant at 0.05.
- Where possible, sample size is kept constant for illustrative purposes. I decided arbitrarily that my samples will consist of a 10% of the dictionary text. A sampling scheme consisting of a simple random selection of pages will always be treated as a basis for comparison with other methods, which will all be evaluated in terms of precision of the estimates i.e. the length of the confidence interval. The narrower the CI, the more precise estimates are. This does not mean that the reverse procedure is not possible i.e. optimizing the sample size (thus costs) to obtain a given precision. In section 5.1.2 I will demonstrate how to calculate sample size needed for reaching

desired precision based on pilot sampling. Assuming the procedure is similar for all sampling designs, it will be exemplified with SRS only.

- Sampling will be made with equal probabilities (or at least it will aim at obtaining equal probabilities).
- All samples are drawn without replacement independently from one another. Sampling is performed using a random number generator (http://www.random.org/sequences) which is claimed to offer truly random numbers.
- In systematic sampling the starting point is always selected at random.
- I assume that a page consists of all the entries beginning thereon, so that all the entries, including those spanning two (or even more) pages stand a chance of being included in the sample.

Another important assumption needs to be made when performing stratified sampling and censuses broken down by letters with a dictionary where a new letter does not start with a new page (as was the case with two of my dictionaries – see 3.3. Dictionaries used in the study). In such a case I doubled those pages, thus increasing the total number of pages in a dictionary. However, when performing other types of sampling I stuck to the real page numbering and if a bordering page was drawn, entries under both letters were included in the sample. This doubling has little[2] effect on entry-based characteristics, but certainly is a source of bias in the case of page-based characteristics. I believe this bias is negligible in larger dictionaries.

However, in small dictionaries this bias might not be negligible anymore. Therefore I am going to test another, more accurate approach which consists in measuring the proportion of pages allocated to the two letters in question with a ruler and use appropriate fractions in the calculations. In Figure 1 one can clearly see that there are differences between the two approaches in a small dictionary. The black bars represent within-letter mean number of entries in PiotrSal, which is one of the studied dictionaries (see section 3.3. Dictionaries used in the study) and which consists of 440.3 pages only (when ruler-adjusted). The overall mean value of 36.26 entries per page is marked with a black hori-

[2]As will be shown in Chapter 5, the total number of pages appears in the formulae for variance in SS, CS and obviously CSC as well. Stratum weights for the estimator formula in SS are also construed using the total number of pages in a dictionary. Therefore page doubling does have an effect on entry-based characteristics but it is not as noticeable as in the case of page-based characteristics.

zontal line. When the bordering pages get reduplicated, their overall number increases to 458, thus reducing the overall mean to 34.91 (gray horizontal line). The within-letter means (gray bars) also differ significantly, especially in the case of "small" letters such as K or V. There differences are, however, easily noticeable in the graph in the case of *any* letter when the ruler adjustment actually did introduce any changes. Hereinafter each graph for this particular dictionary will include ruler-adjusted values only.

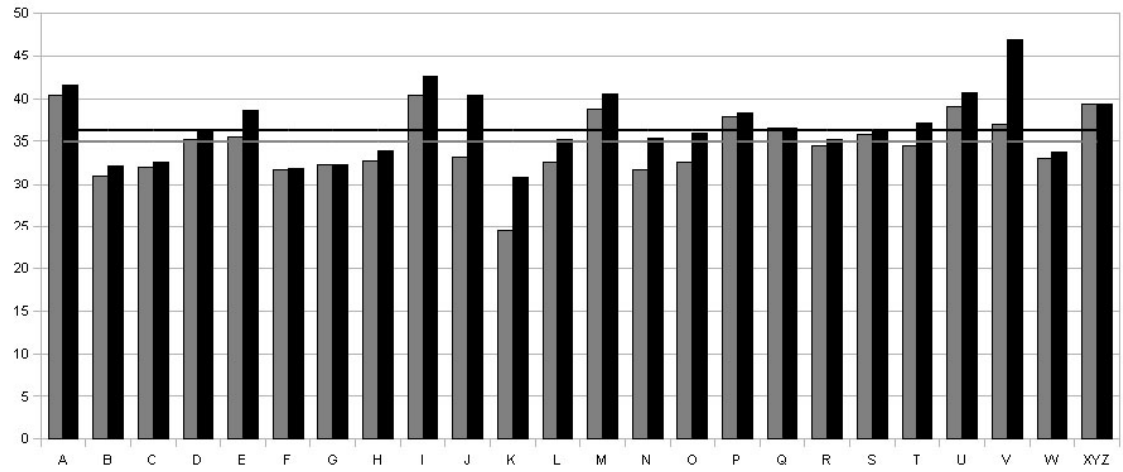

Figure 1: Page adjustment in PiotrSal

3.2. Choice of sampling unit

In metalexicography researchers are most likely to be interested in statistics per dictionary entry. Nonetheless, when sampling a paper dictionary it is usually impossible to construct a sampling frame with which it would be possible to identify every single entry[3]. A numbered list of pages is, however, usually available and all the sampling frames will predictably use these natural units. Depending on the type of characteristic studied, there are two possibilities: in a simpler case dictionary pages can serve as the ultimate SUs, in a more complex one the ultimate SUs must be dictionary entries. In the latter case natural clusters of entries i.e. pages will still be drawn.

[3]Unless the entries are numbered in a dictionary under consideration as e.g. *Słownik Synonimów* by Dąbrówka, Geller & Turczyn.

3.2.1. Page as sampling unit

When each dictionary entry either possesses a given characteristic (1) or not (0) then our natural ultimate SU may be the dictionary page. The variable examined will then be the number of entries on a given page possessing the characteristic of interest. It is a convenient situation mostly because the mathematical description of sampling procedure is relatively simple: it allows e.g. taking an SRS from the whole dictionary. Moreover, population size is always known as it is simply the total number of pages.

Unfortunately such a case may prove rare. It seems that e.g. etymologies may satisfy the condition but usage labels may not as several senses of the word may be labeled.

Of course, when estimating the size of a dictionary, the page will naturally serve as the ultimate SU.

3.2.2. Single dictionary entry as sampling unit

When every dictionary entry possesses a given characteristic but to a different degree (starting with 0), there is no other choice but to accept entries as the ultimate SUs. In most cases, a complete and numbered list of all the dictionary entries is unavailable when sampling a paper dictionary so again pages will be drawn. In this case, however, every sampling must be considered CS, which spells both more complicated mathematical description and, unfortunately, very often biased estimators. Using entries as ultimate SUs will usually be necessary when estimating e.g. number of examples or usage labels per entry.

There is, however, a way of getting around this obstacle other than simply performing cluster sampling. An external numbered word list e.g. from a corpus would be needed to serve as a sampling frame. Apart from enabling SRS, it also makes it possible to use stratified sampling with strata according to criteria not a priori identifiable in the dictionary e.g. word frequency (cf. Xu 2008). It must be borne in mind that words that have been drawn may be absent from the macrostructure of the dictionary studied causing non-response problems. Non-response, which in this case would mean that a given entry drawn from an external list is absent from the dictionary macrostructure, will not

be very troublesome if the task is to draw inferences about the dictionary text alone. If, however, the researcher wants to extrapolate his or her result beyond the dictionary text, it must be borne in mind that non-response may be a serious source of bias because the entries not included in the dictionary may differ significantly from those present in the macrostructure, e.g. there may be more neologisms among them.

It is also worth mentioning that e-dictionary user interfaces may sometimes offer the possibility of copying the word list and numbering it, or even filtering it, e.g. according to part of speech. The most recent versions of LDOCE, OALD and MED display such possibilities (Xu, p.c.).

Despite its inconvenience, CS in metalexicography does have some advantages over taking an SRS with entries as SUs. Namely, when doing comparative research with more than one dictionary it allows the researcher to consider both what is included in the dictionaries as well as what is not.

3.3. Dictionaries used in the study

As already stated above, all the samplings are supposed to be doable manually, but because of a 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).

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.

Now I will present what SGML-tagged versions of the dictionaries in question look like. Below I present an entry for “flank” from Webster

(1) <! p. 567 !>

<hw>Flank</hw> (flă&nsm;k), <pos>v. t.</pos> [<pos>imp. & p. p.</pos> <u>Flanked</u> (flă&nsm;kt); <pos>p. pr. & vb. n.</pos> <u>Flanking</u>.] [Cf. F. flanquer. See <u>Flank</u>, <pos>n.</pos>, and cf. <u>Flanker</u>, <pos>v. t.</pos>] <sn>1.</sn> <def>To stand at the flank or side of; to border upon.</def>

<blockquote>Stately colonnades are flanked with trees.</blockquote> Pitt.

<sn>2.</sn> <def>To overlook or command the flank of; to secure or guard the flank of; to pass around or turn the flank of; to attack, or threaten to attack; the flank of.</def>

There is no specific tag marking the beginning of the entry, but Webster does not have run-on entries so entries are in fact equivalent to headwords. Each headword is marked with a pair of <hw> and </hw> tags. After a section containing information on part of speech, morphology, pronunciation, cross references etc., unfortunately without proper structural tagging, follows a definition tagged with <def> and </def> pair. First sense of a given word is not marked, the following are marked using the sequence <sn>sense number.</sn> as one can also see in the example above. As seen from this example, tagging in Webster in only partially structural. Tags such as <block-quote> - beginning of quotation or <def> - beginning of a definition are purely structural, but unfortunately some tags are typographical (e.g.) and parts of the entry are tagged only using those typographical and not structural tags as e.g. a cross reference to the noun "flank" in the example above. As the reader will probably realize, this is not SGML-tagging proper and presents serious limitations when it comes to automatic searching. These limitations had a considerable influence on the choice of characteristics to be studied (see section 3.5. below). Nonetheless I decided to include this dictionary in the study for several reasons: its availability, age and size.

In the NKFD and PiotrSal files pagination tags were added manually before the first entry that appeared in full on a given page. Webster had already been provided with pagination, but some of the pagination tags were moved if they were originally in the middle of an entry. You can see the page tag in example (1) above just before the entry for "flank". I used the same tagging in the remaining two dictionaries.

Below, in (2) you can see the entry for "flaccidity" in NKFD.

(2) <ntry main>
<hdwd> flaccidity
<pron>
<psgr>
<posp> n.
<gram> U
<sens>
<tran> wiotkość
<tran> zwiotczałość
<sens>

<usge style> przen.
<tran> słabość

NKFD is tagged 100% structurally, which made it easy and convenient to automatically count virtually anything countable in the dictionary. There are no closing tags there. Each main entry starts with <ntry main>, then follows the headword <hdwd>, pronunciation, which I have not replicated here due to the presence of phonetic symbols. Thereafter we always find <psgr> - part of speech description obligatorily containing <posp> - part of speech tag but information on grammar and morphology can also be found there. In the "flaccidity" example we see that it is a noun (<posp> n.) and that is uncountable (<gram> U). Then follow the equivalents. In this case there are two sense subdivisions (<sens>) with two equivalents provided for the first entry, and one for the second. Each equivalent is preceded by a separate <tran> tag. In the second sense we also find a usage label <usge style> przen. informing the user that this sense is figurative in meaning. In Figure 2 you can see what this entry actually looks like in the dictionary.

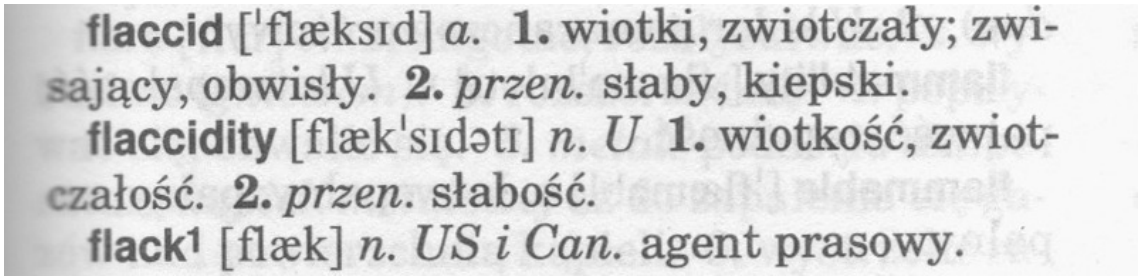
flaccid [ˈflæksɪd] *a.* **1.** wiotki, zwiotczały; zwisający, obwisły. **2.** *przen.* słaby, kiepski.
flaccidity [flækˈsɪdətɪ] *n. U* **1.** wiotkość, zwiotczałość. **2.** *przen.* słabość.
flack[1] [flæk] *n. US i Can.* agent prasowy.

Figure 2: NKFD – sample entry

In PiotrSal tagging is similar but obviously not identical. A sample entry for "saloon" in the SGML-tagged version of PiotrSal is replicated in (3) below.

(3) <entry>
<headword>
<spl>saloon</spl>
</headword>
<hwinfo>
<pos>N</pos>
<pronunciation>
<phonetic>s{\e}{\`}lu:n</phonetic>
</pronunciation>
</hwinfo>
<syntactic>
<semantic>

<subsense>
<trans>limuzyna</trans>
</subsense>
</semantic>
<semantic>
<senseinfo>
<label>
<lang>BR</lang>
</label>
</senseinfo>
<subsense>
<trans>salon</trans>
</subsense>
</semantic>
<semantic>
<senseinfo>
<label>
<lang>US</lang>
</label>
</senseinfo>
<subsense>
<trans>bar</trans>
</subsense>
</semantic>
</syntactic>
</entry>

The main difference in tagging between NKFD and PiotrSal is the existence of closing tags in the latter one. Apart from that it has a very similar structure. Each entry starts with <entry>, then follows the headword. Within this tag there may be more than one spelling version each tagged with <spl>. Then we have a block with information on pronunciation and part of speech. In this case the word is labeled using two geographical labels: <lang>BR</lang> in the second sense informs us that this word means „salon" in the British variety of English, similarly <lang>US</lang> stands for American English. Each equivalent is marked with a separate <trans> tag. Figure 3 shows what it looks like in print. Please note that information on part of speech does not surface in print. I will come back to tagging when discussing characteristics that I used in my study.

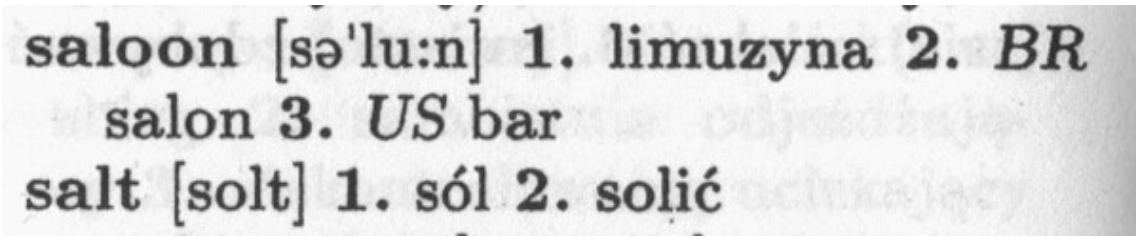

Figure 3: PiotrSal – sample entry

NKFD was the only dictionary where each letter began with a new page. In Webster, pages containing letter boundaries were doubled for the sake of performing stratified sampling according to the assumptions mentioned in 3.1. above. As the dictionary is quite large, I do believe the bias resulting from doubling is negligible. In contrast, bearing in mind the small size of PiotrSal I used fractions of pages in order to get a more reliable and unbiased estimation.

3.4. Procedure

As already mentioned, all the searches and counts were performed automatically. I used self-developed Perl scripts that counted the tags and saved the results in a .csv file. Apart form the script moving the tags marking the beginning of a new page in Webster just before the nearest tag marking beginning of a new entry if they originally happened to be in the middle of an entry, there were basically two scripts: one designed for page-based characteristics, the other one for entry-based characteristics. As input, both received a randomized list of page numbers copied from the (http://www.random.org/sequences) service or a complete list of page numbers in the case of a census. The first script simply produces a .csv file including page numbers followed by the respective count of a given characteristics. All the statistics were then done using a spreadsheet. The second script did more than that. With the same input it produced two output files: one auxiliary file containing information on page number, entry number and a relevant count; another file with a preliminary summary of these data i.e. for each page it provided the information on the number of entries thereon, the mean for the characteristics in question on a given page and variance of this variable.

3.5. Characteristics examined

The characteristics examined have to be easily searchable automatically, thus dependent on tagging. For all three dictionaries I will estimate the total number of entries, as it is often used as an auxiliary statistic and it will serve as the only example of a page-based parameter (with dictionary page as the ultimate SU). For all three dictionaries we have

already seen what the tag marking the beginning of a new entry looks like, so I will not repeat this information here. Apart from that, a number of entry-based parameters will be examined. As each dictionary has different tagging and it is not the aim of this study to compare the dictionaries, there will be a separate set of characteristics for each dictionary.

In Webster, the per-entry rate of quotations will be examined as a characteristic dependent predominantly on lexicographers' modus operandi. Quotations are marked using the <blockquote> tag, which you can see in the sample entry for "flank" presented in (1). Counting these tags means that all quotations were included, regardless of the attribution or the lack thereof. Some entries have been provided with more than one quotation which means that we are dealing with a truly entry-based characteristic. In (4) below one can see the first sense of the entry for "wade", for which two quotations have been provided. Altogether this entry has been provided with as many as five quotations in three senses. This does not mean that one could not make a page-based characteristic out of it i.e. counting the proportion of entries that have been provided with at least one quotation.

(4) <hw>Wade</hw> (?), <pos>v. i.</pos> [<pos>imp. & p.p.</pos> <u>Waded</u>; <pos>p. pr. & vb. n.</pos> <u>Wading</u>.] [OE. waden to wade, to go, AS. wadan; akin to OFries. wada, D. waden, OHG. watan, Icel. va&?;a, Sw. vada, Dan. vade, L. vadere to go, walk, vadum a ford. Cf. <u>Evade</u>, <u>Invade</u>, <u>Pervade</u>, <u>Waddle</u>.]

<sn>1.</sn> <def>To go; to move forward.</def> [Obs.]

<blockquote>When might is joined unto cruelty,
 Alas, too deep will the venom wade.</blockquote> Chaucer.

<blockquote>Forbear, and wade no further in this speech.</blockquote> Old Play.

Apart from quotation provision I will also look at "obsolete" labeling in Webster. Even though there is no specific structural tag, this label is always enclosed in square brackets which makes it easy to extract automatically. "Obsolete" label can also be seen in the above cited entry for "wade"; the first sense defined as "To go; to move forward" is labeled "obsolete" by [Obs.]. Again, entries can obviously have more than one "obsolete" label as the one for "yard" presented in (5) below, in which three out of six senses and one collocation ("under yard") have been labeled as "obsolete".

(5) <hw>Yard</hw> (?), <pos>n.</pos> [OE. yerd, AS. gierd, gyrd, a rod, stick, a measure, a yard; akin to OFries. ierde, OS. gerda, D. garde, G. gerte, OHG. gartia, gerta, gart, Icel. gaddr a goad, sting, Goth. gazds, and probably to L. hasta a spear. Cf. <u>Gad</u>, <pos>n.</pos>, <u>Gird</u>, <pos>n.</pos>, <u>Gride</u>, <pos>v. i.</pos>, <u>Hastate</u>.]

<sn>1.</sn> <def>A rod; a stick; a staff.</def> [Obs.] P. Plowman.

<blockquote>If men smote it with a yerde.</blockquote> Chaucer.

<sn>2.</sn> <def>A branch; a twig.</def> [Obs.]

<blockquote>The bitter frosts with the sleet and rain
 Destroyed hath the green in every yerd.</blockquote> Chaucer.

<sn>3.</sn> <def>A long piece of timber, as a rafter, etc.</def> [Obs.]

<sn>4.</sn> <def>A measure of length, equaling three feet, or thirty-six inches, being the standard of English and American measure.</def>

<sn>5.</sn> <def>The penis.</def>

<sn>6.</sn> (Naut.) <def>A long piece of timber, nearly cylindrical, tapering toward the ends, and designed to support and extend a square sail. A yard is usually hung by the center to the mast. See Illust. of <u>Ship</u>.</def>

<col>Golden Yard</col>, or <col>Yard and Ell</col> (Astron.), <cd>a popular name of the three stars in the belt of Orion.</cd> -- <col>Under yard</col> [i. e., under the rod], <cd>under contract.</cd> [Obs.] Chaucer.

In NKFD I will examine “formal” labeling and the mean number of equivalent disambiguators per entry. Both are illustrated with the entry for “ignite” as seen in (6) and in Figure 4. In this entry, the second sense “rozpalać (się), zapalać (się)” is labeled “formal”, whereas for one of the equivalents in sense 1. i.e. for “wzbudzać” an equivalent disambiguator has been provided informing the user that it can collocate e.g. with controversy (“kontrowersje”).

(6) <hdwd> ignite
<pron>
<psgr>
<posp> v.
<sens>
<usge style> form.
<tran> rozniecać
<tran> wzbudzać
<tlin> np. kontrowersje, niechęć
<sens>

<usge style> form.
<tran> rozpalać (się)
<tran> zapalać (się)
<sens>
<usge dom> chem.
<tran> prażyć

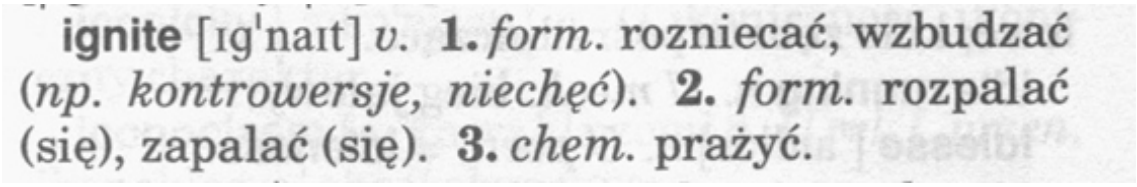
ignite [ɪgˈnaɪt] *v.* **1.** *form.* rozniecać, wzbudzać (*np. kontrowersje, niechęć*). **2.** *form.* rozpalać (się), zapalać (się). **3.** *chem.* prażyć.

Figure 4: NKFD sample entry

The characteristics studied in PiotrSal will include "US" geographical labeling and mean number of equivalents per entry. We have already seen what the tags for both these characteristics look like. Recall that "US" geographical labeling is tagged as "<lang>US</lang>" whereas each equivalent is tagged with a separate <trans> tag. The reader might go back to (3) and Figure 3 if necessary. Here I will just provide an example showing that "US" geographical labeling is in fact a truly entry-based characteristic. It is indeed rare for an entry, especially in a small dictionary like PiotrSal, to have more than one identical geographical label. Nonetheless it is not impossible; the entry for "clerk" in this dictionary has been divided into four sub-senses out of which two ("sprzedawca" and "recepcjonista") have been labeled "US".

Chapter 4: Evaluation of non-random sampling techniques

4.1. Empirical evaluation of single-stretch sampling

As already mentioned before, there is no way of assessing such a sample selection in a theoretical way. Therefore I will now proceed to an empirical evaluation of single-stretch dictionary sampling based on a complete count of the test dictionaries. All the graphs presented herein represent within-letter mean values of the parameters in question based on censuses, all the bars can be treated as nothing else but judgmental single-stretch samples. We already saw above, and more examples will follow, that the characteristics in question can be unevenly distributed throughout the dictionary. Therefore I will not discuss those graphs in detail now but compare them with systematic sampling below.

4.2. Empirical evaluation of systematic sampling

Beside single-stretch sampling, systematic sampling is a technique used with some frequency in metalexicographic research. As Barnett notices “there is also some sort of intuitive appeal in systematic sampling: it seems to 'span the population' in a way that might lead to more 'representative' results than those obtained from random choice” (Barnett 1974: 121). We will see that, provided the starting page is chosen at random, systematic sampling is formally equivalent to CS with just one cluster sampled, which would mean that variance estimators based on systematic sampling are always biased

(see Ardilly and Tillé (2006: 188)). Let us suppose we sample every M-th page. Then we can think of our dictionary as organized in the following way:

$$(7)\quad \begin{matrix} X_1 & X_{M+1} & X_{2M+1}\cdots \\ X_2 & X_{M+2} & X_{2M+2}\cdots \\ \vdots & & \\ X_M & X_{2M} & X_{3M}\cdots \end{matrix}$$

Where for all i X_i designates the i-th dictionary page. It should now be clear that in systematic sampling one row, which constitutes a cluster, is chosen. Despite its intuitive appeal as a method covering the whole alphabetic range, there is little mathematical evidence to back these intuitions up. Nonetheless, systematic sampling can yield effective estimators and, as shown in Ardilly and Tillé (2006: 189), it happens when for *each* cluster the estimated variable is highly dispersed around the mean. These results should be intuitively clear. In fact, the intuitive representativeness depends largely on the way in which the list of SUs is sorted. To illustrate the point let me consider two extreme non-lexicographic examples. First, imagine our list is sorted in either decreasing or increasing manner. In this case systematic sample mean will be a good estimate of the population mean. But the other extreme case is when the arrangement of values on our list resembles a sinusoid and we happen to choose the starting point close to one of the function's extremes and a sampling interval roughly equal to its period: then the results will be highly skewed. Of course, in the case of a dictionary such extremes will probably not be encountered and systematic sampling may produce good results in practice.

Below I present graphs illustrating how systematic sampling worked with my data. As all the graphs herein will follow the same convention, I will briefly outline their structure. Bars illustrate the within-letter mean values of a given characteristic. The continuous black line represents the true mean value of the parameter in question based on a census. Dashed gray lines represent endpoints of the confidence interval for a given sampling method. The formula for calculating variance and, what follows, CIs is identical to the one used in SRS (cf.5.1.1.). Sample mean value will not be included in the graph for clarity's sake. It follows from the formula for calculating confidence intervals that it is always the midpoint of this interval.

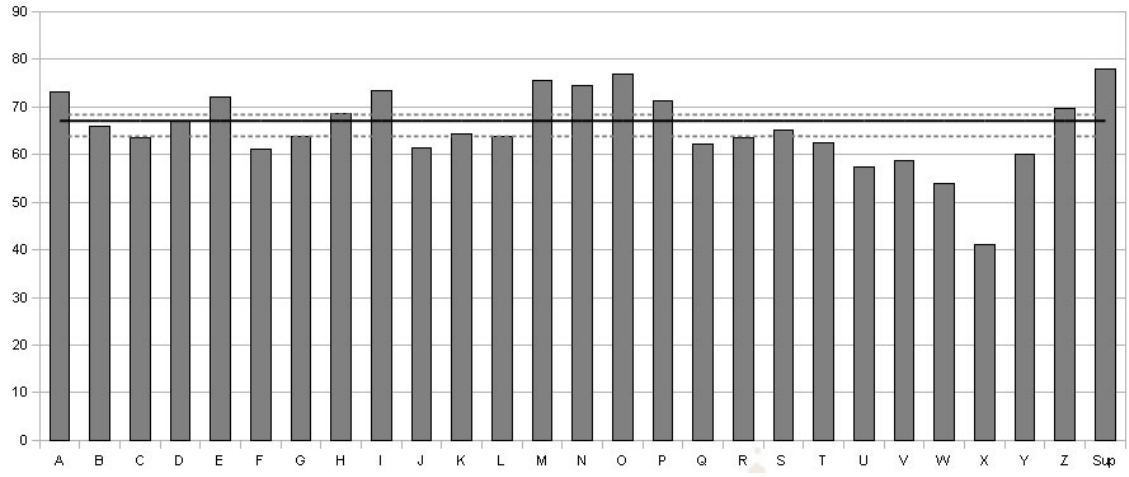

Figure 5: Systematic sampling – entries in Webster

Figure 5 represents the systematic sampling estimate of Webster's mean number of entries per page. As one can see, even though it has quite a uniform distribution, systematic sampling proved more accurate than many of the single-stretch samples consisting of one letter examined in whole. Sample mean is 66.10, whereas the true mean number of entries per page in Webster is 67.06 and it lies within the systematic sampling CI which has a length of 4.53.

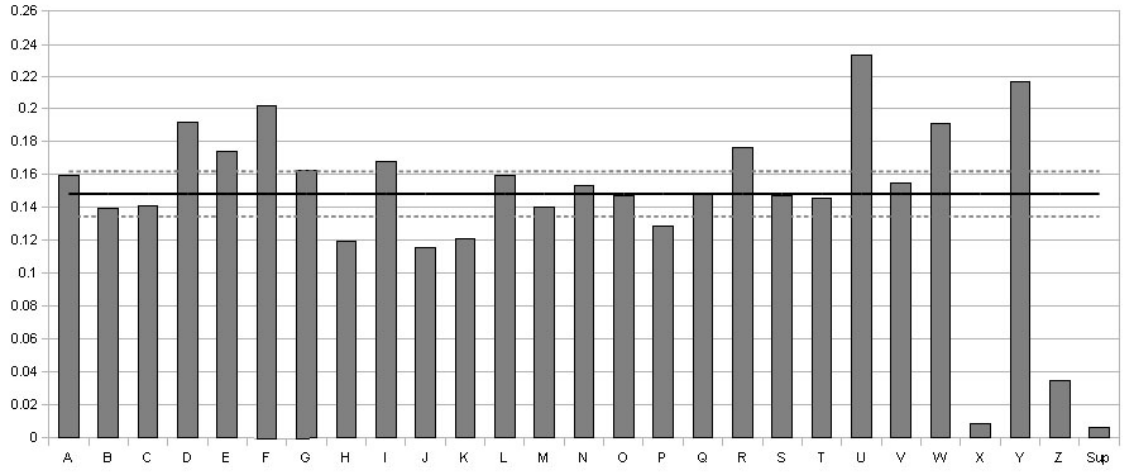

Figure 6: Systematic sampling – "obsolete" labeling per entry in Webster

Some characteristics, however, may be more unevenly distributed throughout the dictionary as it is the case with both "obsolete" labeling and quotation provision in Webster. As one can see in Figures 6 and 7, very few within-letter means come close to the true dictionary mean. In the case of "obsolete" labeling they range between 0.0058 (in the Supplement section) and 0.2338 (in U) labels per entry, with the mean value of

0.1485. Systematic sample mean is surprisingly close to this value (- 0.1484), but the CI proved quite wide (0.0274). We will see whether other sampling designs can improve the precision.

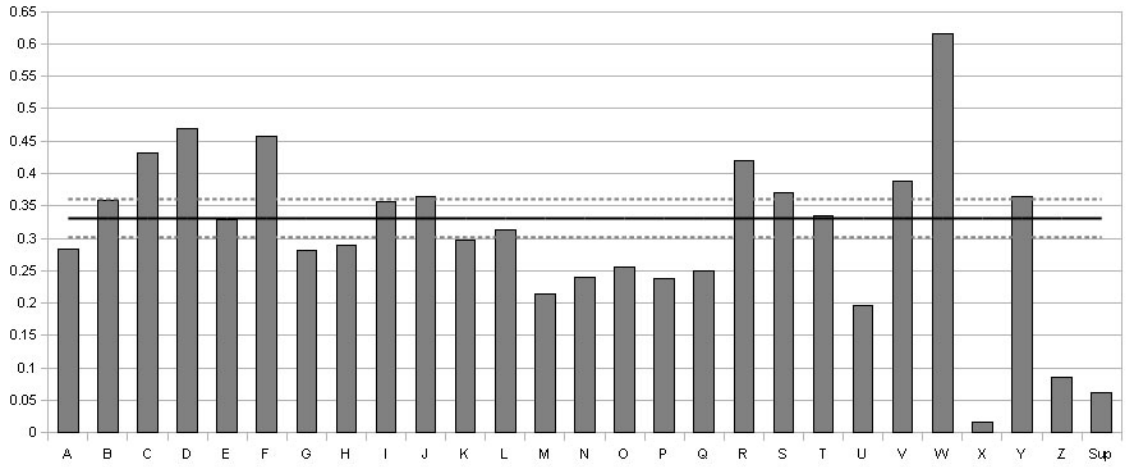

Figure 7: Systematic sampling – quotation provision per entry in Webster

In the case of quotation provision per entry the situation is similar. The distribution looks even more uneven and the letters in the middle of the alphabetic range exhibit some of the lowest within-letter means in the dictionary. As quotation provision is predominantly dependent on lexicographers' modus operandi, these data provide a counter-example for the assumption that the middle of the alphabetic range should best represent the dictionary structure.

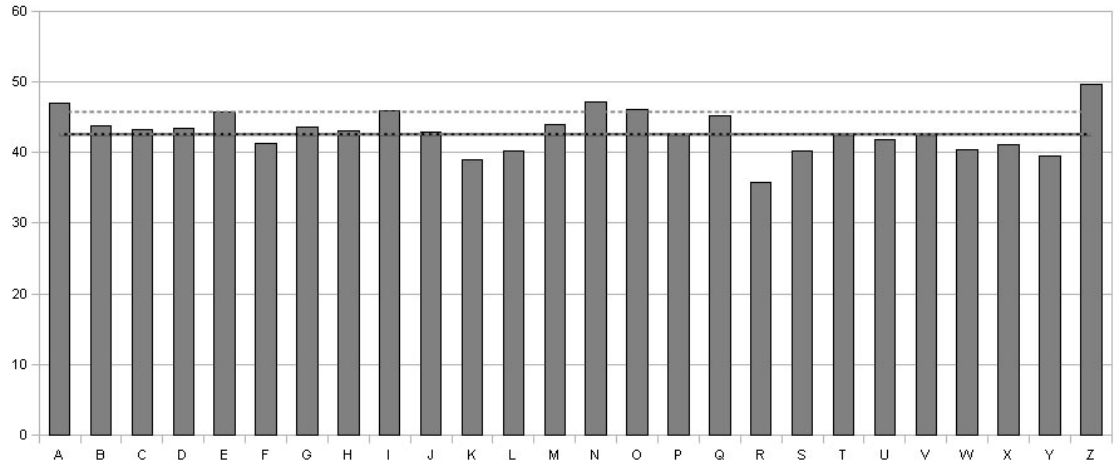

Figure 8: Systematic sampling – entries in NKFD

In the case of entry count in NKFD, sample mean of 44.17 over-represents the whole dictionary content (the true value of 42.58 is not included in the CI (42.60-

45.74)) and it may be difficult to assess whether this over-estimation is more or less serious than the bias resulting from single-stretch sampling (just mind the drop in letter R).

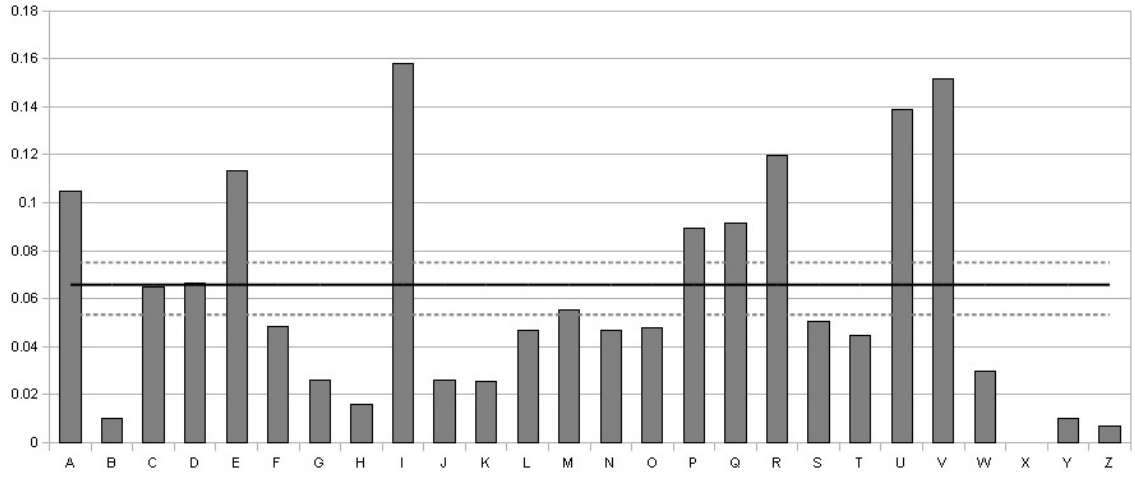

Figure 9: Systematic sampling – "formal" labeling in NKFD

Nonetheless differences begin to be more visible when the characteristic in question has a less uniform distribution. Compare systematic sampling for "formal" labeling in NKFD in Figure 9 with its extremely uneven distribution as well as NKFD's provision of equivalent disambiguators (per entry ratio) in Figure 10.

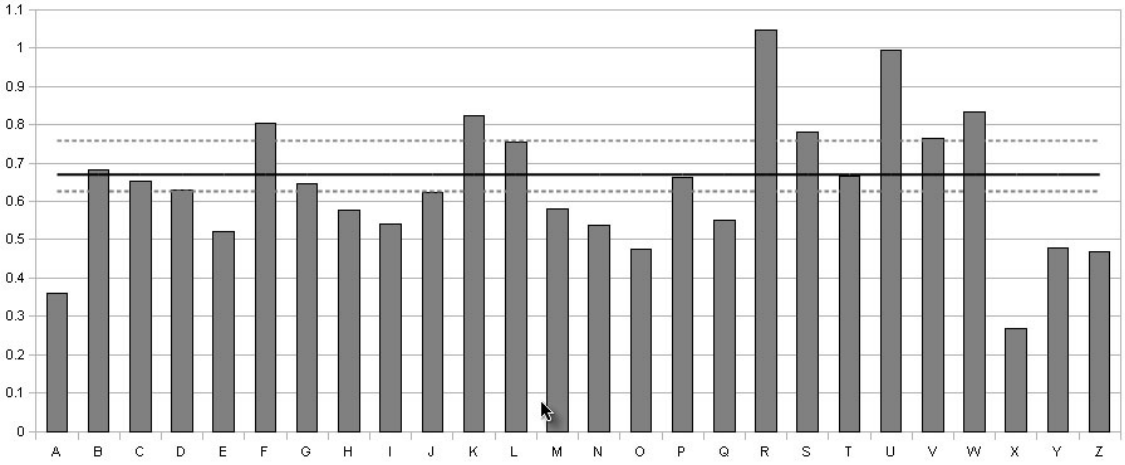

Figure 10: Systematic sampling – equivalent disambiguators in NKFD

Let us now move to PiotrSal. In the case of entry count and mean number of equivalents per entry, systematic sampling also proved more reliable than single-stretch sampling. As one can infer from Figure 11, 13 out of 24 letter categories have a mean that lies outside the systematic sampling CI.

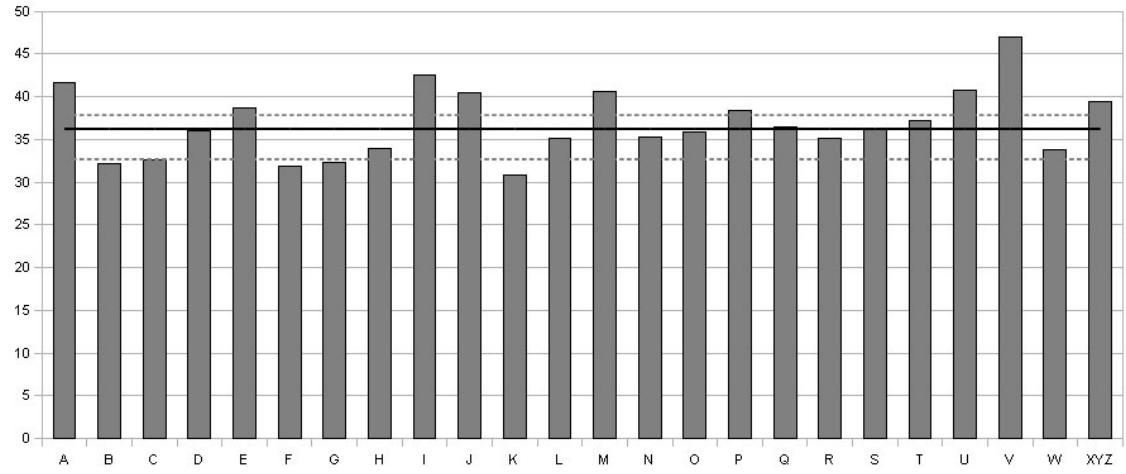

Figure 11: Systematic sampling – entries in PiotrSal

The situation is not very different in the case of mean number of equivalents provided per entry in PiotrSal (Figure12). In this case the distribution is quite uniform when compared with labeling but still systematic sampling is no doubt better than choosing any single letter despite the fact that the sample under-represents the dictionary content (the true mean of 2.3765 equivalents per entry is still below the higher CI endpoint of 2.3855 even though visually they overlap in the figure).

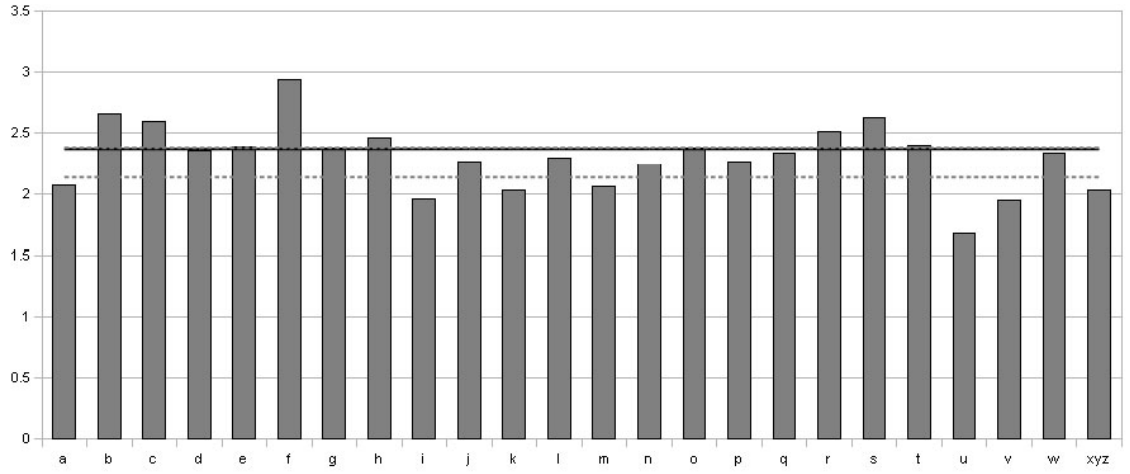

Figure 12: Systematic sampling – equivalents in PiotrSal

Finally, an example clearly calling for more data. Figure 13 illustrates how wide the CI for mean number of "US" labels per entry in PiotrSal is. It is probably caused by both relatively small sample size, and low frequency of labeling (only 31 labels in the sample). Its length (0.0159) 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. This reveals a drawback that systematic sampling has when compared with random sampling techniques, especially with SRS: in such a case the selection of additional pages that are clearly needed may be far more complicated than simply drawing another set of page numbers of desired size (eliminating those that get duplicated). Mind that when an additional systematic cluster is selected, we end up with two clusters which means that we have to apply a different set of formulas for calculating the estimator and its variance. Especially when entry is the ultimate sampling unit, just as it is the case here, we would end up with a two-stage cluster sampling. Thus, the complexity does not result merely from ensuring that the two sequences of page numbers selected do not overlap, but from the change in the statistical model that has to be applied. Besides, this method lacks flexibility with regard to sample size which characterizes random sampling and simple random selection of pages in particular.

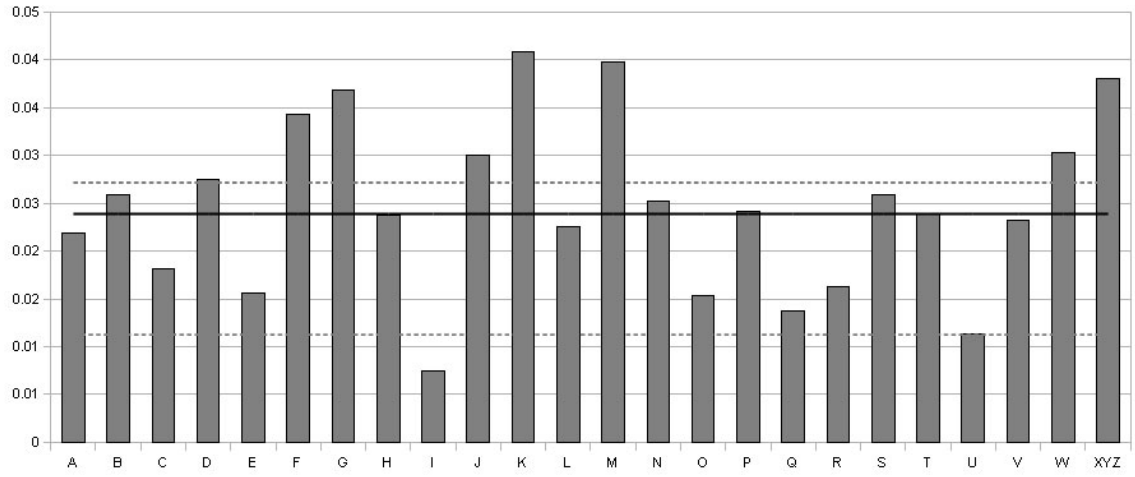

Figure 13: Systematic sampling – "US" labeling in PiotrSal

Summing up, we can see that even though systematic sampling offers only limited possibilities of randomization it proved better than single-stretch sampling in virtually every case presented above. One can argue that in some cases the distance between the within-letter mean and the true value of the parameter in the dictionary was smaller than CI length. Mind, however, that the CI length, although biased in this case, tells us something about the level of heterogeneity of the dictionary matter, thus allowing us to decide whether we need more data or not. Taking just one letter we would never realize that the data are very unevenly distributed. The graphs I presented above are an artificial construct because normally we would know neither the distribution pattern nor the true

mean value. We would just have the sample mean and the CI. In this case it is exactly the CI length that provides an indication of the quality of the sample; and in the case of "US" labeling in PiotrSal, the CI issues a clear warning. Point estimates derived from a single-stretch sample can never do this.

4.3. Sampling in Coleman and Ogilvie (2009) evaluated

In this section I am going to empirically evaluate sampling methods consisting of selecting a number of *initial* entries under each letter as proposed in ((Coleman and Ogilvie 2009): 10f). As already stated in 2.1.4. they use two methods of sampling: selecting first 50 entries under each letter and first 10% of entries under each letter. As one can see neither of the methods is random. My point is to examine whether one can allow for failure to randomize when using stratification. As my strata are letters of the alphabet, a by-product of this would be a first step to checking whether alphabet fatigue applies not only to the whole dictionary text but also to each and every letter separately.

I am going to compare estimates resulting from using these methods with the data from the whole dictionary text and with stratified random sampling, even though it is described in detail only in 5.2. below. Few details and little numerical data concerning stratified random sampling will be presented here; the reader might go back to the present section when I will be discussing stratified sampling in greater detail (in 5.2.). As the methods proposed by Coleman and Ogilvie (2009) are only suitable for dealing with entry-based characteristics, only those will be examined.

As my default sample size for random sampling is 10%, it lends itself for direct comparison with 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 the the total sample size be the same as in the case of random sampling (which is of course 10% of the whole dictionary text). In Webster x proved to be 437 entries for "obsolete" labeling, 436 for quotation provision; in NKFD 294 for both characteristics; in PiotrSal the first 73 entries for mean number of equivalents per entry and 72 entries for "US" la-

beling. Those slight differences result from the fact that, in the case of stratified sampling which served as a basis for comparison, pages and not entries were drawn.

For the "first 50" and "first x" methods I will estimate the overall mean using both arithmetic and weighted mean. I have already raised my concern in 2.1.4. above that allocating the same number of entries to each letter regardless of their original size will lead to an 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 Neyman allocation (cf. Barnett (1974: 94ff) and Deming (1950: 226ff)), which has been shown to be optimal, consists of allocating sample size proportionally to within-stratum variation. It appears that Coleman and Ogilvie method is doing exactly the reverse. Using weighted mean will obviously not eliminate the loss in precision resulting from non-optimal allocation but in this case I will not calculate confidence intervals and therefore I am not interested in precision that much. Weighting will, however, eliminate the bias resulting from uneven representation of different strata. What remains is the bias towards the beginning of each letter which is obviously unknown in general. Therefore, if weighted mean estimate does not improve on arithmetic mean and both estimates differ remarkably from the true value of the parameter, it would mean that there is a considerable bias towards the beginning of the letter. Information on bias resulting from choosing initial entries under each letter will also be provided by the "first 10%" method as in this case the allocation to strata is the same as in my random techniques, only the method of selection within each stratum differs.

As in the section 4.2. , I will start discussing the results with data from Webster. All the figures relating to this sampling method follow the same scheme. The bars from left to right represent the "first 50" arithmetic mean (simply called "mean" in the legend), "first 50" weighted, "first x" arithmetic mean, "first x" weighted mean and finally first 10% mean. For each figure the legend provides the value of x. There is also a black continuous line representing the true mean and black dashed lines representing confidence interval for stratified sampling (SS in the legend).

Figure 14 presents how Coleman and Ogilvie 2009 sampling worked with "obsolete" labeling in Webster. As one can see in this particular case, stratification alone managed to provide remarkably better estimates than single-stretch sampling. Estimates

provided by the "first x" method (regardless of the estimator formula) proved to be quite accurate.

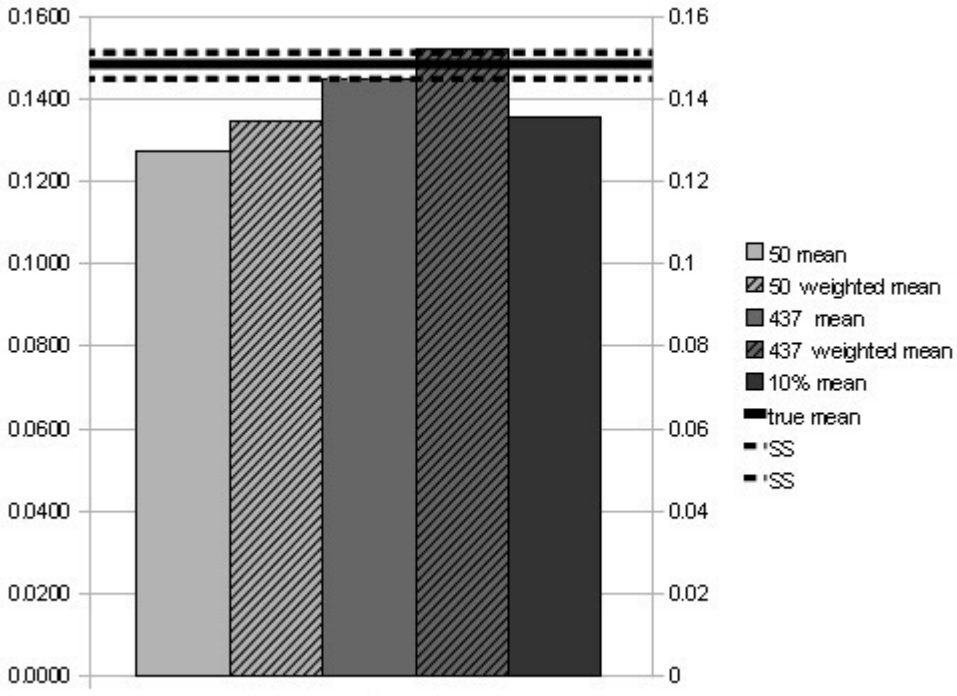

Figure 14: Coleman and Ogilvie 2009 sampling – "obsolete" labeling in Webster

Figure 15 shows that in the case of quotation provision in Webster, bias towards the beginning of the letter results in under-estimation of the mean number of quotations per entry. While in this graph it may not seem that serious, a quick glance at Figure 7 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 letter P for single-stretch sampling, i.e. one of the most serious under-estimates resulting from inaccurate choice of single-stretch sample. Let us remind ourselves that the within-letter mean value of the number of quotations per entry in letter P is 0.2371 whereas the "first 50" estimate (weighted) yields 0.2352. Increase in sample size does help but still we are dealing with considerable under-estimation, this time erring in the region of letter K. All those estimates fall outside the confidence interval for any random technique and for systematic sampling.

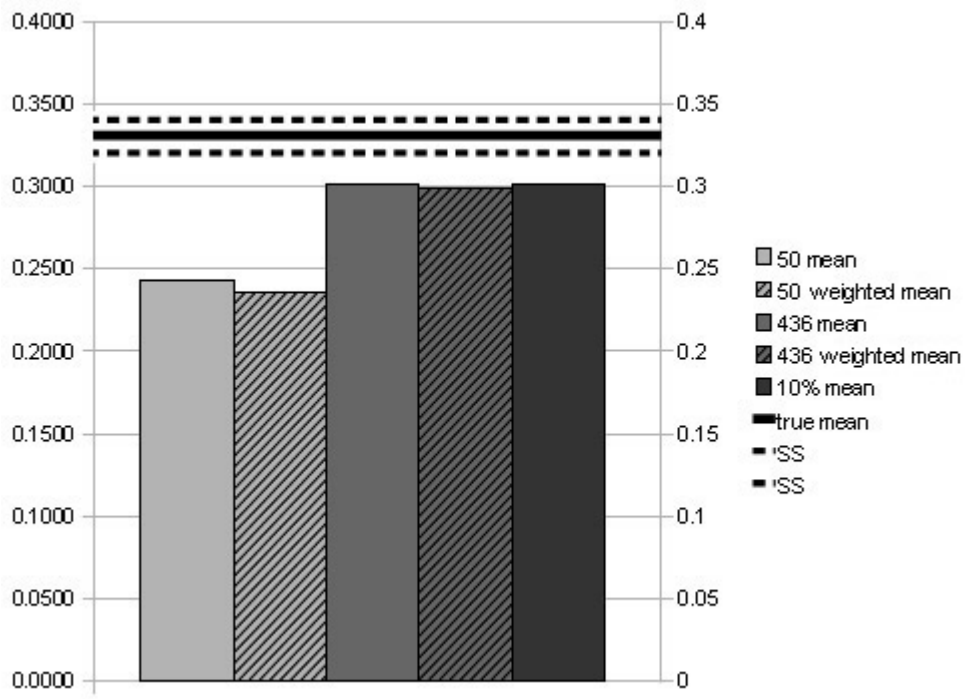

Figure 15: Coleman and Ogilvie 2009 sampling – quotation provision in Webster

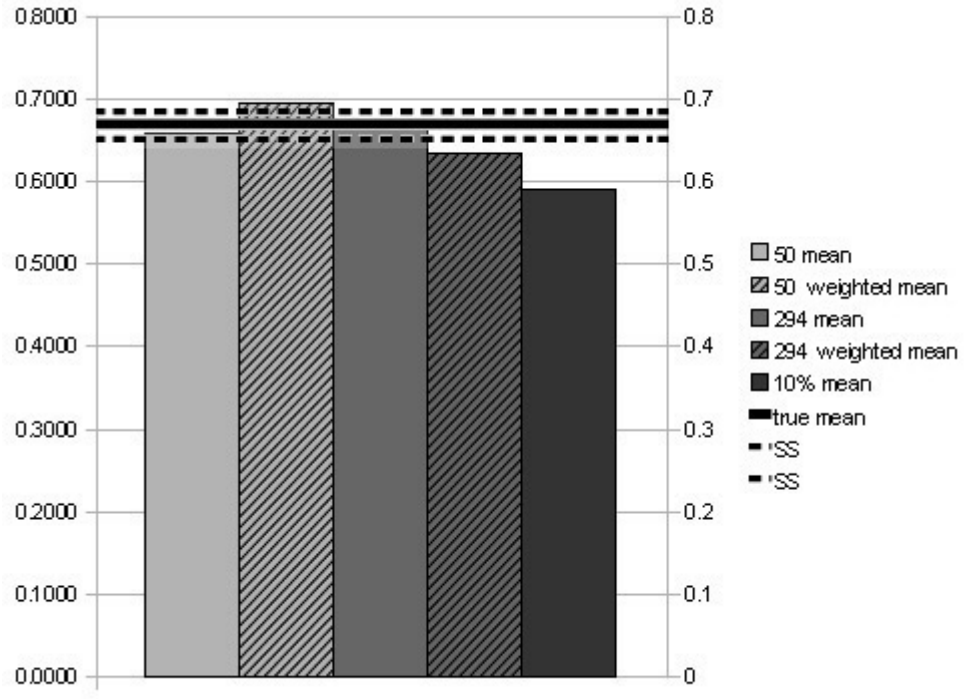

Figure 16: Coleman and Ogilvie 2009 sampling – equivalent disambiguators in NKFD

This kind of sampling does not always yield acceptable results. Figure 17 shows that 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 within-letter mean in M. Figure 17 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 single-stretch samples would yield graver errors in estimation.

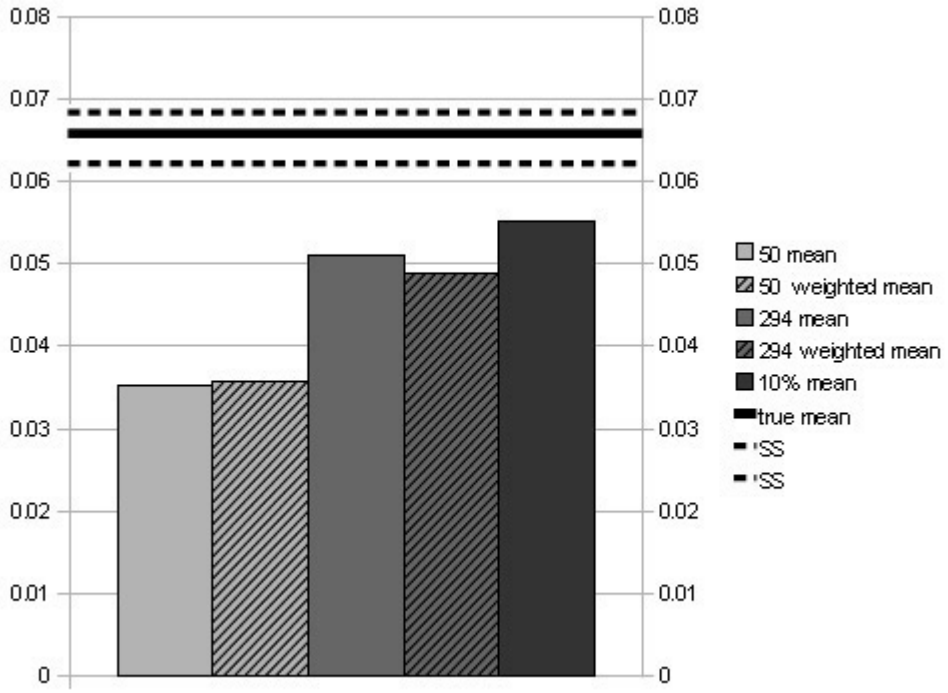

Figure 17: Coleman and Ogilvie 2009 sampling – "formal" labeling in NKFD

Finally, PiotrSal. As shown in Figure 18, 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 of the number of equivalents per entry.

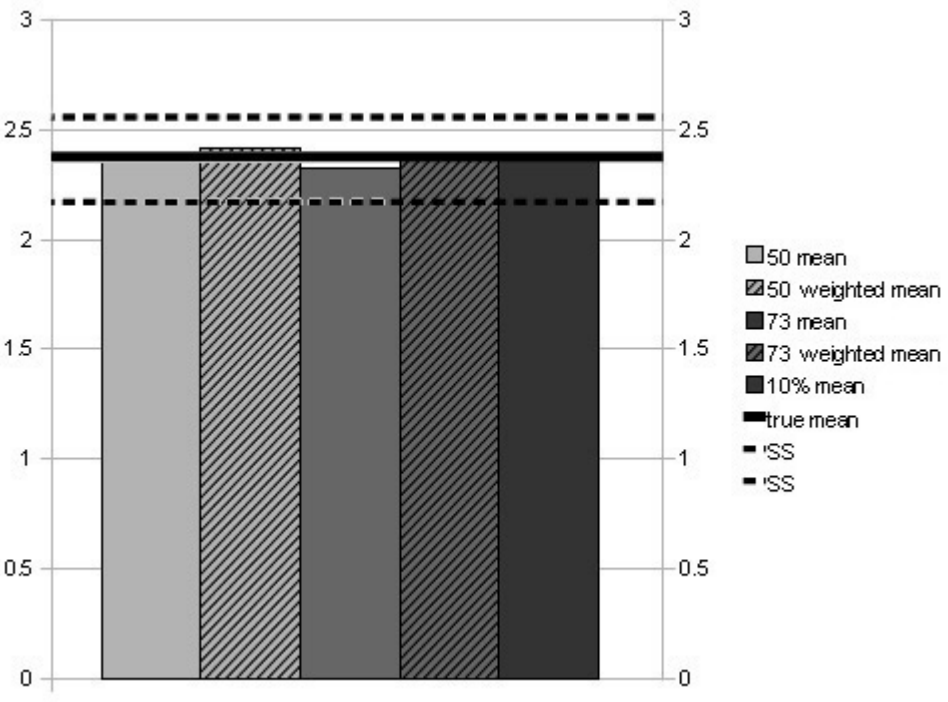

Figure 18: Coleman and Ogilvie 2009 sampling – number of equivalents per entry in PiotrSal

"US" labeling estimation in PiotrSal (Figure 19) presents a very interesting instance of sample size increase having a detrimental effect on estimation. What is surprising in this case is that each successive method that potentially could have been better than the previous 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 (for details see 5.2.3.)

I would dare to draw only one conclusion based on the data presented above: Coleman and 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.

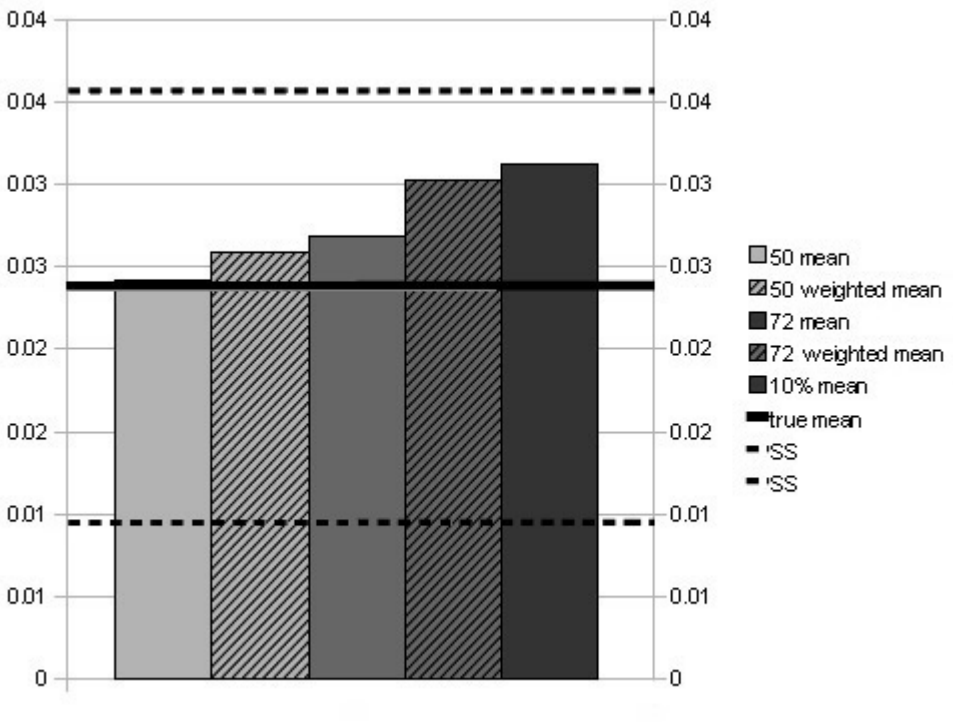

Figure 19: Coleman and Ogilvie 2009 sampling – "US" labeling in PiotrSal

Chapter 5: Random sampling techniques proposed

The description and evaluation of random sampling techniques will be centered around the method in which pages are selected. First, pages will be selected in a simple random way, then stratification will be performed. The general question to be answered here is whether stratification can bring about an increase in precision. Mind, however, the distinction between simple random selection of pages and simple random sampling. Simple random selection of pages does not necessarily mean that we are dealing with SRS as the page need not to be the ultimate SU. When entries are ultimate SUs, we might have simple random selection of pages, but technically speaking we are dealing with cluster sampling.

First, simple random sampling will be analyzed based on the estimation of the mean number of entries per page On the basis of this technique I will also show how to find an appropriate sample size once precision of estimates is predetermined. Further, sampling schemes with simple random selection of pages will serve as a standard of comparison for stratified sampling. It will be analyzed under what theoretical conditions SS can be more efficient than SRS. Next, I will check whether those conditions are satisfied in our test dictionaries.

I personally believe cluster sampling with pages as secondary sampling units is of little use in metalexicographic research and therefore I will not discuss this technique in detail. It sometimes happens that a given researcher samples two clusters, usually two letters, but those are not random samples and therefore will not be analyzed here. Generally, CS is mainly used for convenience, as it rarely yields better estimates than SRS or SS.

5.1. Simple random selection of pages

When one does not need to compare selected stretches of the dictionary text, differences in treatment under subsequent letters, differences in works of two or more editors; simply when there is no valid rationale for stratification, the best possible solution would then probably be to select pages in a simple random manner. The main advantage of this scheme is its simplicity, both in terms of the mathematics behind it and in terms of application. In most cases it should be efficient enough, even though under certain conditions other techniques may yield more precise estimates.

5.1.1. Page as sampling unit – simple random sampling

Let us assume that a dictionary page can be treated as SU. It is a very convenient situation as a variety of relatively simple sampling methods are at the researcher's disposal. It is always the case when one wants to estimate the mean number of entries per page and this very characteristic will be used here to illustrate SRS. But there are more possibilities than that. The general condition is that each entry either possesses the characteristic of interest (1) or not (0), e.g. it either has been copied from the previous version of the dictionary or it is newly entered on the word list. In such a case the researcher might not be so much interested in the mean value as in the ratio of newly entered entries. Nonetheless the sampling scheme is still an SRS. Appropriate formulas for ratio estimation can be found e.g. in Barnett (1974: 38ff).

Suppose our dictionary consists of N pages and one wants to draw n of them. In such a case there is a total of $\binom{N}{n}$ [4]samples from which one is drawn.

When estimating the mean $\bar{X}$, an estimator with intuitive appeal is the sample mean:

(8) $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

[4] i.e. combinations without repetitions. The number of such combinations equals N!/n!(N – n)!

The estimator is unbiased and its variance is to be estimated with the following formula:

(9) $$Var(\bar{x}) = \frac{(1-f)\sigma^2}{n}$$

where *f* is the sampling fraction *n/N*. Typically, however, the values of σ^2 will not be known and must be estimated from the sample. Sample variance s^2 will be its good, unbiased estimate. Now we might wish to construct a CI for $\bar{X}$. When the sample size is large enough (say larger than 40), an appropriate 100%(1− α) symmetric two-sided CI can be written as:

$$\bar{x} - z_\alpha \sigma \sqrt{(\frac{(1-f)}{n})} < \bar{X} < \bar{x} + z_\alpha \sigma \sqrt{(\frac{(1-f)}{n})}$$

(10)

or, more generally as:

(11) $$\bar{x} - z_\alpha \sqrt{(var(\bar{x}))} < \bar{X} < \bar{x} + z_\alpha \sqrt{(var(\bar{x}))}$$

where z_α stands for the two-tailed α - point of N(0,1) (normal standardized distribution). In practice, σ in (10) will not be known, but with larger samples replacing σ with *s* is reasonable. If the sample is small, however, it would be safer to use Student's t-distribution with n – 1 degrees of freedom rather than the normal distribution. The CI is then as follows:

(12)
$$\bar{x} - t_{(n-1)}(\alpha) s \sqrt{(\frac{(1-f)}{n})} < \bar{X} < \bar{x} + t_{(n-1)}(\alpha) s \sqrt{(\frac{(1-f)}{n})}$$

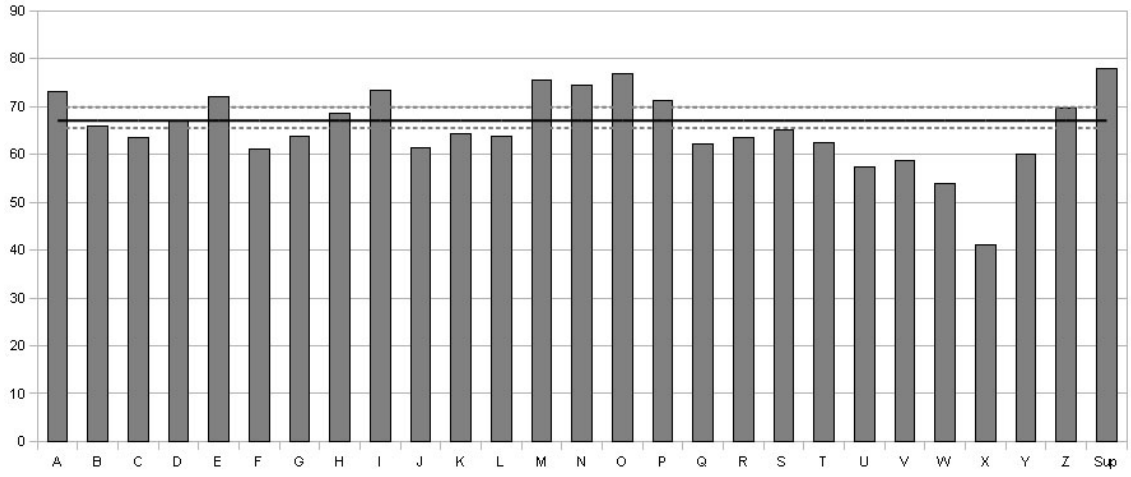

Figure 20: SRS – entries in Webster

Now, let us see how SRS worked with our dictionaries when estimating mean number of entries per page. In general, similarly as in the case of systematic sampling, SRS proved definitely better than single-stretch judgmental sampling. Let us now move to particular dictionaries to see the details.

Even though we cannot speak of glaring over- and under-treatment in Webster, very few (only four out of twenty seven) within-letter means lie within the SRS CI. SRS offers quite precise estimation with CI length of 4.3369, which is a little bit more precise than in the case of systematic sampling (CI length 4.5329). Most importantly however, we are using *strictly unbiased* estimates in this case.

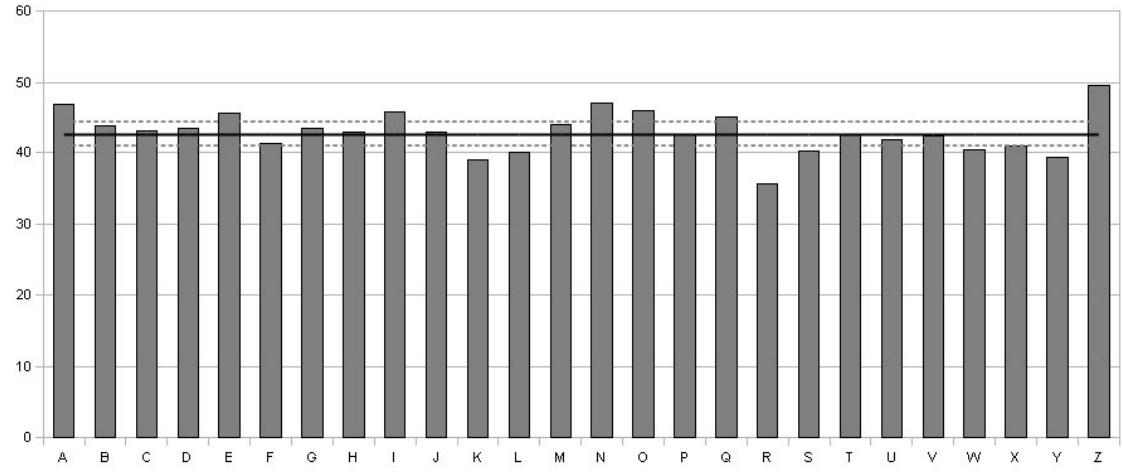

Figure 21: SRS – entries in NKFD

As can be seen in Figure 21, SRS does not hold such an advantage over single-stretch sampling in NKFD because the data are almost uniformly distributed there. The

CI length is 3.3306 which in this case did not translate into an improvement in precision when compared with systematic sampling (3.1439)

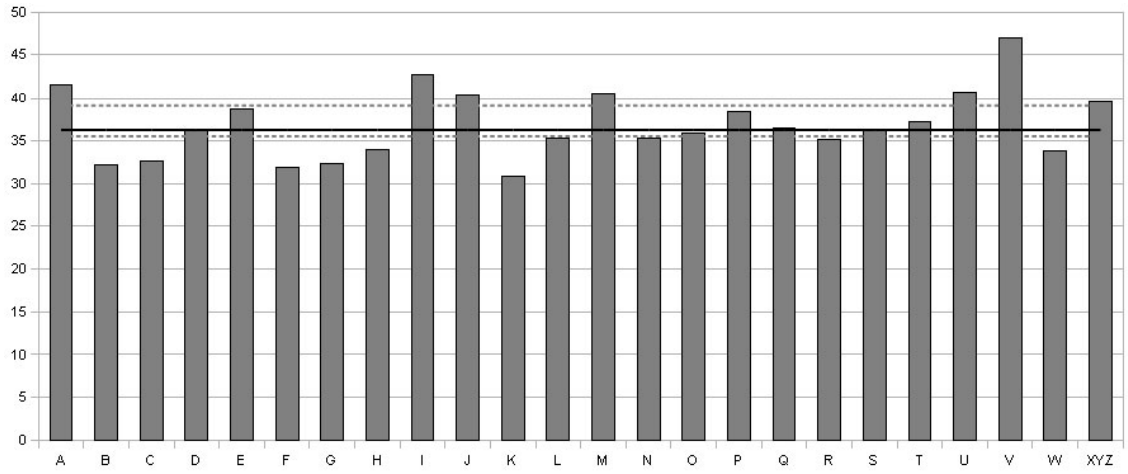

Figure 22: SRS – entries in PiotrSal

Results from PiotrSal also come as no surprise. In this case a CI of 3.59 proved to be considerably narrower than in systematic sampling (5.16). Mind that PiotrSal is not even one fourth the size of NKFD and only about 13.5% the size of Webster. Therefore the amount of data analyzed in the sample is much smaller but the estimate shows similar precision as in the case of these much bigger dictionaries.

From the data presented above we cannot infer that SRS performed considerably better than systematic sampling. Nonetheless, this method has the advantage of offering strictly unbiased estimates and the possibility of easy selection of additional pages if more data should be needed. Therefore, it should still be preferred over systematic sampling.

5.1.2. Matching sample size for predetermined precision.

In the previous section as well as in the sections to follow, sample size was predetermined and estimators examined in terms of their precision. This was done, among others, for illustrative purposes but very often it is the precision that is predetermined and sample size needs to be found.

As precision is related to estimator variance, it should be intuitively clear that a provisional variance estimate is needed in the search of an appropriate sample. The

most reliable approach would be to run a preliminary sample and then to run a complementary one.

Using the method described in Barnett (1974: 32ff), a minimum sample size will be calculated. Suppose a 100%(1 – α) CI for means narrower than *d* is desired. This is equivalent to saying that

(13) $$P(|\bar{X}-\bar{x}|>d)\leq\alpha$$

having standardized it we obtain:

(14) $$P\left(\frac{|\bar{X}-\bar{x}|}{S\sqrt{\frac{(1-f)}{n}}}>\frac{d}{S\sqrt{\frac{(1-f)}{n}}}\right)\leq\alpha$$

assuming that SRS sample mean has a normal distribution we require that:

(15) $$\frac{d}{S\sqrt{\frac{(1-f)}{n}}}\geq z_{\alpha}$$

which is equivalent to:

(16) $$n\geq\left(\frac{S z_{\alpha}}{d}\right)^{2}\left[1+\frac{S z_{\alpha}}{Nd}\right]^{-1}$$

This presupposes that S is known, which is unlikely to be the case. That is why a preliminary sample is needed to estimate it.

Let us now suppose that the 4.3359 length of the CI for the mean number of entries per page in Webster does not satisfy the researcher's needs. For some reason one wants the CI to be no wider than 2 entries. The sample variance that has already been calculated previously is 238.01. N equals 1749 as this is the number of pages in Webster. Therefore the sample size that would probably be needed to meet these requirements would be $n\geq\left(\frac{1.96\sqrt{(238.01)}}{2}\right)^{2}\left[1+\frac{1.96\sqrt{(238.01)}}{(1749*2)}\right]^{-1}$ After calculating and

rounding we learn that we would need 227 pages altogether, which means that an additional 54 pages will have to be drawn to meet the desired precision.

5.1.3. Single dictionary entry as sampling unit – cluster sampling

When sampling a paper dictionary drawing pages is often the only possible way out, but the researcher might be interested not in pages themselves but in individual entries. Naturally, a page constitutes a cluster of entries and that is why we should discuss cluster sampling in more detail now. In this case sampling takes place in one stage – once the pages are drawn the sample is determined. Obviously, each page includes a different number of entries, so it is CS with different-sized clusters. Let us assume there are M clusters (i.e. pages) of sizes $N_{1,} N_2 \cdots N_M (\sum_{i=1}^{M} N_i = N)$ in the dictionary. Cluster means $\bar{X}_i$ and variances σ_i^2 are defined in the usual way. The overall dictionary mean can then be written as follows (after Barnett (1974:123)):

$$\bar{X} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i} X_{ij}}{\sum_{i=1}^{M} N_i} = \frac{\sum_{i=1}^{M} X_{iT}}{N} \tag{17}$$

The problem here is that we may not know the dictionary's total number of entries, which is needed to construct strictly unbiased estimators of $\bar{X}$ (see Barnett (1974: 124f) if you are interested in constructing such estimators). Therefore we will have to satisfy ourselves with a slightly biased estimator of the mean as presented in Barnett (1974: 124)

$$\bar{x_{cl}} = \frac{\sum_{i=1}^{m} x_{iT}}{\sum_{i=1}^{m} n_i} \tag{18}$$

Fortunately, the bias of this estimator is weak if the number of pages included in the sample is big as it is of range m^{-1} (see Barnett (1974: 53) for a proof). Its variance can be approximated with the following formula given in Barnett:

$$(19) \qquad Var(\bar{x}_{cl}) \approx \frac{(M-m)M}{(M-1)m} \sum_{i=1}^{M} \left(\frac{N_i}{N}\right)^2 (\bar{X}_i - \bar{X})^2$$

Note that this variance is dependent on variation between cluster means – and unlike in stratified sampling[5] it is more efficient than SRS if the between-cluster variation is relatively small. Obviously we are unlikely to know $N, \bar{X}_i, \bar{X}$. The latter two will have to be replaced by their sample equivalents and N may be replaced by sample estimate $\frac{Mn}{m}$ to yield:

$$(20) \qquad var(\bar{x}_{cl}) \approx \frac{(M-m)m}{M(m-1)} \sum_{i=1}^{m} \left(\frac{n_i}{n}\right)^2 (\bar{x}_i - \bar{x}_{cl})^2$$

As CS concerns entry-based characteristics, it would be perhaps more interesting to learn how it worked with our data. Here again, let us start with the two characteristics analyzed in Webster.

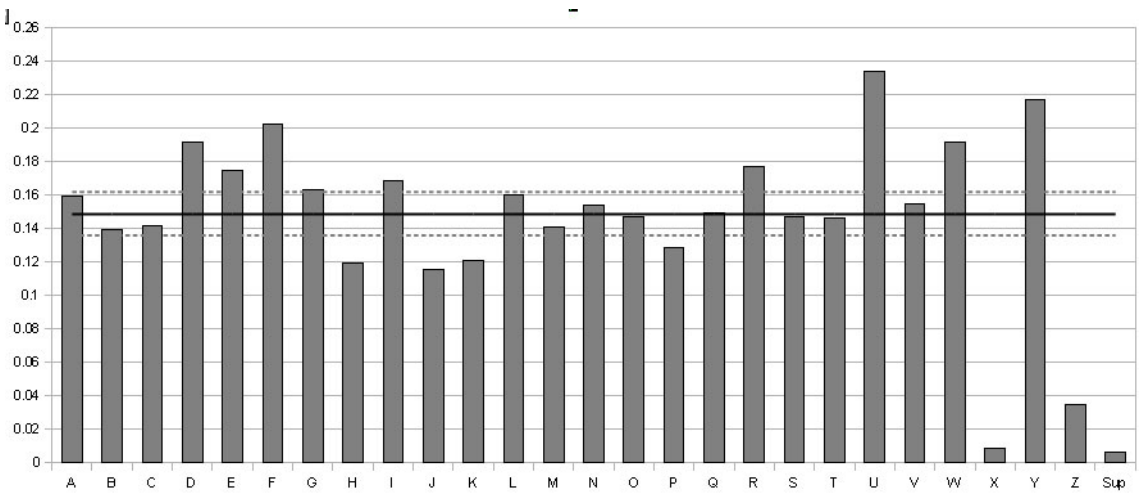

Figure 23: CS – "obsolete" labeling in Webster

In this case sample estimate of the mean proved to be nearly equal to the true mean value. The results are similar to those obtained from systematic sampling both in

[5]See section 5.2.2. for a comparison between SRS and SS

terms of accuracy and precision and, as already stated in 4.2. , definitely more reliable than single-stretch sampling.

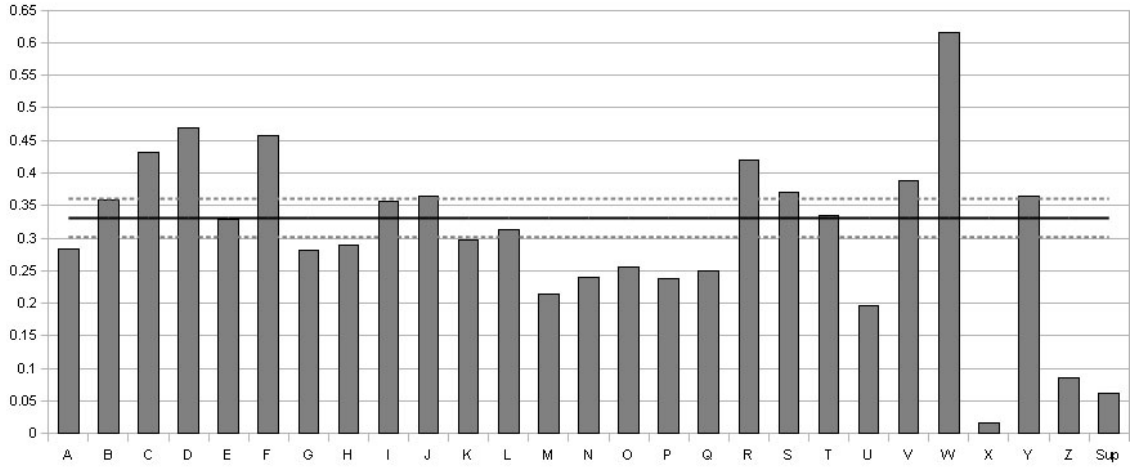

Figure 24: CS – quotation provision in Webster

Figures 24 and 25 represent the analysis of quotation provision in Webster and provision of equivalent disambiguators in NKFD. As one can see in both cases simple random selection of pages yielded acceptable results.

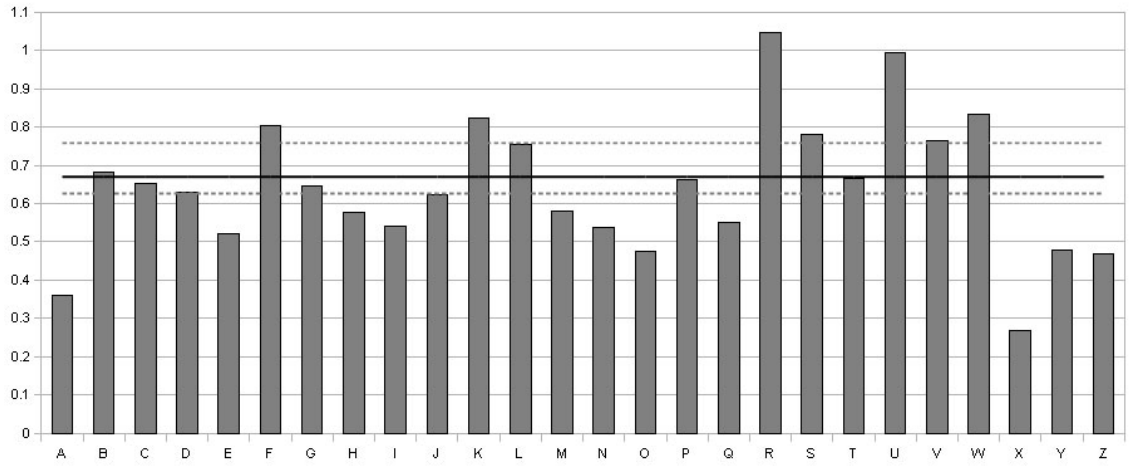

Figure 25: CS – equivalent disambiguators in NKFD

Figure 26 illustrates how CS worked with "formal" labeling in NKFD. Similarly as in the case of systematic sampling, the CI is relatively wide (of almost equal length in both cases) which is only to be expected with so unevenly distributed data. We will further see whether stratification can improve on precision.

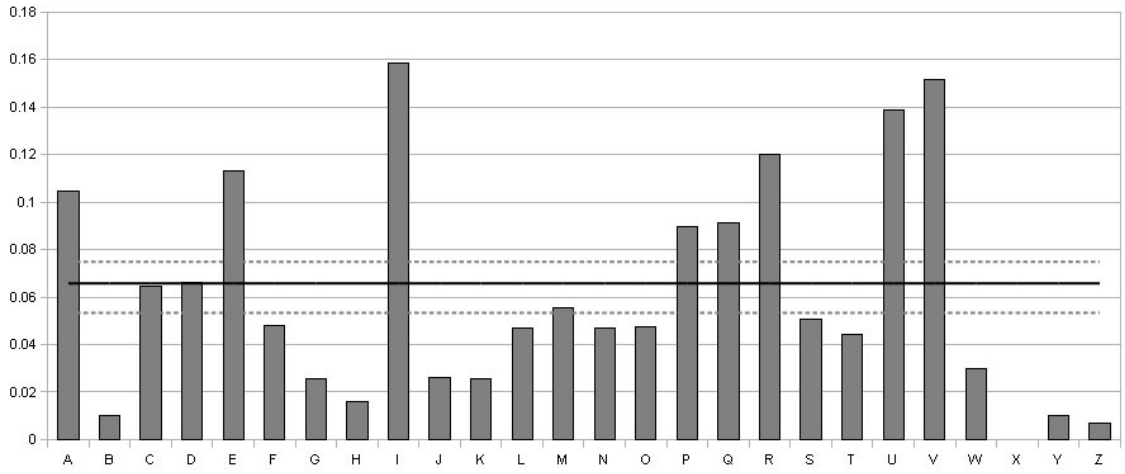

Figure 26: CS – "formal" labeling in NKFD

Estimating mean number of equivalents per entry in PiotrSal (Figure 27) was by no means surprising. When it comes to "US" labeling (Figure 28) in this dictionary, similarly as in the case of systematic sampling the CI proved wide when compared with the range of within-letter means. It shows a bit of an improvement when compared with systematic sampling. The CI of 0.0135 is 85% the length of the CI in systematic sampling, but the precision reached is still not satisfactory. Definitely more data is needed in this case.

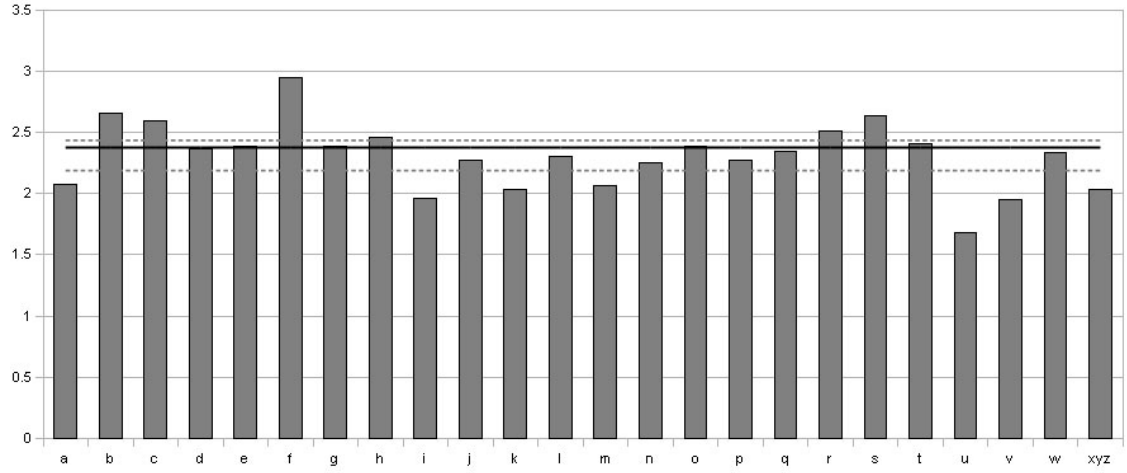

Figure 27: CS – mean number of equivalents per entry in PiotrSal

Summing up, I would like to stress what I already pointed to in 4.2. In all of the cases analyzed here both systematic sampling and sampling with simple random selection of pages, be it SRS or CS, proved definitely more reliable than single-stretch sampling. Both methods allow for control over reliability of results. In the case of "US"

labeling in PiotrSal we clearly see that we failed to reach a satisfactory level of precision. This knowledge presents yet another advantage of random (or at least partially random) methods over judgmental sampling.

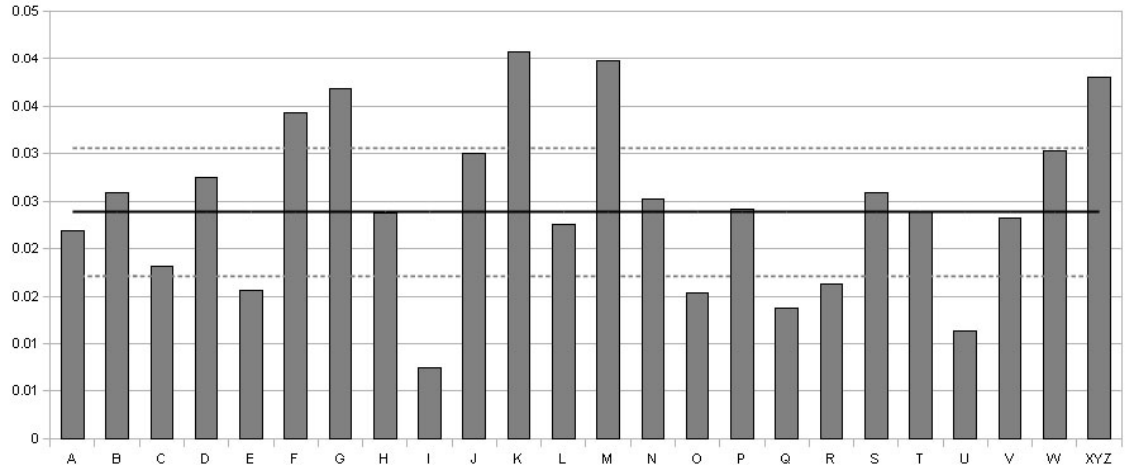

Figure 28: CS – "US" labeling in PiotrSal

In the majority of cases simple random selection of pages did not lead to improvements in terms of precision when compared with systematic sampling. There are however two facts, already mentioned before, that in my opinion still give priority to simple random selection of pages: lack of estimator bias and simplicity of selecting additional pages when needed. Whereas the former is mostly formal, the latter has a practical dimension.

5.2. Stratified selection of pages

Originally, the idea behind SS was to reduce sample variance and therefore provide more precise estimates for the same sample size. As it will be shown in the section to follow (5.2.2), under certain conditions stratification may (but need not – see Deming (1950: 241) for some common misconceptions regarding stratification) lead to an increase in precision. Stratification does however more than just this. Very often one might be interested not only in the dictionary as a total but also in its separate parts, be it letters, be it stretches edited by different editors or any other divisions the researcher might think of as useful for the study. The important assumption, however, is that this is a division into non-overlapping groups that must cover the entire dictionary (or at least

the part of the dictionary under investigation e.g. when interested in the treatment of adverbs it suffices that it covers all the adverbs). Apart from that, we should know a priori what the strata are, what is the number of elements in each one and what is the sampling fraction in each stratum. In general f_i does not need to be identical for each stratum, but for the sake of simplicity I sometimes assume it is. This is called stratified sampling with proportional allocation in the literature[6].

In this work, however, proportional allocation (or rather the estimator formulas intended for stratified sampling with proportional allocation, because I will try to allocate proportionally) will not be used as preliminary research showed that the bias resulting from the false assumption of proportionality was serious, especially in PiotrSal because of its small size. For instance, the letter F consists of 24 pages, letter L of 15. When taking a 10% sample, rounding resulted in sampling exactly 2 pages in both cases. In large dictionaries where letters consist of hundreds of pages, this bias is no doubt negligible but as using precise weight does not add much to the complexity of the calculations, I will also use exact weighting when stratifying Webster and NKFD.

5.2.1. Page as sampling unit – stratified sampling

Suppose that our dictionary is divided into *k* strata and we follow the notation in Appendix 1. The overall mean takes the following form:

$$(21) \qquad \bar{X} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{X}_i = \sum_{i=1}^{k} W_i \bar{X}_i$$

where $W_i = (\frac{N_i}{N})$ is termed the weight of the i-th stratum. When sampling a paper dictionary with letters serving as strata this would be the number of pages under a given letter over the total number of pages in the dictionary. The overall variance (after Barnett (1974: 78)):

[6]In fact proportional allocation is not optimal. Allocation that minimizes the variance of stratified mean estimator, thus giving the highest possible precision for a given sample size (assuming the unit sampling are constant in each stratum) is called Neyman allocation after Jerzy Neyman, who gave the proof of its optimality. However, using Neyman allocation requires pre-knowledge or pre-estimate of each stratum variance and therefore is rather tedious to implement. Moreover, as shown in Barnett (1974: 94ff) and Deming (1950: 226ff) the gains from using it may be quite modest. The present author's intuition is that it is of little use in dictionary sampling.

(22)

$$\sigma^2 = \frac{1}{(N-1)} \sum_{i=1}^{k} \sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i + \bar{X}_i - \bar{X})^2 = \frac{1}{(N-1)} \sum_{i=1}^{k} (N_i - 1)\sigma_i^2 + \sum_{I=1}^{k} N_i (\bar{X}_i - \bar{X})^2$$

Now we take a SRS from each stratum. Obviously one needs estimators of the above values. The so-called stratified sample mean is a commonly employed unbiased estimator of $\bar{X}$ and it is defined as:

(23) $$\bar{x_{st}} = \sum_{i=1}^{k} (\frac{N_i}{N}) \bar{x}_i = \sum_{i=1}^{k} W_i \bar{x}_i$$

Mind that in general this is not identical to the arithmetic mean unless the allocation is proportional. If this is not the case, the arithmetic mean is a biased estimate (as already stated above).

Its variance and variance estimate are given as:

(24) $$var(\bar{x_{st}}) = \sum_{i=1}^{k} \frac{W_i^2 (1 - f_i)\sigma_i^2}{n_i}$$

and

(25) $$s^2(\bar{x}_{st}) \sum_{i=1}^{k} \frac{W_i^2 (1 - f_i) s_i^2}{n_i} = \frac{1}{N^2} \sum_{i=1}^{k} \frac{N_i (N_i - n_i) s_i^2}{n_i}$$

respectively.

As we can see, similarly as in SRS, σ_i^2 will not be known so it needs to be replaced by its unbiased estimate s_i^2 . CIs are analogues to SRS for N(0,1) distribution.

It may sometimes happen that selecting a SRS from specific strata is not possible. The case of stratification by letters of alphabet is straightforward, but a researcher might want to use strata that are not readily available, e.g. stratifying according to part of speech in the case of a paper dictionary. The researcher may not be able to determine to which stratum an element belongs until it has been drawn. In such a case, there is no other method as to select a SRS from the whole dictionary and subsequently assign the drawn elements to different strata. Such an approach is called post-hoc stratification (cf. Barnett (1974: 100)). Note that the number of elements in each stratum is in itself a random variable. If the sample size is big, post-hoc stratification should approach SS with proportional allocation. Nonetheless this assumption should be applied with caution and

treating the sample as SRS when calculating the estimates is always a safer way out if post-stratification is done.

Another word of caution is needed: stratification should not be used as an excuse for failure in obtaining randomness. Therefore if a researcher takes a sample consisting of, say, 10% initial entries under each letter as e.g. in Martínez Egido (2002) or similarly in Rodríguez-Álvarez and Rodríguez-Gil (2006) but the technique is not uncommon and believed to yield representative samples) or the first 50 entries under each letter e.g. Coleman and Ogilvie (2009), it is by no means a stratified random sample. As a consequence, the guidelines presented here do not apply. This is not only a theoretically faulty approach but it may have serious practical consequences, as alphabet fatigue may not only apply to the whole dictionary text but also to each and every letter separately. We have already seen in 4.3. that failure to randomize, even if using stratification, may result in considerable and unpredictable bias.

5.2.2. Stratified sampling and simple random sampling compared

Let us first start with some theoretical background. As mentioned in 5.1, there are certain conditions under which stratification may bring about an increase in precision. To examine this possibility let me, similarly as in Barnett (1974: 83ff), compare $\bar{x}$ and $\bar{x_{st}}$ in the same situation. For simplicity's sake I will assume that we are dealing with proportional allocation. In fact in bigger dictionaries what we have is very close to strict proportionality. Given the constant sample size *n,* an increase in precision would mean that the overall sample variance would be smaller in the case of SS than in SRS. In other words, the difference $Var(\bar{x}) - Var(\bar{x_{st}})$ has to be positive.

(26) $$Var(\bar{x}) - Var(\bar{x_{st}}) = \frac{(1-f)}{n}(\sigma^2 - (\frac{1}{N})\sum_{i=1}^{k} N_i \sigma_i^2)$$

Now we shall assume that the stratum sizes N_i are large enough for the following conditions to be roughly satisfied

(27) $$\frac{(N_i - 1)}{(N-1)} \approx \frac{N_i}{N} \approx \frac{N_i}{(N-1)}$$

Therefore we can replace (22) with the following formula

$$(28) \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{k} n_i \sigma_i^2 = \sum_{i=1}^{k} N_i (\bar{X}_i - \bar{X})^2$$

Then

$$(29) \qquad Var(\bar{x}) - Var(\bar{x}_{st}) = \frac{(1-f)}{Nn}\sum_{i=1}^{k} N_i (\bar{X}_i - \bar{X})^2 \geq 0$$

Therefore it seems that SS will always be at least as efficient as SRS. However, when the SUs are dictionary pages and the dictionary matter is stratified by letters, assumption (27) will not be tenable. In such a case the difference

$$Var(\bar{x}) - Var(\bar{x}_{st}) = \frac{(1-f)}{(n(N-1))}\left(\sum_{i=1}^{k} N_i (\bar{X}_i - \bar{X})^2 - \frac{1}{N}\sum_{i=1}^{k} (N - N_i)\sigma_i^2\right)$$

can be negative as well.

After further calculations (for details please consult Barnett (1974: 84) or Deming (1950: 230ff)), it turns out that "[t]he stratified sample mean will be more efficient than SRS sample mean if variation between the stratum means is sufficiently large compared with within-strata variation (Barnett (1974: 84)". Obviously the same results are applicable for estimating the total.

Therefore, when using the dictionary page as SU, one should think twice before stratifying. SRS is not only simpler, but may be more efficient. In the sections to follow we will see whether in the case of dictionary sampling stratification can bring about an increase in precision.

The situation changes a bit when we are able to adopt a single dictionary entry as SU and draw an SRS or an SS. Then (27) would probably be tenable and SS will certainly not be less efficient than SRS. This, however, may not be technically possible even with electronic dictionaries.

Now let me proceed to examining whether in the case of our three dictionaries stratification helped in achieving greater precision. As usual, I will start with data from Webster. Figure 29 presents an estimation of the mean number of entries per page based on stratified sampling. The confidence interval contains the true mean. It is 5.04 entries

wide, 0.71 entries wider than in the case of simple random sampling which represents a 16% loss in efficiency. This may be just a random effect but we also see that this characteristic has a relatively uniform distribution which probably accounts for the lack of increase in efficiency.

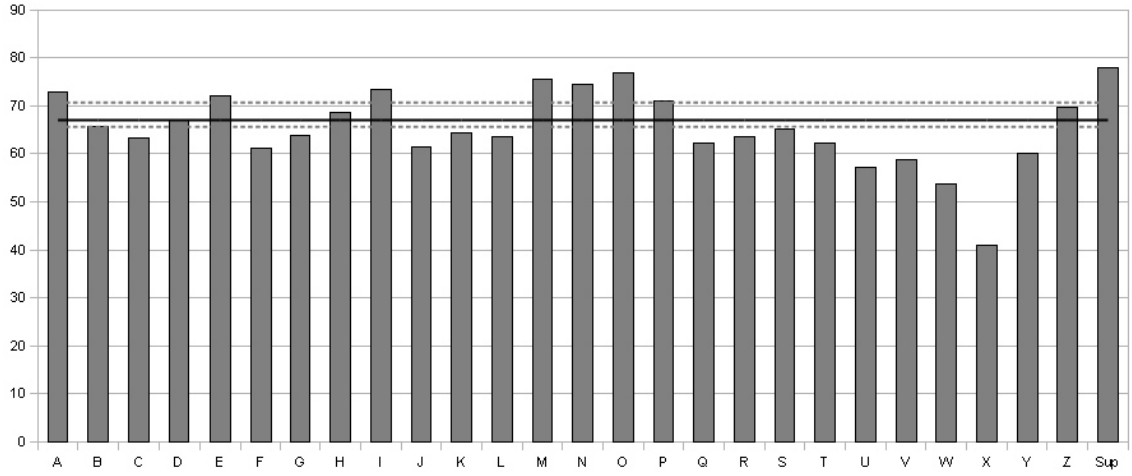

Figure 29: Stratified sampling – mean number of entries per page in Webster

In Figure 30 we can see the confidence interval for mean number of entries per page in NKFD. In this case, just as in Webster, we observe a loss in efficiency. This time the CI is 0.04 entries wider which translates into a 1.16% loss, which I personally consider insignificant.

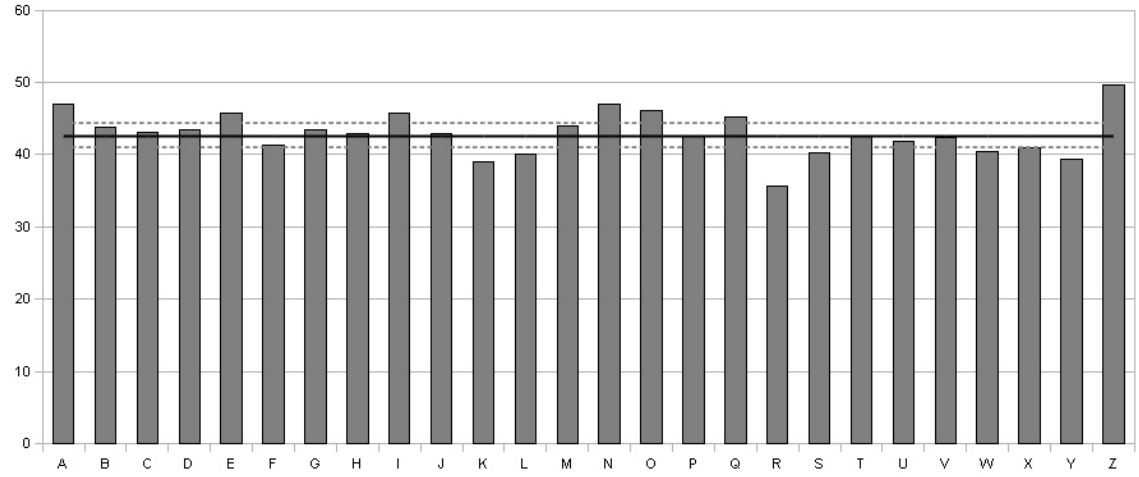

Figure 30: Stratified sampling – entries in NKFD

Finally, let us take up PiotrSal. In this case I exceptionally present additional data in Figure 31. Besides the usual continuous black line representing the true dictionary mean and the dashed gray lines representing the confidence interval I also intro-

duced the fine dashed black line which illustrates the confidence interval around mean estimate which has not been calculated using formula (23) but using the arithmetic mean. By doing so I wanted to illustrate how important assumptions underlying any calculations are. In general this sampling is not satisfactory as the true value of the parameter is not included in the confidence interval but what is interesting is the huge difference between both estimators calculated on the basis of *identical* data.

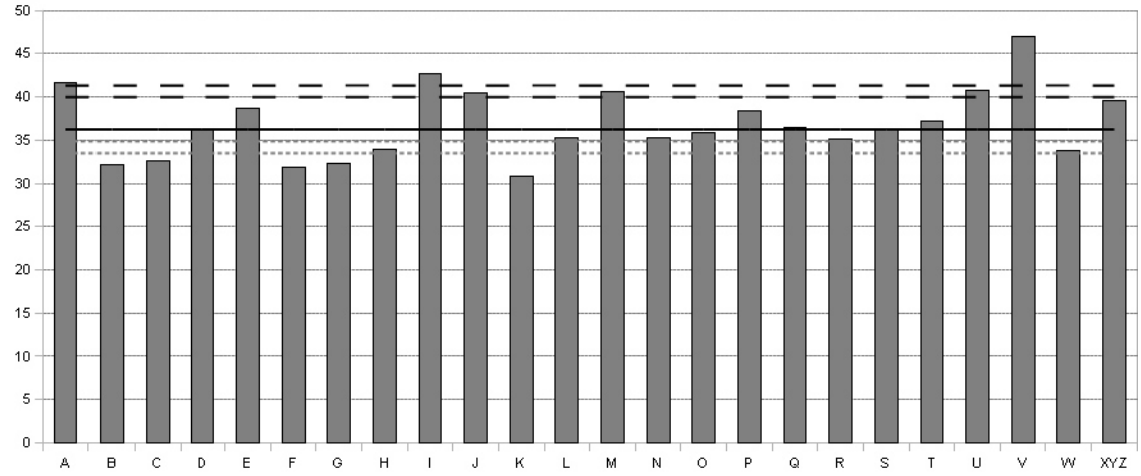

Figure 31: Stratified sampling – entries in PiotrSal

PiotrSal is a small dictionary and represents peculiar problems that might be encountered when dealing with small-sized strata. Therefore more research would be needed to answer the question whether other methods of stratification would yield better results or whether it is safer not to divide a small dictionary into strata. As we have not observed an increase in efficiency with the two larger dictionaries, the first questions remains valid for them as well.

5.2.3. Single dictionary entry as sampling unit – stratified cluster sampling

Now let us imagine that similarly to the section above, a single dictionary entry must be considered the ultimate SU but for some reason stratification is needed. The situation becomes slightly more complicated. Remember that multi-stage sampling consisted of first choosing a number of clusters and then again sub-sampling these clusters. I will use this scheme to get the desired stratification. This time, however, I will give formulas for

estimation of the total and not mean for simplicity reasons. Once the total is estimated, there are several methods of estimating the mean of which I will discuss only one.

I will use three-stage sampling to derive estimators for stratified cluster sample.

Let us assume our dictionary is divided into M letters, the i-th letter is itself divided into N_i pages consisting of L_{ij} entries. X_{ijk} will be a variable of interest attributed to k-th entry on j-th page of the i-th letter. It is not hard to realize that the whole dictionary total of all X_{ijk} variables will be the following:

$$X_T = \sum_{i=1}^{M} \sum_{j=i}^{N_i} \sum_{k=1}^{L_{ij}} X_{ijk} \tag{30}$$

Obviously we are not going to make a census of a dictionary so we need an estimate of the total. Denoting the variable connected with samples with lower case letters (see Appendix 1 for all notation details), an unbiased estimator of X_T is given as follows.

$$x'_T = \frac{M}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{N_i}{n_i} \sum_{k=1}^{l_{ij}} \frac{L_{ij}}{l_{ij}} x_{ijk} \tag{31}$$

But it is not a three-stage sampling that is needed but SCS. Please notice that the difference between those two sampling designs is that in both the first (selection of letters) and the third (selection of entries within a page) stage, not several but all elements are included in the sample. That is why after substituting *M* for *m* and L_{ij} for l_{ij} in (31) an estimator for X_T in SCS is obtained:

$$x_T = \sum_{i=1}^{M} \sum_{j=1}^{n_i} \frac{n_i}{N_i} \sum_{k=1}^{L_{ij}} x_{ijk} \tag{32}$$

Applying the same trick to the formula for x'_T 's variance in multi-stage sampling given in Deming (1950: 156) I obtained:

$$var(x_T) = \sum_{i=1}^{M} \left(\frac{N_i}{n_i}\right)^2 \left[\frac{N_i - n_i}{N_i - 1} n_i s_i^2\right] \tag{33}$$

One is, however, more likely to be interested, not in the total but in the mean. Nonetheless, having estimated the former, the latter should pose no problem as obviously $\bar{X} = \frac{X_T}{L}$, where L is the total number of entries in the dictionary. Estimating mean's variance is less straightforward but nonetheless not complicated: it suffices to apply the following property of variance:

(34) $\forall a \in \mathbb{R}\, var(aX) = a^2 var(X)$

Therefore

(35) $var(\bar{x}) = \frac{1}{l^2} var(x_T)$

Stratified cluster sampling and stratified sampling differ only in the way units within each stratum are selected. Therefore, after having applied formula (35), we should arrive at a formula almost identical to the one for stratified sampling, i.e. (25), with one basic difference: s_i^2 in (33) obviously stands for within-stratum variance for cluster sampling, whereas in (25) it is for simple random sampling. Let me now proceed to examining stratified sampling with the six entry-based characteristics in my dictionaries.

As one can see in Figure 32, stratified sampling proved very precise in the case of "obsolete" labeling in Webster. The confidence interval is not only neatly symmetrical around the true mean but also very narrow. Its length is only 0.0063 which translates into an increase in precision of more than 410% when compared with simple random selection of pages where the confidence interval length was 0.0259.

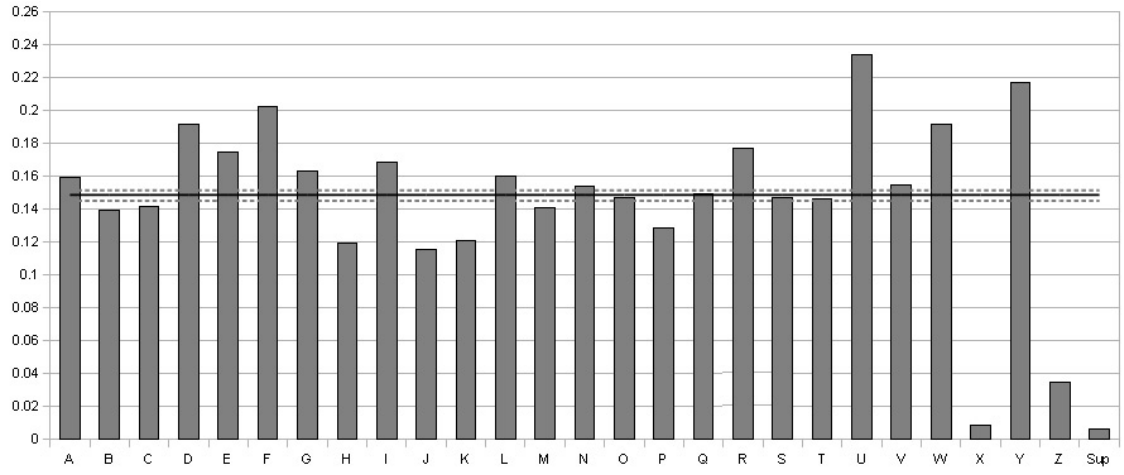

Figure 32: Stratified sampling – "obsolete" labelling in Webster

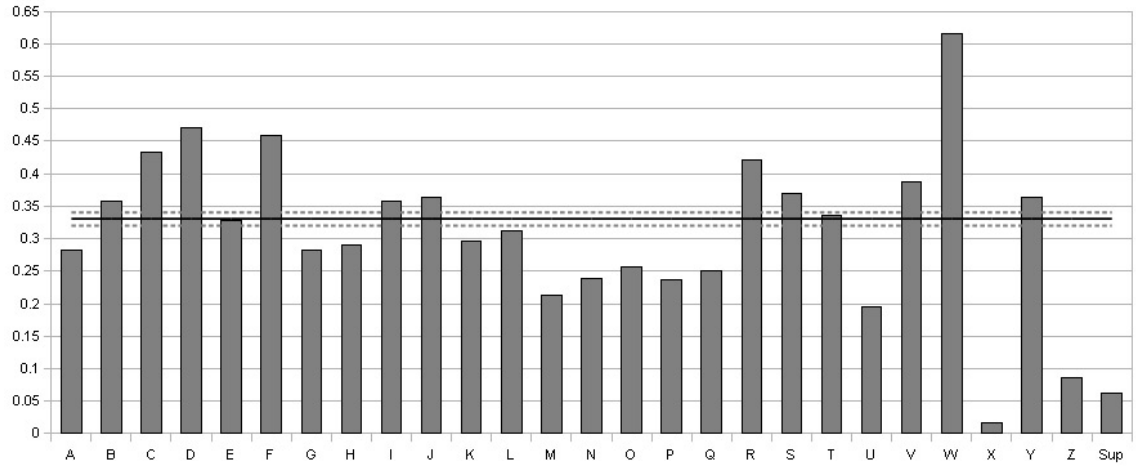

Figure 33: Stratified sampling – quotation provision in Webster

When it comes to NKFD, the results are equally satisfying. Figure 34 presents the estimation of the mean number of equivalents disambiguators per entry in this dictionary. Again, it proved much more precise than simple random selection of pages with an increase in precision of more than 360% (0.0332 vs 0.1203)

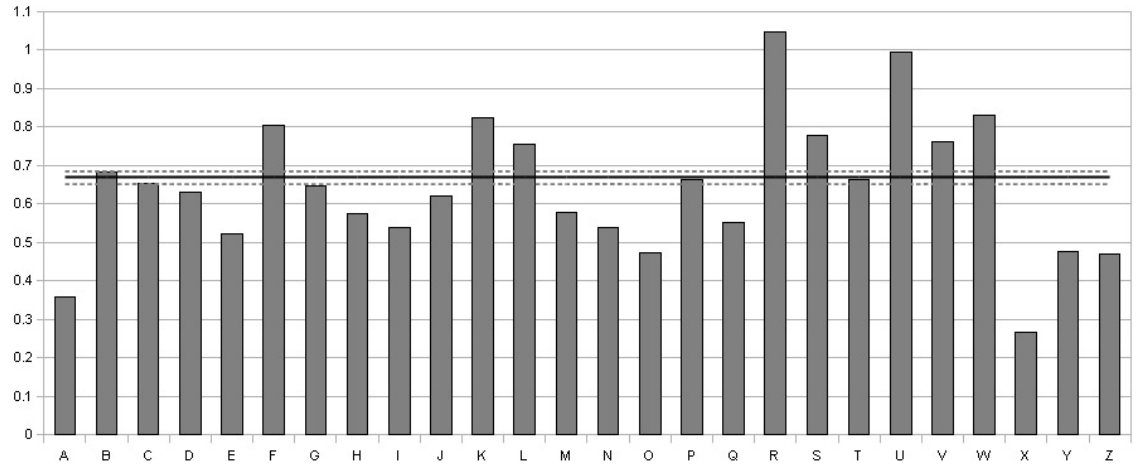

Figure 34: Stratified sampling – equivalent disambiguators in NKFD

The estimation of mean number of "formal" labels per entry in NKFD allows us to appreciate the value of stratification. As we already know, in this case there is a great deal of variation between within-stratum means. Despite this variation, stratified random sampling allows us to determine quite precisely where the true mean is. The length of the confidence interval is 0.0063 which translates into a 340% increase in precision

(the length of the confidence interval in the case of simple random selection of pages was 0.0215)

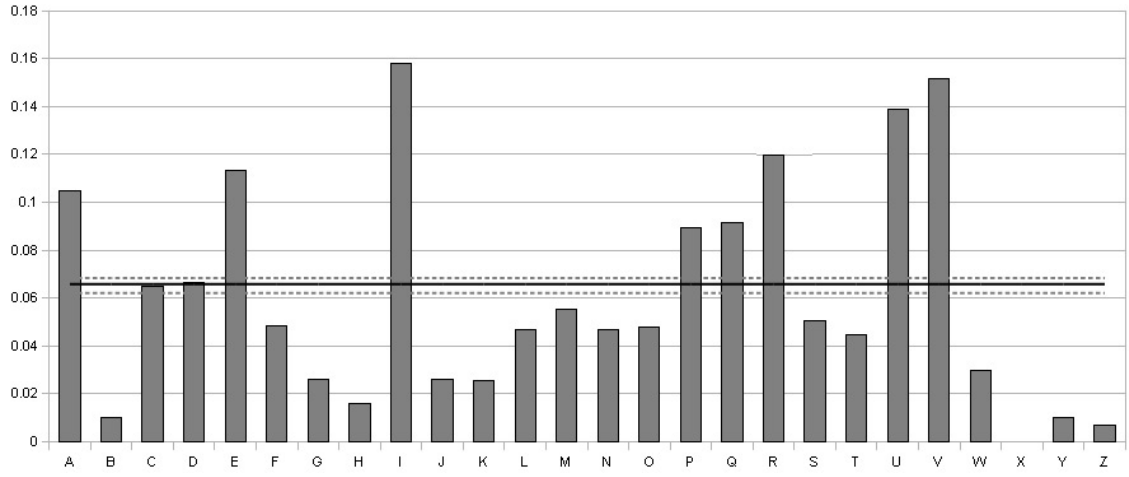

Figure 35: Stratified sampling – "formal" labeling in NKFD

Unfortunately, stratification in the case of PiotrSal proved quite disappointing. In the case of the mean number of equivalents provided per entry in PiotrSal the confidence interval is not only not narrower than the 0.2451 confidence interval in simple random selection of pages but it is almost 57% wider (0.3848).

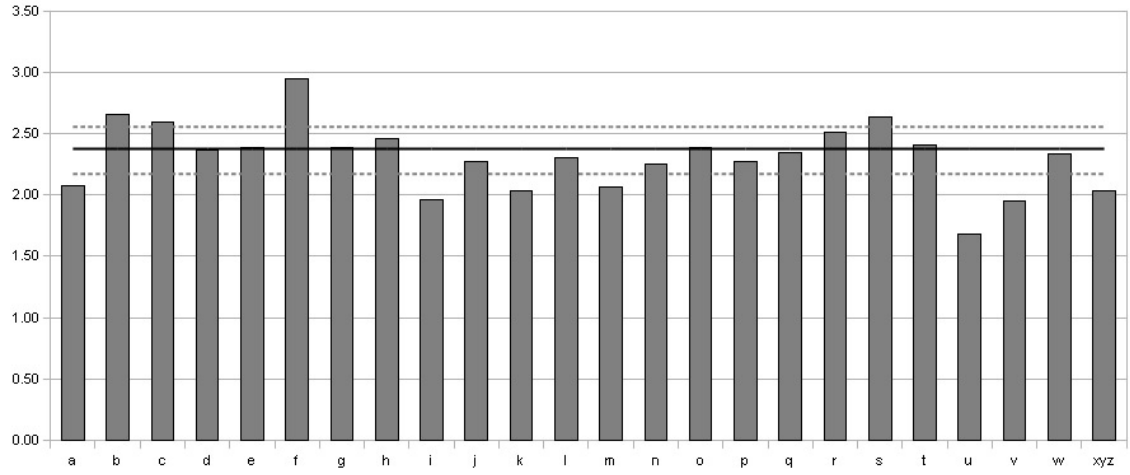

Figure 36: Stratified sampling – number of equivalents per entry in PiotrSal

Stratified sampling did not do better than simple random in the case of "US" labeling as one can see in Figure 37. As the reader will probably remember from section 5.1.3. , the confidence interval was very wide there. Unfortunately, here the estimate is even worse. The confidence interval is 94% longer with the length of 0.0262 which means that it is slightly more than 78% of the range of within-letter means. Obviously,

this is far from satisfactory. Nonetheless there is still some positive side to it. First, I will repeat what I already pointed out in 5.1.3. : this method tells us that the results are not satisfactory and forces us to search for other solutions. Secondly, unlike in single-stretch sampling, the point estimate is not very different form the true dictionary mean. The difference between the two is only 0.0013 and one can see it by the relative symmetry of confidence interval endpoints around the true dictionary mean. We only learn that the confidence with which we can draw these inferences is very low.

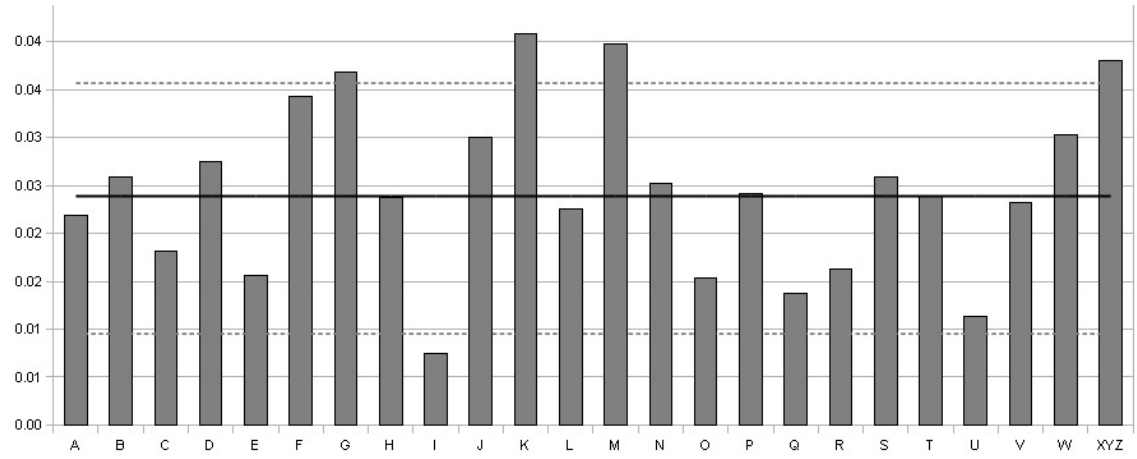

Figure 37: Stratified sampling – "US" labeling in PiotrSal

By means of conclusion to this section one can say that when dealing with a large dictionary and entry-based characteristics, stratification is worth the trouble. With page-based characteristics, stratification did not bring about an increase in precision in the dictionaries under investigation. Two factors may account for that: the relative uniformity of the distributions of this characteristic throughout the dictionary combined with the untenability of condition (27). By contrast, with entry-based characteristics strata get large enough for (27) to be tenable. A small dictionary, PiotrSal in our case, seems to present another set of problems. When estimating the mean number of entries per page, i.e. a page-based characteristic, the true mean fell outside the confidence interval, whereas with both entry-based characteristics, the estimation was remarkably less precise than with simple random selection of pages. It remains an open question whether this was just a random effect or whether one could improve on this by choosing another method of stratification, e.g. by clustering letters.

5.3. Summary – all random techniques compared

It this section I would like to offer a brief overview of all random (including systematic) sampling techniques examined with the use of data from all the dictionaries. Basically two criteria will be used for this final evaluation: whether the true mean is contained in the confidence interval, and the length of the confidence interval. From these criteria it follows that neither single stretch sampling nor the Coleman and Ogilvie 2009 method will be summarized here.

Figures 38, 39 and 40 present the lengths of the confidence intervals in the form of bars for each characteristic examined in Webster, NKFD and PiotrSal respectively. The cases where the true mean was not contained in the confidence interval are marked with hatching. Unfortunately, the figures do not show the distance between the confidence interval endpoint and the true mean in such cases.

What follows from these figures is that in two cases, systematic sampling of entries in NKFD and stratified sampling of entries in PiotrSal, the true mean number of entries per page is not included in a suitable confidence interval. As there were altogether nine characteristics examined, this fell below my expectations. Recall that the α level adopted in the study was 5%. Nonetheless, the distance between the endpoint of the confidence interval and the true mean in the case of systematic sampling in NKFD was indeed very close to zero. By contrast the lack of accuracy of estimation in PiotrSal was considerable.

Apart from that, these figures neatly summarize and visualize what I have already stated in 5.2.3. Namely, we can see clearly in what cases stratification proved to be worth the trouble. As we can see, in none of the cases did it yield better estimation of the mean number of entries per page. In both Webster and NKFD the confidence interval proved longer than in systematic and simple random sampling. In PiotrSal stratification did manage to reduce variance but at the same time it yielded considerably skewed estimates and therefore cannot be considered an improvement over simple random or systematic sampling. When we move to entry-based characteristics we can see that in large dictionaries i.e. Webster and NKFD, stratification managed to reduce the length of the confidence interval significantly with no detriment to accuracy. In a small dictionary, PiotrSal, the results are exactly the reverse: with both characteristics the confidence interval for stratified sampling proved considerably wider.

It seems that in the case of a small dictionary condition (27) might not be tenable as this condition guarantees that stratified sampling yields estimates at least no worse than simple random sampling. Nonetheless, intuitively one could expect that in small dictionaries letters would exhibit relative uniformity at least when compared with the whole dictionary text. This, on its part, should contribute to a higher degree of precision (cf. Section 5.2.2.). Therefore we can conclude that short stretches of dictionary text do not exhibit the expected uniformity. One can ponder on whether this is for language internal reasons or it is still connected with alphabet fatigue.

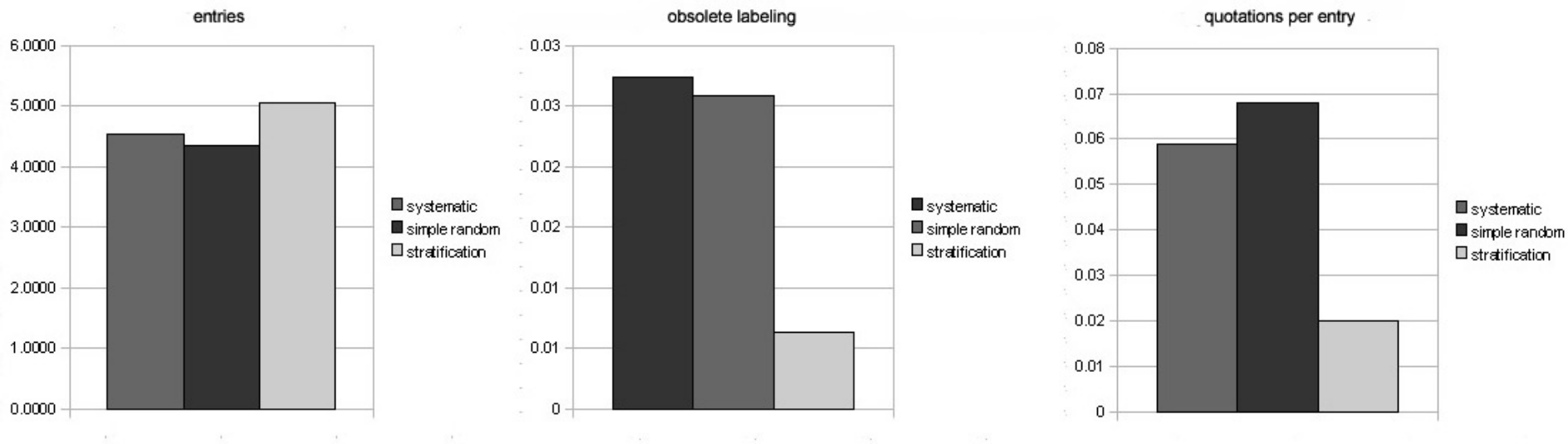

Figure 38: Webster – comparison of CI lengths

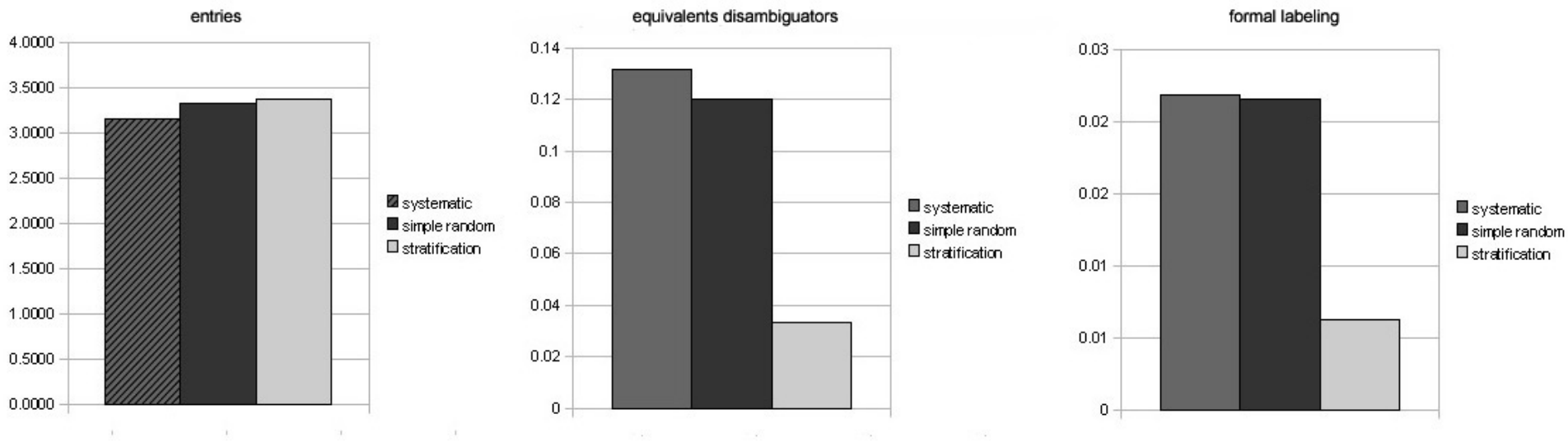

Figure 39: NKFD – comparison of CI lengths

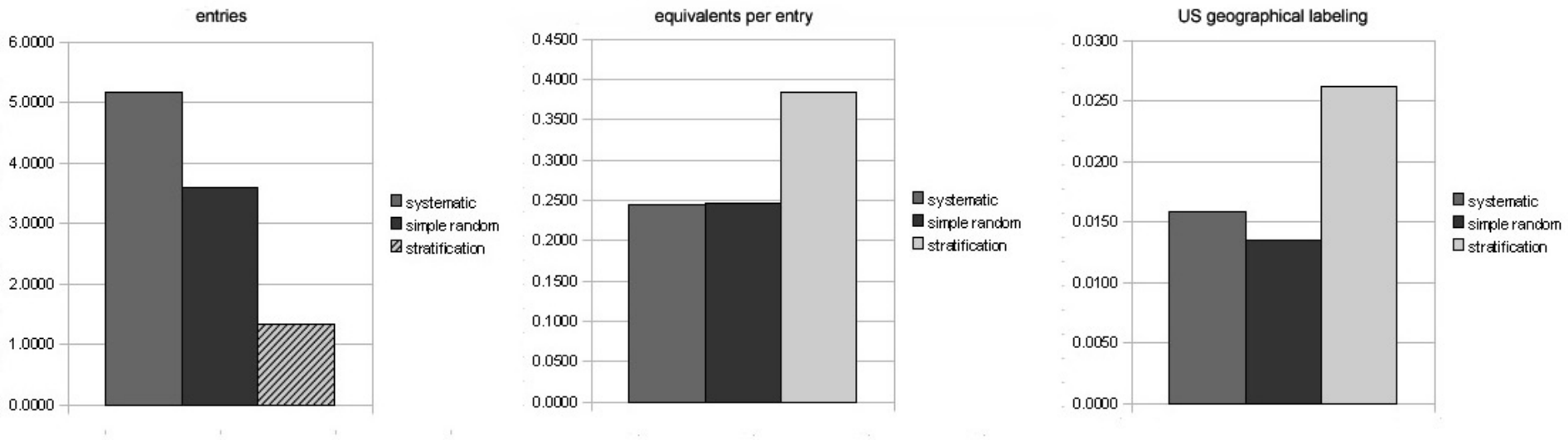

Figure 40: PiotrSal – comparison of CI lengths

Conclusion

The aim of the current study was to assess the reliability (or the lack thereof) of sampling methods currently used in metalexicographic research. As my hypothesis was that the most commonly used method i.e. judgmental selection of one letter in the dictionary is not reliable enough, I also propose random sampling techniques commonly encountered in statistics i.e. systematic sampling, simple random sampling and stratified sampling and I evaluate the precision with which inferences can be drawn by using confidence intervals.

I hope I managed to show that single-stretch samples based on a judgmentally chosen letter of the alphabet, all too often encountered in metalexicographic research, present a substantial threat to the reliability of research results. For such methods to yield acceptable results, the characteristic in question would have to exhibit uniform distribution throughout the whole dictionary text. As I have shown, this condition is *never* satisfied and usually one cannot even claim that it is *roughly* satisfied. In the majority of cases the distribution of within-letter means exhibits unpredictable irregularities either due to inconsistencies during the dictionary compilation process, or for language internal reasons. Therefore, thought must be given to the way scientists select material for metalexicographic research so that this selection bias could be eliminated.

Selection bias could be effectively eliminated using random sampling techniques or systematic sampling. My data show that there are no considerable differences in precision between systematic sampling and simple random selection of pages. The next step was to examine whether stratification i.e. dividing the dictionary text into non-overlapping parts and selecting a sample independently form each of them would bring about an increase in precision. In this case the answer is less straightforward: it turned out that, at

least with our three dictionaries and with stratification according to letters of the alphabet, it may depend both on the type of characteristic examined and dictionary size. In the case of page-based characteristic, stratification turned out to be no better than simple random sampling. Considerable improvements were noted for entry-based characteristics in the two larger dictionaries, i.e. Webster and NKFD. PiotrSal, a small dictionary, revealed a set of problems that may well be typical of dictionaries of that size. First of all, it showed the importance of using exact weighting in estimator formulas. Second of all, not only no improvements in precision have been observed but sample variance proved greater than in simple random selection of pages. Stratified estimate of the mean number of entries per page proved highly inaccurate. Additionally, replicating Coleman and Ogilvie 2009 sampling demonstrated the importance of randomization within strata.

It must be borne in mind that the simulations were performed with three dictionaries and nine characteristics only. Therefore the results must be treated with caution. I would like to remind the reader, however, that there is a large body of mathematical theory behind random sampling. Therefore doubts should pertain mostly to empirical assessment of non-random sampling techniques i.e. single-stretch sampling and the Coleman and Ogilvie approach.

Yet another reservation might concern the characteristics. Those were chosen because they were easily identifiable automatically in the SGML-tagged dictionary text and easily quantifiable. I think that there is no reason to expect a more uniform distribution that would justify judgmental sample selection in the case of other characteristics. I am convinced that even when studying characteristics of less quantitative and more qualitative nature, the same guidelines would apply to get a non-distorted and comprehensive picture of a given dictionary. Nonetheless, formal and precise evaluation of sampling techniques would not be possible methodologically if characteristics of a qualitative nature would have been chosen.

There are also several limitations to the present study and questions that need to be addressed in further research. First of all, when examining stratified sampling I have been preoccupied with estimating the overall mean in the whole dictionary. Nonetheless, stratification might be performed not only in order to get a clearer picture of a dictionary as a whole but also in order to compare different sections of the dictionary. Therefore estimating within-stratum means would have to be examined in more detail. Second of all I assumed that one paper dictionary is to be sampled. Obviously one may

want to sample electronic dictionaries or a correlated sample from several dictionaries might be needed. I will now outline the specific problems related to these two issues.

When doing comparative research on several dictionaries an approach of intuitive appeal would be to select a sample randomly from one dictionary, and then match the comparator text in the remaining dictionaries. As seen in e.g. Ardilly and Tillé (2006: 15), there is more to it than just intuition. Even though statisticians treat cluster sampling more as a necessity than a way of improving research results, I think that in dictionary comparison, cluster sampling might present substantial advantages. Taking a dictionary stretch as a sample, be it a page or any other unit, allows the researcher not only to compare what is present in the two dictionaries, but also what is absent from one of them. Straightforward as this approach might seem, there are nonetheless several questions to be addressed. First of all, suppose our dictionaries differ considerably in size and type of alphabetization. Sampling a larger dictionary first presents the advantage of obtaining a possibly fine-graded picture. On the other hand, however, when matching the comparative text in smaller dictionary one might then encounter "interval endpoint absent" problem. In extreme cases a given stretch in a large dictionary might even translate into an empty set in a smaller one. Starting with the smaller dictionary minimizes the risk of encountering such a problem, but having a small number of larger clusters might result in a considerable bias (as the estimator for cluster sampling is only asymptotically unbiased). Yet another problem would be dealing with differences in alphabetization or heavy cross-referencing (as in the case of the Dictionary of Regional American English where entries consisting of a cross-reference only account for more than a third of the total number of entries[7]). Finally, let as imagine that at least one of the dictionaries is subject to heavy over- and under-treatment (in De Schryverian terms). Obviously, if pages were sampling units in the first dictionary studied, it will certainly not be the case in the remaining ones. This will not be much of a problem in well balanced dictionaries; as we might assume we have the same number of clusters as we have pages in the first dictionary and the cluster size will roughly be the same. If at least one of the dictionaries compared is heavily out of balance, the matched stretches in the second dictionary will have considerably different sizes.

[7]Point estimate derived from a 3% stratified sample of DARE

Streszczenie po polsku

Tematem prezentowanej pracy jest zagadnienie próbkowanie w badaniach metaleksykograficznych. Wstępna analiza tego typu badań wykazała, ze temat ten jest często lekceważony przez czołowych badaczy, a najczęściej spotykaną próbą jest próba celowa, co w przypadku słownika przekłada się na wybranie jednej litery alfabetu, którą badacz uważa za reprezentatywną i przebadanie jej w całości. W intuicji autora niniejszej pracy tego typu metodologia stwarza poważne zagrożenie dla wiarygodności wyników badań. Z tego powodu poza empiryczną oceną tego typu próbkowania na tekście trzech słowników przeprowadzony został eksperyment mający na celu ocenę probabilistycznych metod próbkowania w kontekście metaleksykograficznym.

Zakłada się, że badanie dotyczy przede wszystkim próbkowania słownika papierowego, jednakże dla celów eksperymentalnych użyto elektronicznej siatki istniejących słowników wydanych w wersji papierowej: The New Kościuszko Foundation Dictionary – (NKFD) English-Polish (Nowy Słownik Fundacji Kościszkowskiej, Angielsko - Polski), Webster's Revised Unabridged Dictionary (Webster), oraz New English-Polish Dictionary (Nowy Słownik Angielsko-Polski) (PiotrSal). Do celów eksperymentalnych wybrano parametry charakteryzujące stronę jak i takie, które charakteryzują poszczególne hasła słownikowe, gdyż od tego podziału uzależniony jest wybór modelu matematycznego użytego do estymacji. Jedynym przykładem parametru charakteryzującego stronę w tym badaniu jest średnia liczba haseł na stronę. Wybrane parametry charakteryzujące hasła to etykiety oznaczające słowa i wyrażenia przestarzałe (“obsolete”) oraz średnia ilość cytowań przypadających na hasło w Websterze, etykiety oznaczające słowa i wyrażenia o charakterze formalnym (“formal”) oraz średnia liczba wyjaśnień uściślających znaczenie ekwiwalentów (equivalent disambiguators) w NKFD, wreszcie

średnia liczba ekwiwalentów przypadających na hasło i etykiety wskazujące na amerykański charakter danych słów w słowniku PiotrSal. Dla każdego z wyżej wymienionych parametrów przeprowadzono cenzus a następnie wybrano próbę systematyczną, prostą i warstwową stron danego słownika. Dane z próby posłużyły do wyliczenia estymatorów średniej oraz ich przedziałów ufności. Moim celem było uzyskanie estymatorów o możliwie małej wariancji przy jednoczesnej eliminacji obciążenia będącego rezultatem nielosowego doboru próby, który dominuje w obecnych badaniach metaleksykograficznych. Estymatory te, wraz ze swoimi przedziałami ufności, zostały porównane do estymatorów pochodzących z prób jedno-literowych. Zreplikowano także próbkowanie z badania Coleman – Ogilvie (2009), w celu udzielenia odpowiedzi na pytanie, czy można dopuścić brak losowości w przypadku stosowanie próbkowania warstwowego.

Zgodnie z oczekiwaniami wewnątrz-literowe średnie prezentowały bardzo niejednorodne rozkłady, w związku w czym estymacja na podstawie próbkowania jednej wybranej litery alfabetu narażona jest na znaczące obciążenie. Zarówno próbkowanie systematyczne jak i metody losowe okazały się dostarczać zadowalających wyników we wszystkich przypadkach. Nie zaobserwowano znaczących różnic w długości przedziału ufności między próbą systematyczną a prostą. W przypadku parametrów charakteryzujących hasła w dużych słownikach, czyli w Websterze i w NKFD, zaobserwowano około trzy- do czterokrotny wzrost precyzji oszacowań będący wynikiem pobrania próby warstwowej (dokładniej mówiąc grupowo-warstwowej). Nie zaobserwowano podobnego efektu w przypadku średniej liczby haseł na stronę oraz w małym słowniku czyli w PiotrSal. Słownik ten natomiast ukazał problemy, które mogą być charakterystyczne dla próbkowania słowników o podobnym rozmiarze: bardzo szerokie przedziały ufności w przypadku niektórych parametrów oraz konieczność dokładnej weryfikacji założeń; w szczególności założenia o proporcjonalności alokacji w próbkowaniu warstwowym. Replikacja badania Coleman – Ogilvie (2009) wykazała, że losowość próby wewnątrz warstw ma kluczowe znaczenie; wyliczone estymatory rzadko znajdowały się wewnątrz przedziału ufności dla losowej próby warstwowej.

Pomimo jasnych wytycznych dotyczących próbkowania słowników papierowych, które udało się ustalić dzięki obecnemu badaniu, nie rozwiązuje ono problemów związanych z próbkowaniem słowników elektronicznych, czy też z pobieraniem prób do badań porównawczych nad tekstami kilku słowników, w szczególności jeśli różnią się one znacząco rozmiarem i sposobem alfabetyzacji.

References

Dictionaries:

Adamska-Sałaciak, Arleta. 2003. *New Kościuszko Foundation Dictionary, vol. I English-Polish* [NKFD]. New York. Kraków: Universitas

Dąbrówka, Andrzej, Ewa Geller and Ryszard Turczyn. 1993. *Słownik Synonimów*. Warszawa: MRC.

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

Piotrowski, Tadeusz and Zygmunt Saloni. 1999. *New English-Polish, Polish-English dictionary: Part 1, English-Polish* [PiotrSal]. Warsaw: Spotkania.

Porter, Noah (ed.). 1913. *Webster's Revised Unabridged Dictionary* [Webster]. Springfield: G & C. Merriam Co.

Rundell, Michael. (ed.). 2002. *Macmillan English Dictionary for Advanced Learners* [MED]. Oxford: Macmillan Education.

Summers, Della. 2003. *Longman Dictionary of Contemporary English* [LDOCE]. London: Pearson Education Ltd.

Wehmeier, Sally. 2000. *Oxford Advanced Learner's Dictionaries of Current English* [OALD]. Oxford: Oxford University Press.

Other sources:

Ardilly, Pascal and Yves Tillé. 2006. *Sampling methods: exercises and solutions*. New York: Springer.

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

Coleman, Julie and Sarah Ogilvie. 2009. "Forensic dictionary analysis: Principles and practice", *International Journal of Lexicography* 22, 1: 1-22.

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

Cormier, Monique and Heberto Fernandez. 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, 4: 479-507.

De Schryver, Gilles-Maurice. 2005. "Concurrent over- and under-treatment in dictionaries – The Woordeboek van die Afrikaanse Taal as a case in point", *International Journal of Lexicography* 18, 1: 47-75.

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

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

Martínez Egido, José Joaquín (2002). *La obra lexicográfica de Lorenzo Franciosini: "Vocabulario italiano-español, español-italiano" (1620) [Lexicografic works of Lorenzo Franciosini: "Vocabulario italiano-español, español-italiano" (1620)]*. Unpublished doctoral dissertation.

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

Ogilvie, Sarah. 2008. "Rethinking Burchfield and world Englishes", *International Journal of Lexicography* 21, 1: 23-59.

Random.org. (http://www.random.org/sequences) (date of access: October 2009).

Roberts, Roda. 2007. "Dictionaries and culture", in: Henrik Gottlieb and Jens Erik Mogensen (eds.), *Dictionary visions, research and practice: Selected papers from the 12th International Symposium on Lexicography, Copenhagen 2004*. Amsterdam: John Benjamins, 277–297 .

Rodríguez-Álvarez, Alicia and María Esther Rodríguez-Gil. 2006. "John Entick's and Ann Fisher's dictionaries: An eighteenth-century case of (cons)piracy?", *International Journal of Lexicography* 19, 3: 297-319.

Wood, Anthony, Paul Fletcher and Arthur Hughes. 1986. *Statistics in language studies*. Cambridge: Cambridge University Press.

Xu, Hai. 2005. "Treatment of deictic expressions in example sentences in English learners' dictionaries", *International Journal of Lexicography* 18, 3: 289-311.

Xu, Hai. 2008. "Exemplification policy in English learners' dictionaries", *International Journal of Lexicography* 21, 4: 395-417.

Zgusta, Ladislav. 1971. *Manual of Lexicography*. Prague: Academia.

Appendix 1

Notation used

The tables below present the notation used throughout the thesis. Those are given here mostly as a reference point for further reading.

Table 1: Notation for simple random sampling

Name	**Population**	**Sample**
Number of SUs	N	n
Sampling fraction	$f=(n/N)$	
Population of x-characteristic	$X_i(i=1\cdots N)$	$x_i(i=1\cdots n)$
Mean	$\bar{X}$	$\bar{x}$
Total	X_T	x_T
Variance	$\sigma^2=\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})^2$	$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

Table 2: Notation for stratified sampling

Name	**Population**	**Sample**
Number of SUs	N	n

Number of strata	k	
Stratum size	$N_i (i=1 \cdots k) \sum_{i=1}^{k} (N_i = N)$	$n_i (i=1 \cdots k) \sum_{i=1}^{k} (n_i = n)$
Sampling ratio	$f_i = \frac{n_i}{N_i} (i=1 \cdots k)$	
Population of x-characteristic	$X_{ij} (i=1 \cdots k\,;\, j=1 \cdots N_i)$	$x_{ij} (i=1 \cdots k\,;\, j=1 \cdots n_i)$
Stratum means	$\bar{X}_i (i=1 \cdots k)$ $\bar{X}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} X_{ij}$	$\bar{x}_i (i=1 \cdots k)$ $\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$
Intra-stratum variance	$\sigma_i^2 (i=1 \cdots k)$ $\sigma_i^2 = \frac{1}{(N_i - 1)} \sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i)^2$	$s_i^2 (i=1 \cdots k)$ $s_i^2 = \frac{1}{(n_i - 1)} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$

Table 3: Notation for cluster sampling

Name	**Population**	**Sample**
Number of elements (e.g. dictionary entries)	N	n
Number of clusters	M	m
Average number of elements per cluster	$\bar{N}$	$\bar{n}$
Number of elements in a cluster	$N_i (i=1 \cdots M)$ $\sum_{i=1}^{M} N_i = N$	$n_i (i=1 \cdots m)$ (obviously $n_i = N_i$)
Sampling ratio	$F = \frac{n}{N}$	
Cluster mean	$\bar{X}_i (i=1 \cdots M)$	$\bar{x}_i (i=1 \cdots m)$
Cluster variance	$\sigma_i^2 (i=1 \cdots M)$	$s_i^2 (i=1 \cdots m)$
Cluster total	$X_{iT} = \sum_{j=1}^{N_i} X_{ij}$	$x_{iT} = \sum_{j=1}^{N_i} x_{ij}$

Table 4: Notation for stratified cluster sampling

Name	**Population**	**Sample**
Number of PSUs	M	mm
Number of SSUs in the i-th PSU	N_i	n_i
Number of TSUs in the j-th SSU	L_{ij}	l_{ij}
Total number of SUs	L	l
Population of x-characteristic	$X_{ijk}(1\cdots M;1\cdots N_j;1\cdots L_{ij})$	$x_{ijk}(1\cdots m;1\cdots n_j;1\cdots l_{ij})$
Internal variance per SSU within the i-th PSU	$\sigma_i^2(i=1\cdots k)$ $\sigma_i^2=\frac{1}{(N_i-1)}\sum_{j=1}^{N_i}(X_{ij}-\bar{X}_i)^2$	$s_i^2(i=1\cdots k)$ $s_i^2=\frac{1}{(n_i-1)}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$