The Distribution of English Isograms in Google Ngrams and the British National Corpus

Breit, Florian

Opticon1826

Unpublished: 01/01/2017

Peer reviewed version

The Distribution of English Isograms in Google Ngrams and the British National Corpus

Florian Breit
Research Department of Linguistics, University College London, 2 Wakefield Street, London WC1N 1PF, UK. Email: florian.breit.12@ucl.ac.uk

Abstract
The study of isograms—words in which each letter occurs the same number of times—has thus far largely been limited to manual search for examples in sources such as dictionaries, and accounts have principally limited themselves to simply listing the known isograms of various categories. This paper presents the results of a corpus study of English isograms from Google Ngrams (ca. 1 trillion words, ~13 million types) and the British National Corpus (ca. 100 million words, ~6 million types). The paper discusses methodological issues relating to the automated mining of isograms, explores the distribution of isograms in relation to word-length and frequency, and presents several new isograms, which have so far gone unnoticed in the literature. Moreover the paper describes the resultant dataset of English isograms and the tools used to create it, which are made freely available and can be used to further study the distribution of isogramy in English and other languages.

Keywords
Isograms, orthography, glottometrics, logology, English

Introduction

Isogramy

The term *isogram* was first introduced by Borgmann (1965:125), who applied the term to 'a word that uses no letter of the alphabet more than once'—otherwise also labeled as 'nonpattern words'. Some examples of English nonpattern words are given in (1) below.

(1)	*word*	*plain*
	juniper	*bread*
	balcony	*monkey*

As opposed to the more strict requirements of non-repetition in nonpattern words such as those in (1), the term isogram can be construed more widely to capture the notion of any

word in which each letter of the alphabet occurs exactly the same number of times (cf. Borgmann 1974). The amount of times each letter occurs indicates the order of isogramy. Thus, a word in which each letter occurs exactly once is a first-order isogram, one in which each letter occurs exactly twice is a second-order isogram, and so forth (cf. Crystal 2007b). Thus, nonpattern words are a special case of isograms, namely first-order isograms. Some examples are given in (2) below.

(2) a. *plain* *bread*
b. *bilabial* *deed*
c. *deeded* *geggee*

For convenience, we can refer to an isogram of order *n* as an *n-isogram*. The examples in (2a) are 1-isograms, (2b) 2-isograms, and (2c) 3-isograms. Second- and third-order isograms are also sometimes referred to as *pair* and *trio isograms* respectively, but I will not use these terms here.

Purpose and Goals

As far as it is discernible from the available literature, methodological investigations of isogramy have hitherto relied on introspection and manual searches of sources such as dictionaries and atlases. Moreover, while it has been variously stipulated that length and order of isogramy are inversely related to the number of words which fit the criteria, these relations have never been quantitatively investigated. The same applies to the relations between isogramy, tautonymy and palindromy discussed presently. The present paper seeks to provide at least tentative answers to some of these questions by employing computational methods to mine word lists from Google Ngrams (henceforth simply Ngrams) and the British National Corpus (BNC) for isograms. Presenting a first foray into the computational mining of isograms, the paper moreover seeks to make available to the wider logological community a set of tools, which can be easily adapted to apply the methods of this study to different languages and datasets.

The results and datasets from this study add value beyond logological study by providing a first foundation for further quantitative research on pattern-restricted subsets of text (such as isograms, palindromes, tautonyms, etc.). One way in which this might be applied in the future is to study in how far such special subsets of text follow the same distributional regularities as are known from English and other language text more generally (cf. e.g. Altmann 1988, Wimmer & Altmann 1996, Smith 2012). The results can also be used as a basis for studying whether isograms may be processed differently from non-isogrammic words (possibly depending on their order of isogramy), something which has been proposed previously for palindromes (Shillock, Kelly & Monaghan 1998).

Isograms versus Palindromes and Tautonyms

Borgmann (1974), Grant (1982), and other sources note early on that there is a not negligible overlap between isogramy, palindromy and tautonymy. They go as far as to make the point that isograms which are also palindromes or tautonyms should be disregarded, on the presumption that they are isograms principally by virtue of being palindromes or tautonyms. Consider for instance the isograms in (3), taken from Grant (1982):

(3) a. *terret*
gnipping

b. *beriberi*
tartar

The examples in (3a) are both palindromes and, according to the definition of isogramy given on page 1, 2-isograms. Another example from Borgmann (1974) is *detannated*. The examples in (3b) are tautonyms and, according to the definition used in this paper, also 2-isograms.

While tautonymy has sometimes been restricted to mean reduplication as specification—for instance in the naming of a genus such as the *bison bison* (the Latin name for the American buffalo)—it is frequently more widely applied to include any case of full repetition and would therefore possibly also include morphologically simplex forms such as *mama* and *papa*. In actuality, a case could be made for being more selective about which forms to exclude based on whether they are morphologically complex, such as the term *bison bison*, or forms which, at least for English speakers are lexically treated as a simplex form, which would on closer inspection include both *tartar* and *beriberi*. While *beriberi* is originally indeed a case of morphological reduplication from the Sinhalese simplex *beri* 'weak', this information is of course not available to a modern English speaker acquiring the form only as *beriberi*, never knowing that a form such as *beri* 'weak' even exists. In other instances, such as the terms *mama* and *papa*, this form of reduplication is more or less accidental.[1] Nonetheless, while English is not one of the languages that makes systematic use of morphological reduplication, there exist such forms that can even then be treated as morphologically complex reduplication, for instance *poo* versus *poopoo*. The example of *poopoo* shows moreover that tautonymy, in neither the restricted nor the loose sense, necessitates isogramy. A tautonym is only an isogram if the repeated string is also an isogram—although given the increased incidence of lower order isogramy among short words the density of second order isogramy can be expected to be somewhat higher among tautonyms than other forms.

While palindromy, especially where the number of letters is even (meaning that the medial letter must be repeated), may also increase the likelihood that a word is a 2-isogram, this is not necessarily the case. For instance, a word of the form *abbccbba*, although having an even number of letters, would not constitute an isogram. Palindromes with an uneven number of letters have by definition one fewer occurrence of the medial grapheme than they can possibly have of any others and thus cannot ever be isograms.

While an initial assumption then may be that tautonymy and palindromy, which both feed into isogramy and at the same time make these words interesting for different reasons, may make them less interesting as isograms, there is actually a good case that they are of special interest if considered against the list of tautonyms and palindromes. One may wonder about palindromes of even length which are not isograms and, to a lesser extent, about tautonyms which are not isograms, or about isograms which are morphologically simplex tautonyms versus those which are morphologically decomposable into an actually reduplicated lexical item. This exclusion has also sometimes been extended to forms in which any significant sequence of letters repeats, for instance, in *senescence* (cf. Grant 1982). Borgmann (1985a) is, however, quick to note that overly harsh exclusion of items in any of these three categories will exclude nearly all long higher-order isograms and if isograms of a given length or order are sufficiently rare they are interesting nonetheless; in addition these forms may be even more interesting in that many of them also exhibit cadence.[2] In the

remainder of the paper, while I will occasionally remark on such matters, I will, given the above considerations, refrain from making any exclusions of items either based on repetition or palindromy.

Isograms and Word Length

In his 1974 overview, Borgmann discusses at length the issue of which are the longest isograms of each order of isogramy. Pertaining to 1-isograms, he suggests that the level of rarity, at which examples become interesting, is around 15 letters. The only attested single word example of length 15 he gives here (and in Borgmann 1965) is *dermatoglyphics*, the study of fingerprints. He further suggests several possible coinages, such as *uncopyrightable* from the attested 14 letter isogram *copyrightable*, **misconjugatedly* from attested *misconjugated* and *hydropneumatics* from attested *hydropneumatic*, although note that both *uncopyrightable and hydropneumatics* have since been attested and feature in the Oxford English Dictionary, with quotations going back to 1927 and 1887, respectively, thus pre-dating Borgmann's suggestion.[3] Borgmann (1985a) gives a long list of 15 letter isograms; however, a considerable amount of these are of coinages or words otherwise of questionable attestation. One additional example of interest from this list is *endolymphaticus*, because it occurs exclusively as part of the Latin anatomical term *ductus endolymphaticus*.

At 16 letters, the only example Borgmann (1974) finds is a place name, '*South Cambridge, N. Y.*',[4] but he suggests the coinage of ?*uncopyrightyables* from ?*copyrightables* as a possibility, a term which has likewise seen limited logology-independent attestation since being coined. Borgmann (1985a), Wolpow (1991) and Eckler (1997) report an example of 17 letters, ?*subdermatoglyphic*, attested at least in one independent source, namely a dermatology paper by Goldsmith 1990—although, as Wolpow points out, he personally discussed the term with Goldsmith before and so its attestation is not actually fully independent of the logological literature. The only longer examples in Borgmann (1985a) and elsewhere are personal names, real or hypothetical.

For 2-isograms, Borgmann (1974) suggests that 10 letter examples are already of interest. He gives numerous examples, repeated in (4), although some of these examples are also place names such as *Succasunna* and orthographically loose compounds (i.e. forms usually written as separate words without hyphenation), such as *tool steels* and *swing wings*, which I omit here for reasons of simplicity.[5] In addition to (4), Grant (1982) reports 10 letter *well-wooded*.

(4)

arraigning	*notionists*	*tromometer*
concisions	*reproposes*	*horseshoer*
insciences	*rereigning*	*intestines*
ma'amselles	*retardated*	*Superpures*
tessellata		

At 12 letters he reports eight examples, repeated here as (5), with the exception of *Transnistria*, the name of a former Romanian administrative division. Grant (1982) also adds *charactereth* and *Tukitukipapa*, which is a place name of Maori origin.

(5) *cancellanses* *interinserts*
cicadellidae *shanghaiings*
gradgrindian *trisectrices*
happenchance

At 14 letters length, Borgmann (1974) reports *scintillescent*, *unsufficiences* and *Taenidontidae* (originally reported by Darryl Francis, see Eckler 1971) and also notes that the verb phrase *are integrating* could be counted as a potential 14 letter example. Borgmann (1985b) adds *inaccidentated*, an attested theological term relating to the theory of transubstantiation, and then unattested coined *unconstructors*, intended to refer to 'those who tear down what others have laboriously erected' (Borgmann 1985b:142), but which has since seen limited attestation with a different meaning in object-oriented programming. At 16 letters and more, Borgmann (1974) reports only French *antiperspirantes* 'antiperspirant (ADJ.FEM)' and several coinages, but Borgmann (1985c) gives the now marginally attested *noninstallations*, occurrences on which an otherwise planned installation did not take place. At more than 16 letters, Borgmann (1985b) only suggests place names and coinages.

3-isograms have been little studied thus far and Borgmann (1974) reports only two 6 letter examples: *deeded* and *geggee*, the victim of a hoax. Grant (1982) adds *feffee*, a variant spelling of *foeffee*, and *seeses*, an OED attested early spelling variant of *ceases*. Grant (1982) also makes mention of the two OED attested 9 letter forms *sestettes*, a variant spelling of *sestets*, and *sheeshehs*, a variant spelling of *shishas*, both of which he attributes to Darryl Francis.

Both Borgmann (1974) and later Crystal (2007a) report being unaware of any fourth or higher order isograms, although Borgmann (1984b) notes such instances as ?*Zzzz*, the common comic strip representation of the sound of snoring, the exclamation '*Ay! Ay! Ay! Ay!*' and the possible coinage **dedeeded* (from attested *deeded*)—none of these are of course serious contenders for attested English language 4-isograms.[6]

Methodology

Data

Sources

The data sources, which the remainder of this study is based on, were obtained in the form of word lists from two sources. All the files labelled *a*–*z* from the 1-grams, Version 20120701, were downloaded from Google Ngram in Gzip compressed format and stored in a single directory; the numeric, *other*, *pos* and *punctuation* files were excluded.[7] For the BNC, the word frequency list made available by Adam Kilgarriff *all.al.gz*, also Gzip compressed, was obtained via the University of Brighton's public FTP server.[8]

Preparation

Both data sources were pre-processed by a Python script to produce a uniform word list in a single file for each source.

For the Ngrams, the script read all files from a specified directory in alphabetical order. Each file was then read line by line. On encountering the first headword in the file, a variable of the type *list* was instantiated with fields keeping a tidied version of the headword (described in the next subsection), the original untidied version of the headword (including the Part of Speech [POS] tag), the token count and the volume count. For each further line, if the headword was the same as before, the values for token and volume count were added and the next line read. As soon as a new headword was encountered, the current list was written to an output file with each field separated by a single tab stop before starting the same process of summing up token and volume counts for the next headword. Headwords which after tidying did not consist solely of alphabetical characters were discarded. The process was repeated for all files in the directory and the resulting word list was written into a single uncompressed text file. If and only if totals were provided in the source data, an additional file with the same name as the output file plus the suffix "`.totals`" was written which included the total token and volume counts for the file; this data was later used to calculate normalised frequencies and volume counts. From the script made available (see end of Methodology section), this function can be accessed by running the command

(6) `isograms --ngrams --indir=INDIR --outfile=OUTFILE`

where `INDIR` specifies the directory containing the Gzip compressed 1-gram files from Google Ngram and `OUTFILE` the path the resultant word list should be written to.

The BNC word list was pre-processed in a similar fashion. The script read the Gzip compressed word list line by line and combined lines where the tidied-up version of the headword was identical into a single entry, with the token and volume counts summed for all items with an identical tidied headword. The output was again written to a single file with a tidied headword first, followed by the original untidied headword supplemented by the POS tag of the first headword (i.e. in the format `word_POS`), the token count and the volume count. In addition to excluding strings where the tidied version had non-alphabetic characters, this script also excluded headwords where the untidied version contained one of the characters '`&`','`_`','`%`','`/`' or '`:`'. Such exclusion was necessary because the source list appeared to have a number of issues with sources including XML entities, which were not properly parsed and this produced tokens such as '`&aacute;ngel`', many instances of which were also broken (e.g. missing a closing semicolon) and so could not be reasonably resolved. For the BNC a "`.totals`"-file was always written, since the totals data is inherent to the source word frequency list used. The BNC preparation function can be accessed by running the command

(7) `isograms --bnc --infile=INFILE --outfile=OUTFILE`

where `INFILE` is the path to the Gzip compressed *all.al.gz* file, and `OUTFILE` the path the to the file to write the output to.

String Tidying

As described in the previous subsection, headwords in the word list were tidied during pre-processing. The function used for this (named `tidyString()` in the script) first stripped any characters including and following the first underscore in the string, which among other

things is used in the data sources to attach POS tags. The function then employed Unicode normalization to produce the Normalization Form KD (NFKD).[9] The NFKD provides a uniform mapping of all special characters, ligatures and diacritics decomposed as far as possible. This normalised form was then encoded into ASCII with Python's built-in `str()` function, with the error directive set to `ignore`, meaning that characters not in the ASCII character map were effectively stripped, resulting in a string which only retains the bare ASCII characters, minus any diacritics and with all ligatures decomposed. The string was subsequently lower-cased and all non-alphanumeric characters were stripped. The resultant all-lowercase alpha-numeric only string is the canonical representation which was employed as the tidied headword in the pre-processed word lists and which was taken as input for determining isogramy in the script mining these word lists for isograms described below.

Procedure

Extraction of Isograms

After preparing uniform word lists for the Ngrams and BNC as described above, a script was run on each of these word lists to extract those entries which are isograms from them. To achieve this, the script again read each file line by line and passed each headword to an evaluation function `isogram(candidate)`, which returns the order of isogramy of a `candidate` string. Three counters are kept for the total number of isograms, tautonyms and palindromes, which are evaluated and counted for each headword before entering the routine described below. The counters together with totals from the `.totals`-file (if available) is then written to a new additional file also named with the same filename as the output file with "`.totals`" appended.

The algorithm by which `isogram()` evaluates a candidate is as follows: first, an empty variable of the type dictionary (`d`) is constructed. Then the candidate string is read letter by letter. For each letter, if the letter is not in `d`, a new dictionary entry is made for it and assigned the integer value `1` (one). However, if the dictionary already contains an entry for the given letter, the integer value of the entry is incremented by `1` instead. After the end of the candidate string is reached, a new variable of the type integer to determine the order of isogramy (`n`) is instantiated and set to the value of the first entry of `d`. A loop evaluates every entry of `d`, and if at any point the value of any entry is different from that in `n`, the function returns `0` (zero) as it is clear that the candidate cannot be an isogram. If no entry of `d` has any different value from `n`, however, then each letter must occur `n` times and so the function returns the value of `n`.

If the value for a headword is equal to `0` (i.e. it is not an isogram), the next line is evaluated immediately. However, if `isogram()` returns a value greater than `0`, then the headword is additionally passed to two functions, `isPalindrome(candidate)` and `isTautonym(candidate)`, both of which return a Boolean value of `true`, if the candidate string is a palindrome or tautonym respectively, and `false` otherwise. `isPalindrome()` simply reverses a copy of the string and compares it to its original, while `isTautonym()` slices a string in half and compares whether both halves are identical. A new line is then written to the output file, which contains the fields in (8), each separated by a tap stop—given here together with the label I will henceforth adopt for them:

(8)

Order	*Label*	*Description*
1.	isogramy	Order of isogramy
2.	length	String length
3.	word	Tidied headword
4.	source_pos	Original, untidied headword + POS
5.	count	Token count (combined)
6.	vol_count	Volume count (combined)
7.	count_per_million	Normalised token count per million
8.	vol_count_as_percent	Norm. volume count per hundred
9.	is_palindrome	Palindromy (1 or 0)
10.	is_tautonym	Tautonymy (1 or 0)

After writing the line to the output file, the next line from the input file is evaluated. (Note that fields 7 and 8, the normalised count and volume count are calculated with the total number of tokens and volumes from the original source data. If this information is not available in the original source, the fields are filled with zeros instead.)

The functionality by which this extraction process was carried out can be accessed from the published script as follows:

(9)
```
isograms --batch --infile=INFILE --outfile=OUTFILE
```

where INFILE is the path to one of the word lists which has been prepared by the method described above, and OUTFILE the path the to the file to write the output to. To evaluate a single string for isogramy, the script can be called as follows:

(10)
```
isograms -i STRING
```

where STRING is any candidate string and the script will print the order to isogramy to the console output, or 0 (zero) if the string is not an isogram.

Storage in Database

For convenience of handling and manipulating the large dataset which resulted from running the script extracting isograms on the BNC and Ngrams word lists, both datasets were imported into an SQLite3 database.

Each list of extracted isograms was imported into a separate table, named ngrams and bnc respectively with column labels as detailed in (8) above. The .totals files for both of these were imported as bnc_totals and ngrams_totals, respectively. Since accessing data across both datasets can take a considerable amount of time, five further tables were then constructed from the data contained in the bnc and ngrams tables to facilitate faster queries for specific kinds of subsets of the data. The first is a table which combines all the entries from the bnc and ngrams tables by simple union, i.e. appending one list to the other. This table is named combined. Following this, since all three of the bnc, ngrams and combined tables feature a number of repeated headwords, compacted versions of these tables were created, in which entries are grouped by the word column, thus

generating tables in which each tidied headword has exactly one entry. These tables have been named `bnc_compacted`, `ngrams_compacted` and `combined_compacted`, respectively. Lastly, a table was created to give an intersection between the isogram lists from Ngrams and the BNC. This was achieved by selecting only those entries from the `bnc_compacted` and `ngrams_compacted` tables where the headword (i.e. the `word` column) is identical. This table is named `intersected`. Note that for all three compacted tables and the intersected table the values of both `count` and `vol_count` columns in the new table are the sum of all the original entries for the headword, and the values of the both `count_per_million` and `vol_count_as_percent` columns are summed for the compacted Ngrams and compacted BNC but averaged (not summed) for both the combined compacted and intersected dataset. This means that token and volume counts in these tables are only representative in relation to the data in their own tables, as there is likely to be significant overlap between original sources, which cannot be traced and excluded with this design, and the normalised counts per million and as percent may in some cases sum up to exceed their denominator.

The database itself (named `isograms.db`), together with the two isogram lists generated by the script (named `ngrams-isograms.csv` and `bnc-isograms.csv` respectively), are made available publicly (see Breit 2017).

Tools

Computing Environment

All work was conducted on a 64-bit workstation running Microsoft Windows 7 Enterprise.

Data Processing

All data processing was conducted using a single script written in Python (Python Software Foundation 2011, Version 3.2.5), as described above. The script (named `isograms.py`) is made available and can be used to reproduce the final dataset from the original frequency data by following the procedure described above.

Data Storage

Results from the `isograms.py` script were stored in simple CSV files. These were then imported and collated into a SQLite (Hipp, Kennedy & Mistachkin 2010, Version 3.7.3) database, as described above. A SQL script (named `create-database.sql`), which can be run in the SQLite console application to reconstruct the database from the script's CSV output, is also made available.

Statistical Analysis

Statistical analysis was conducted in R (R Core Team 2013, Version 3.0.2). The packages DBI and RSQLite were used to access the datasets directly from the SQLite database. An R script (`statistics.r`) is provided, which can be used to reconstruct all the tables, numbers, statistical tests and figures reported in the Results and Discussion section below.

Public Data Repository

All the scripts and data (except source data) referred to above are made freely available from a public repository, see Breit (2017).

Results and Discussion

Type Totals

A combined total of 5,176,456 isograms were extracted from both the BNC and Ngrams lists together; 5,064,274 from Ngrams and 112,182 from the BNC. After combination and compacting, the total combined number of isograms extracted was 1,160,507, with 92,849 isograms occurring in both the Ngrams and BNC results.

Noise

From visual inspection of the resulting dataset it is readily apparent that noise is a big problem. For the Ngrams data this is especially apparent in the form of long sequences of one or more characters, such as a sequence of 50 *A*'s or 5 *A*'s followed by 5 *B*'s and so forth. Together with long sequences of *L*'s and similar items it becomes readily apparent that these trace back to problems with the OCR system employed by Google Books, the source of the Ngrams. To a lesser extent, letter sequences constituting items such as Roman numerals, sequences of nucleotides, examples such as 'ababab' for things such as grammars and automata and the like pervade the data, which of course are not noise in the sense that they represent actual text which has been correctly scanned by the OCR system employed for Google Books. The BNC data seems to be generally of much better quality with few notable issues, such as those with Ngrams, or otherwise unwanted items. As may perhaps be expected, the table containing an intersection of the BNC and Ngrams results appears to have the least problems with noise overall. While many of the issues of one or two letters repeating in Ngrams could possibly be excluded or resolved by a script, no such filtering was implemented as this was judged to be beyond the goals and immediate scope of the present study.

Systematic study of the problems of noise that appear in this dataset, ways to reduce noise, and especially comparison to the level of noise in the underlying corpora overall would however be an interesting way to extend this study in the future.

Isogram Distribution by Length

Before turning to a discussion of the specific examples of 'long' isograms in the data, let us give some consideration to the relationship between isogramy and length. In the opening parts of this paper it was discussed that there is quite a natural assumption that with increasing string length isograms become increasingly rare (just as word length generally is inversely related to frequency). In addition to this, given that the combinatorial possibilities of a limited alphabet are more constrained the shorter a string is, we might also expect to find a drop in the number of isograms of a very short length. But what exactly is the margin where we find the greatest fraction of English isograms?

Dataset	Min	Max	Median	Mean	SD
Ngrams	1	81	6	6.29	1.71
BNC	1	28	6	6.11	1.86
Combined	1	81	6	6.30	1.72
Intersected	1	19	6	5.97	1.82

Table 1: Median and mean length of isograms in the different datasets.

The longest isograms in the different datasets (all noise) are 81 letters for Ngrams, 28 letters for the BNC and 19 letters for the intersected dataset. The first examples, which are actual English words, are 15 letters in all cases. All the examples above 45 letters are repetitions of a single letter and above 16 letters almost all examples are either repetitions of a few letters or repetitions of a single word or interjection, for instance *hahahahahahaha* or *thankyouthankyouthankyouthankyou*—though a notable set of isograms is also constituted by Roman numerals, examples such as 'ababab' from mathematical or computing related texts and variations of the entire alphabet written as a single string (we find the entire alphabet forward, reversed, and various parts and shifted patterns of the alphabet). From manual browsing of the data, it appears that the longest real single word examples are at 16 letters, all of which are exclusively found in the Ngrams dataset, and all of which are of German origin, e.g. *Dialektforschung* 'dialect research', *Kampfschilderung* 'fight narration', *schwerpunktmäßig* 'pertaining to the main focus', and *standardisierten* 'standardised (ADJ.PL)'. In all three datasets (Ngrams, BNC, and intersected) the longest English examples are 15 letters and these include most of the already familiar examples discussed above as well as a number of new examples, discussed presently. As can be seen from **Table 1**, while there is great variability in maximal isogram length, both the median and mean length of isograms are very stable, with the centre of gravity falling around 6 letters in all of the datasets.

Figure 1 illustrates the distribution of isograms in the compacted BNC and Ngrams datasets, irrespective of order of isogramy, in relation to string length. The length distribution of the combined compacted and the intersected datasets are nearly identical to those of the compacted Ngrams and BNC in **Figure 1** respectively. We see that this approaches normal distribution between 1 and 13 letters length, with a peak between 5 and 8 letters length. This is true of both the BNC and Ngrams data, although we can also see that especially in this region, where the bulk of isograms are to be found, the sheer size difference of the corpora makes a major difference in the actual number of isograms found, while below 3 and above 11 letters, their coverage is very similar. As we would expect from the stable centre of gravity, a two-sample Kolmogorov-Smirnov Test ($D=0.65$, $p<0.001$) confirms that both samples come from the same distribution, making it likely that the size difference seen in **Figure 2** can be attributed to the effect of corpus size.

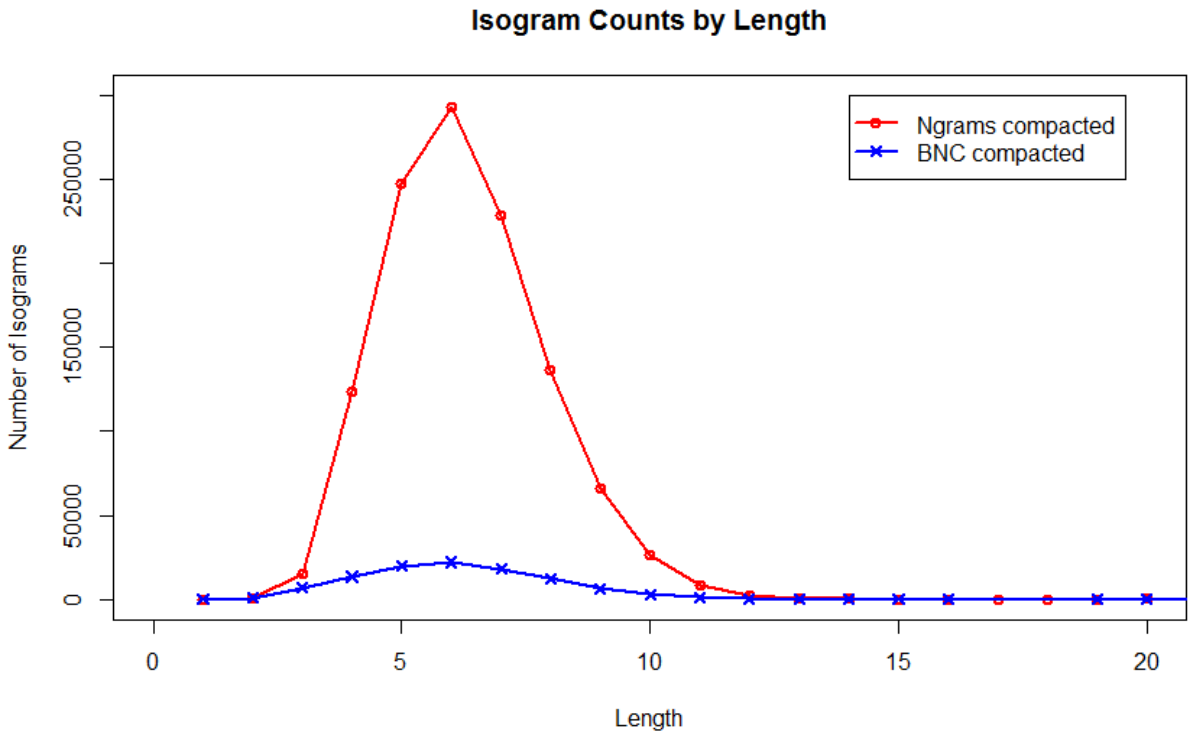

Figure 1: Isogram counts by length and by corpus.

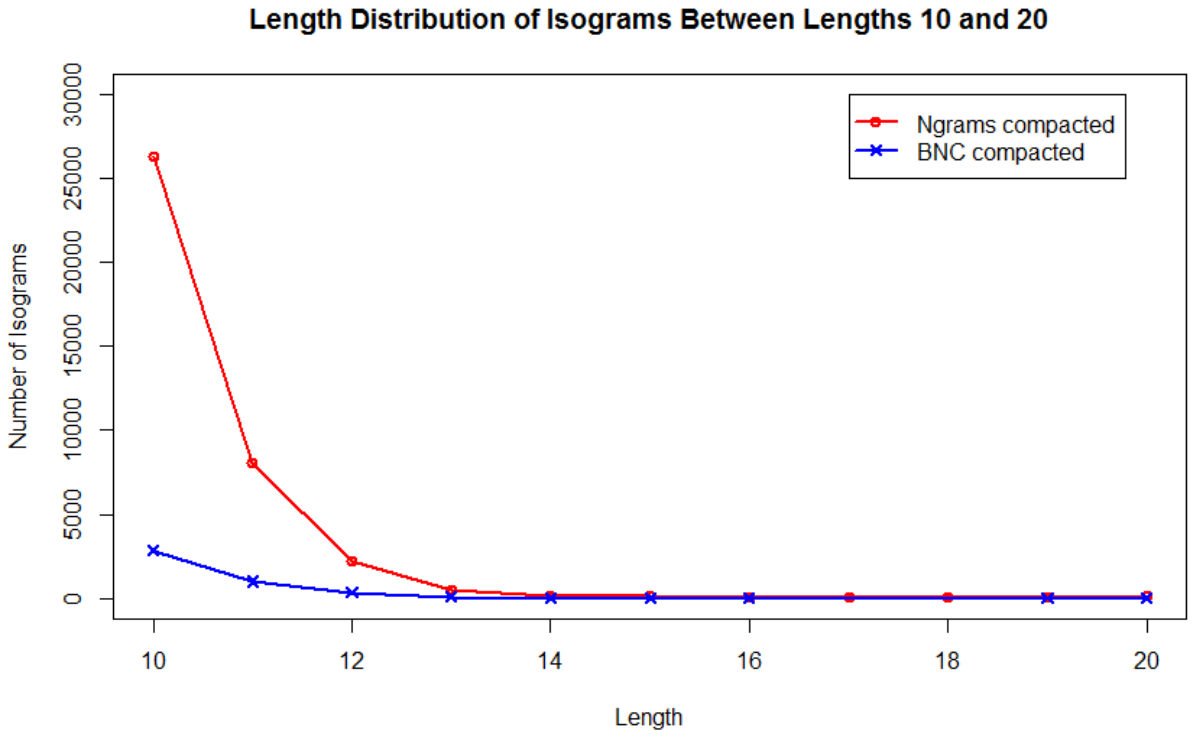

Figure 2: Distribution of isograms by length above 10 letters.

Borgmann (1974) initially gauged that the particular level at which 1-isograms become interesting because of their length is 15 letters. **Figure 2** above illustrates the number of isograms by length for both the maximal dataset (the combination of both BNC and Ngrams) and the minimal dataset (the intersection of the two). It shows that Borgmann's 15 letter limit does indeed coincide quite well with the area in which isograms become increasingly rare. This pattern also clearly reflects the well-documented general trend that the number of distinct words in any language decreases with overall word length (Rothschild 1986, Smith 2012). For instance Smith (2012) documents a similar decline in overall English word length distribution beginning at around length 12 with words of 17 or more letters length being increasingly rare, only marginally behind the decline seen in the isogram data here. Indeed, the sharpest drops can be found at above 10 letters, and examples with 14 or 15 letters are both similarly rare, with 176 and 121 examples in the combined compacted set respectively. In the combined compacted data we find 52 examples at 16 letters, and then 30 or fewer for 17, 18 and 19 letters length. In the intersected dataset we fall below 10 examples at 14 letters length, and from 16 letters on find only single examples.

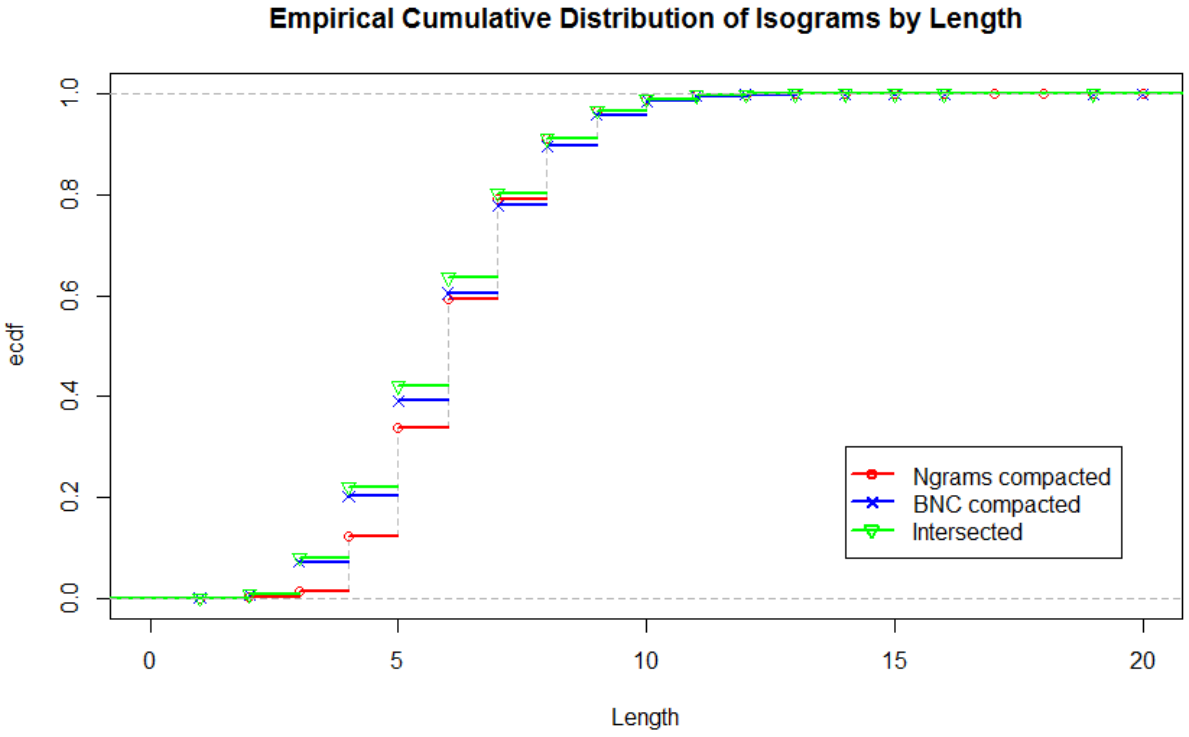

Figure 3: The empirical cumulative distribution function of length for the compacted Ngrams, BNC and intersected datasets.

On a mainly subjective level, as applied in much of the previous literature on isograms, one could take these facts to indicate that examples of length 14 or above should be of particular interest by virtue of the rarity of examples of their length. However, it would be useful to find a more objective criterion for what such a subjective level of rarity actually refers to. A starting point here might be to ask what percentile of isograms is delimited by a

previously assumed boundary such as Borgmann's 15 letters. We can then use a percentile around that mark as a measure for rarity to be applied,which we can also compare to other common percentiles at the right edge of a distribution, such as the 95th and 99th. **Figure 3** shows a plot of the empirical cumulative distribution function (ecdf) of length for the compacted Ngrams, BNC and intersected datasets. The ecdf of length for the combined compacted dataset matches that of the compacted Ngrams dataset so closely (although it is not identical) that it would not be visible as a distinct item on the graph and therefore has been left out. The ecdf shows clearly that, in any of the datasets, even examples of 10 letters length contribute only very marginally to the body of samples. The ecdf of length at 15 letters (rounded to the first significant digit) is 0.9995 for the compacted Ngrams and combined compacted, 0.9999 for the compacted BNC and 1.0 for the intersected dataset. **Table 2** shows the 95th, 99th and 99.95th percentile of the length distribution of isograms in all four compacted datasets. Both the 95th and the 99th percentile are much lower than what the previous literature has regarded as the level of interest, but the 99.95th percentile, based on looking at the ecdf of length at 15 letters length for the compacted Ngrams dataset above, shows a level from 13 letters in the compacted BNC and intersected datasets and 16 in the larger (and also noisier) compacted Ngrams and combined compacted datasets. This suggests that the 99.95th percentile could be a good measure to objectively assess the level of rarity previously approached more subjectively in the logological literature.

Dataset	95th Percentile	99th Percentile	99.95th Percentile
Ngrams	9	11	16
BNC	9	11	13
Combined	9	11	16
Intersected	9	11	13

Table 2: 95th, 99th and 99.95th percentile of the length distribution of isograms.

A similar question to that of long isograms arises in connection with brevity: do we exhaustively know all English isograms of lengths 1, 2 and 3? The combinatorial possibilities of a 26-letter alphabet give us 26,650 and 15,600 possible 1-isograms for length 1, 2 and 3 respectively. While we find all 26 single letters both in the combined and intersected datasets, the possibilities at 2 letters are only exhausted in the Ngrams and combined data while the BNC and intersected datasets only have 628. At 3 letter length, the combined data has nearly all possibilities, at 15,372, and the intersected dataset just over half of the possibilities, with 6,768 3-letter 1-isograms; although of course the vast majority of these 1-3 letter isograms even in the intersected set are not what we would call a solid word.[10]

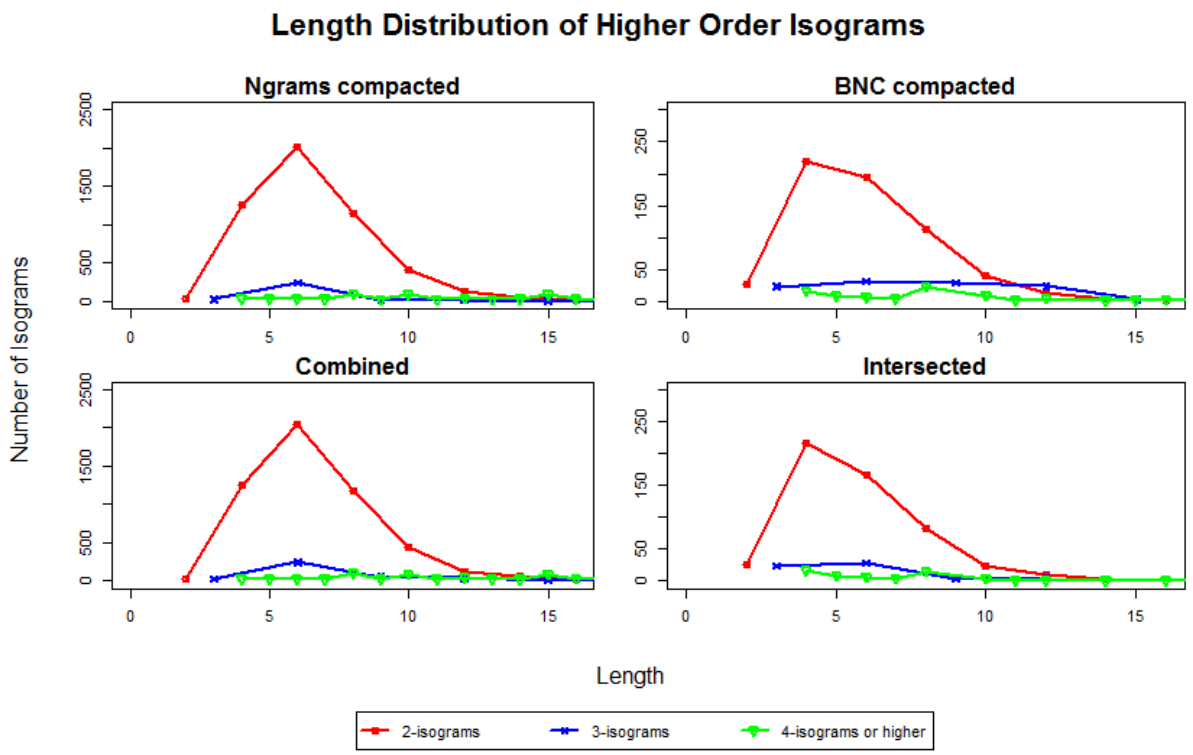

Figure 4: Distribution of 2-, 3- and 4+-isograms by length in the compacted Ngrams, compacted BNC, combined compacted and intersected dataset.

As regards higher order isograms, Borgmann (1974) initially set the level of interest at 10 letters or longer. We may usefully also ask how many isograms of a particular length there are not just in general and for 1-isograms, but also specifically for 2-, 3-, 4-isograms, and so on. As can be seen from **Figure 4** above, 2-isograms show the most pronounced relationship between their distribution and length, and indeed the 10-letter mark here coincides with a marked drop in the number of isograms below a certain level. In the case of the large combined compacted dataset, this is from 1178 instances of 8-letter 2-isograms, to 432 instances at 10 letters. At 12 letters this drops further to only 126 examples and then 55 at 14 letters and only 22 of 16 or more. For the intersected dataset, the overall high-point is at only 216 examples (for 4-letter 2-isograms), much below the number of examples at 10 letters in the combined dataset. Examples drop to double-digits from 8 letters and to single digits for 12 and 14 letters and there are no longer examples. **Table 3.1** shows summary statistics for isograms of different orders in the intersected dataset (which was chosen because it is the least noisy). Their distributions again show a very stable mean and median around length 6 (cf. **Table 1**), but notably standard deviation increases as a function of isogramy and the mean for isograms of order 4 and higher is larger than for any other order of isograms. This difference may be due to the fact that noise in the data (e.g. repetitions of the single letter 'a') proportionally increases with order of isogramy, making this category more susceptible to such noise than the others, even in the intersected dataset.

Isogramy	Min	Max	Median	Mean	SD
1-isograms	1	15	6	5.97	1.81
2-isograms	2	14	6	5.61	2.12
3-isograms	3	12	6	5.38	2.50
4-isograms or higher	4	19	6	6.90	3.10

Table 3.1: Median, mean and standard deviation for the length of different order isograms in the intersected dataset.

Isogramy	95th Percentile	99th Percentile	99.95th Percentile
1-isograms	9	11	13
2-isograms	10	12	14
3-isograms	12	12	12
4-isograms or higher	12.3	17.29	18.91

Table 3.2: 95th, 99th and 99.95th percentile of the length distribution of different orders of isograms in the intersected dataset.

Table 3.2 applies the same percentiles as before to isograms of different orders in the intersected dataset. This reveals that the previously proposed level of 10 letters does not correspond to the same percentile for higher order isograms as appears from investigating the distribution of 1-isograms and from all isograms irrespective of isogramy. Somewhat surprisingly perhaps, this measure would suggest that higher order isograms become rare much later than lower order isograms (note especially the outlier values of isograms of order 4 and above). Again this skew may be partially due to noise in these categories, but is likely in part also due to the fact that the distributions of isograms above length 2 are much flatter than those of lower orders (cf. **Figure 4**). These measures might be improved if the data is manually reassessed to eliminate noise from the higher order isogram data, or if measured against distributions of previously known and published higher order isograms, although this is beyond the scope of what is presented here.

Both 3-isograms and isograms of the orders 4+ are generally exceptionally rare in the data, even without excluding much of the noise such as *thankyouthankyouthankyou* or interjections such as *tatatata* (possibly a dental click, also written '*tsk*' or '*tut*'). In the intersected dataset, there are no more than 28 examples at any particular length for isograms of order 3 or above. However, in the combined compacted dataset, which contains much more noise, there is a marked rise of the number of examples of isograms of orders 4 or higher with letter lengths of 16 or more, as is apparent from the slight repeated raises in all but the bottom right chart of **Figure 4** above—this increase is largely due to the aforementioned repletion of single or two letter sequences, mainly brought in by the Ngrams dataset. Due to its marked exceptionality and particular adherence to the much noisier non-co-indexed dataset, the occurrence of items in these regions might be usefully exploited to

serve as a metric of noise. Their characterisation as noise is further corroborated by visual inspection of these examples, which confirms that none of them appear to be anything approaching the likeness of an actually attested (or even attestable, cf. the distinction made by Hale and Reiss 2008) English word.

Isograms by Frequency

The last section discussed the distribution of isograms by their length, which directly relates to two of the main questions arising from previous studies of isogramy, i.e. (i) how many isograms of length *x* are there in the data?, and (ii) which are the longest isograms in English? However, the nature of the corpus data raises two related but different questions, namely (i) which are the most frequent isograms in English? and (ii) what is the frequency distribution of English isograms?

As already noted in the methodology section, each token in the word lists has two fields related to frequency: token count (`count`); and volume count (`vol_count`)—what Adelman, Brown & Quesada (2006) term *contextual diversity*. Token count refers to the total number of times that a given token (the isogram in this case) occurs in the entire corpus. Volume count on the other hand refers to the total number of individual sources in which the token occurs. For example, if a word *abc* is found in three books of the corpus and is mentioned five times in each of these books, it would have a token count of 15 and a volume count of 3. In looking at the frequency data, two issues arise. The first problem is one of comparability of corpus size. Because the corpus which Ngrams is based on (i.e. Google Books) is much larger than that of the BNC, we must rely on a normalised measure instead of absolute counts. For this reason the following discussion of frequencies are based on frequency per million words for token count and cumulative contextual diversity per hundred volumes for volume count. The second problem is to do with overlap. This issue most markedly affects the intersected and combined datasets, where both token counts have been added up and volume counts have been averaged across, but the problem is amplified with each level of compacting including the initial data preparation as described, because from the counts we cannot reconstruct whether two homographs that are later joined to a single entry have overlap in the volumes they occur in. This indifferentiability means that the normalised volume count is frequently above 100%, and that this measure is still susceptible to corpus size to a certain degree. Consequentially the normalised counts in these datasets must be expected to be somewhat skewed toward exaggerating the frequency of high frequency items and the size of the gaps in frequency can only cautiously be accepted to represent any order of magnitude.

The ten most frequent isograms by normalised token count are *the*, *of*, *and*, *to*, *in*, *a*, *is*, *for*, *it*, and *as* (all over 12,000/million) in the compacted Ngrams and *the*, *of*, *and*, *to*, *a*, *in*, *it*, *is*, *was*, and *I* (all over 9,000/million) in the compacted BNC dataset. The ten most frequent isograms by normalised volume count are *I*, *in*, *a*, *no*, *one*, *on*, *s* (presumably from the contracted *'s* in forms such as *it's*, *he's*, and so forth), *so*, *to*, and *d* (presumably from the contracted *'d* forms, e.g. in *he'd* for *he would* etc.; all over 650/hundred) in the compacted Ngrams and *as*, *s*, *right*, *over*, *set*, *to*, *used*, *like*, *left*, and *put* (all over 290/hundred) in the compacted BNC dataset.[11] As to the least frequent items, it is notable that there are no isograms in the Ngrams and intersected dataset which are hapax legomena (i.e. items that occur only once in a corpus), while there are 38,055 isograms in the compacted BNC and 7,973 isograms in the combined compacted datasets which are hapax legomena. There are

45,079 isograms which occur in only a single volume for the BNC, while there are no isograms in the Ngrams dataset which have a volume count of only one. Inspection of the 10 least frequent isograms by normalised token and volume count yields mainly noise, with very few possibly marginal forms, such as ?*bodywashing* (which, from browsing Google Books appears to sometimes be written as one word in reference to the act of body washing in nursing) and ?*boghten* (found for instance in Chaucer's Legend of Good Women, line 211); both of these are in the intersected dataset.

Dataset	Min	Max	Median	Mean	SD
Ngrams	0.0001	113338.42	0.0009	0.95	151.34
BNC	0.01	61814.86	0.03	6.52	280.67
Combined	0.0001	87576.64	0.001	0.77	116.86
Intersected	0.005	87576.64	0.06	9.51	413.06

Table 4.1: Summary statistics for normalised token frequency of isograms (rounded to either two digits or the first significant digit).

Dataset	Min	Max	Median	Mean	SD
Ngrams	0.0009	898.46	0.006	0.56	8.77
BNC	0.0242	373.91	0.05	1.60	9.87
Combined	0.0009	583.40	0.006	0.36	5.75
Intersected	0.0126	583.40	0.16	4.20	19.92

Table 4.2: Summary statistics for normalised volume frequency of isograms (rounded to either two digits or the first significant digit).

Table 4.1 and **Table 4.2** give summary statistics for the normalised token and volume frequency data across all categories of isograms. The table shows that while most isograms exhibit a relatively low token frequency, the data are highly variable, more so for token frequency than contextual diversity. The fact that mean and median normalised volume frequencies are relatively high compared to mean and median normalised token frequencies suggests that even low frequency isograms show great contextual diversity (rather than for instance all being high repetition items in very few volumes).

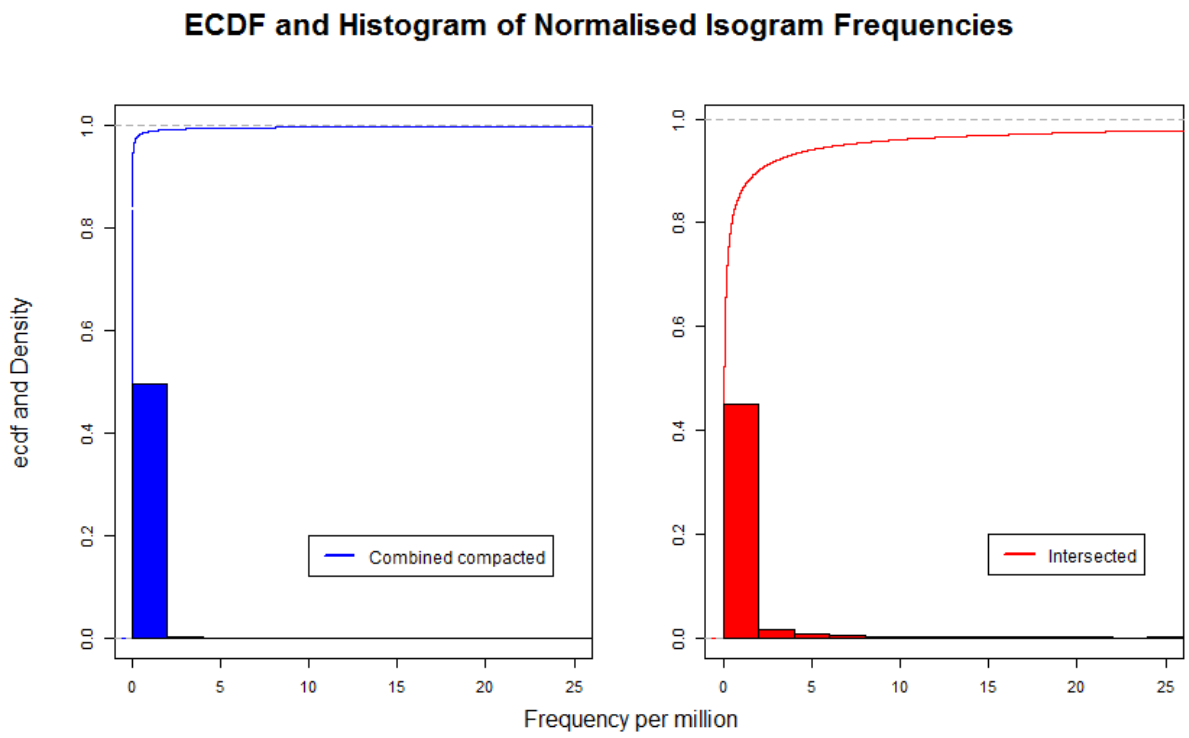

Figure 5: Frequency density (histogram) and empirical cumulative distribution function (line) of normalised token frequencies of isograms in the combined compacted and intersected datasets. The x-axis is cut off at 25, although the maximal frequency is just over 87576/million (cf. Table 4.1).

In **Figure 5**, we see that the number of isograms with any given token frequency decreases rapidly with increasing token frequency. It is apparent from the histogram that the distribution is extremely right skewed, i.e. there are relatively more low frequency isograms than there are high frequency isograms. As indicated by the empirical cumulative density function values in **Figure 5**, isograms with a frequency close to zero make up the vast majority of the two datasets, though this trend is less extreme for the intersected dataset than for the combined compacted.

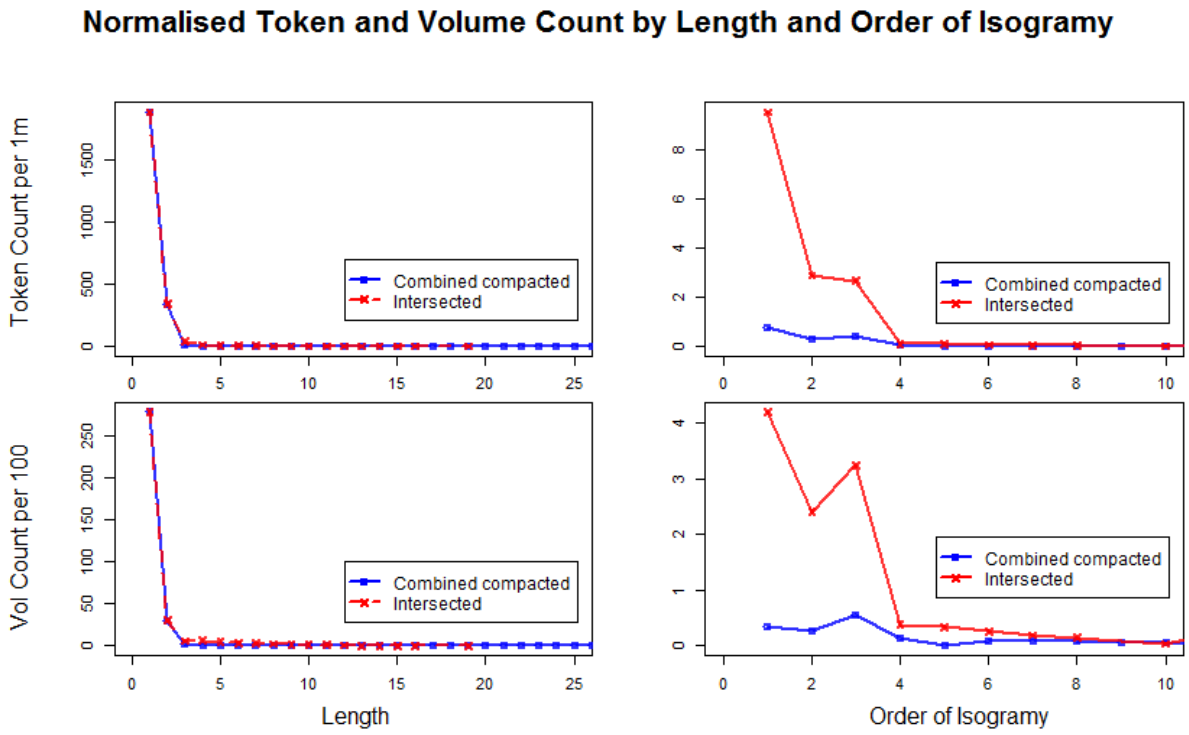

Figure 6: Mean of normalised token frequency per million grouped by length (top left) and order of isogramy (top right), and mean of normalised volume count per hundred grouped by length (bottom left) and order of isogramy (bottom right).

The association of isogram distribution with length and order of isogramy discussed in the previous section may lead us to expect that we find a similar association between token count and length, and token count and order of isogramy. Indeed, as can be seen from the four charts in **Figure 6**, shorter isograms tend to be both much more frequent and contextually diverse than long isograms and isograms of higher orders tend to be less frequent and less contextually diverse than isograms of lower orders, although it must again be noted that most of the higher order isograms represent noise in the underlying corpora. Nonetheless, there is a significant drop from the mean token count of 1-isograms to that of 2- and 3-isograms in both datasets; perhaps indicative that generally speaking 2- and 3-isograms can be expected to be less frequent based on their order of isogramy. A tie-corrected non-parametrical Spearman's correlation of isogramy and length moreover shows that there is a significant negative correlation between the length and order of isogramy (r^2=-0.018, $p<0.001$), meaning that higher order isograms (which are lower frequency) are also generally longer (a factor also associated with lower frequency).

Isogramy, Palindromy and Tautonymy

In the introduction I advanced some arguments in favor of studying the overlap between palindromy and tautonymy with isogramy, rather than simply disregarding isograms which fit one of those categories. While I will further remark on particular isograms which are palindromes and tautonyms in the next section, let us here briefly consider this question: how many of the isograms in the data are also palindromes, tautonyms, or both?

Dataset	Isograms	Tautonyms	Palindromes	Both
Ngrams	1,149,631	3,005	1,420	360
BNC	103,725	438	235	59

Table 5.1: Summary of isograms which are also palindromes, tautonyms or both in the compacted Ngrams and BNC datasets.

Dataset	2-Isograms	Tautonyms	Palindromes	Both
Ngrams	5,060	2,591	724	26
BNC	612	389	137	26

Table 5.2: 2-isograms which are also palindromes, tautonyms or both.

Dataset	3+-Isograms	Tautonyms	Palindromes	Both
Ngrams	1,080	414	644	334
BNC	197	49	72	33

Table 5.3: Isograms of order 3+ which are also palindromes, tautonyms or both.

Dataset	Isograms	Tautonyms	Palindromes
Ngrams	5,064,274	3,020	5,096,894
BNC	112,182	183	112,781

Table 5.4: Total count of all isograms, tautonyms and palindromes (the latter two irrespective of whether they are isograms or not) in the Ngrams and BNC dataset.

Table 5.1 shows that only a very small proportion of the isograms in the dataset are tautonyms or palindromes and even fewer isograms are both a tautonym and a palindrome at the same time, underlining the suggestion in the introduction that isograms which are also tautonyms and/or palindromes are actually interesting in their own right based on the fact that they are increasingly rare in the dataset. While it is the case that over half of the 2-isograms in both corpora are tautonyms, even for 2-isograms, the number of palindromes and tautonymous palindromes is exceedingly rare (cf. **Table 5.2**). As can be seen from **Table 5.3**, for isograms of order 3 or above, both tautonymy and palindromy are common, though restricted to less than half of cases. Comparison of the data for individual orders of isograms to the overall summaries for tautonymy and palindromy in **Table 5.4** show that tautonymy is rather rare overall and indeed appears to entail a significant trend toward isogramy, while the number of total palindromes is comparable to the number of isograms in the data, suggesting that the two may be more independent from one another, although both of these trends would justify further investigation.

New Isograms

After discussing the various patterns of distribution, frequency, tautonymy and palindromy in the data, one important question that remains is whether the study has turned up any hitherto unnoticed isograms and, if so, which these are. In this section I will give a brief list of such isograms, which have been manually taken and cross-checked from the data. I adopt a practice similar to Gooch (1998) and indicate palindromes with a *p* and tautonyms with a *t*. The letters N and B represent presence in the Ngrams and BNC respectively. Isograms are given first by order, then by length. Forms with a prefixed superscript question mark are attested but still marginal or questionable.

1-isograms

The following lists give a number of 1-isograms of 15 and 14 letters length, which have thus far not been reported in the relevant literature.[13] There are a number of interesting foreign language examples which I have for reasons of space omitted here; but I do give some such examples for higher order isograms.[14]

15 Letters

(11)	*uniformly-spaced*	N		

14 Letters

(12)	*black-uniformed*	N	B	(Not a headword in OED, but contained in a quotation under *fascist salute*)[12]
	capsule-forming	N		
	counterdisplay	N		
	cytomegalvirus	N		Apparent alternate spelling of *cytomegalovirus*
	double-tracking	N		
	flame-producing	N		
	flesh-producing	N		
	?*formula-weights*	N		Chemical unit, very marginally hyphenated
	heavy-producing	N		
	hydromagnetics	N		(in OED)
	hyperabducting	N		
	hyperabduction	N		(in OED)
	latex-producing	N		
	leftward-moving	N		
	low-disturbance	N		
	metal-producing	N		
	neurolymphatic	N		
	outspreadingly	N		
	problem-causing	N	B	
	pseudochivalry	N		
	pseudomythical	N		(in Webster's New International Dictionary)
	semibankruptcy	N		
	shame-producing	N		

slate-producing		N	
slave-producing		N	
stackunderflow		N	Computing term, opposite of *stackoverflow*
steam-producing		N	
sweat-producing		N	
symbol-creating		N	
undiscoverably		N	(in Webster's New International Dictionary)
unphagocytized		N	From biology, not yet engulfed by a phagocyte
vesiculography		N	Medicine, the imaging of seminal vesicles
victory-flushed		N	
waste-producing		N	
wheat-producing		N	

2-isograms

The following lists give a number of hitherto unreported 2-isograms. While I have included some foreign language examples found in the data, I have limited listing these items to the longest example of each language.

16 Letters

(13)	*standardisierten*		N	German, 'standardised (ADJ.PL)'

14 Letters

(14)	*ammortizzatori*		N	Italian, 'shock absorbers'
	aristocráticos		N	Spanish, 'aristocratic (ADJ.PL)'
	benzhydroxamic		N	
	concomitantiam		N	Latin, 'concomitant/accompanying'
	economic-minded		N	
	Schizotrypanum		N	(In Webster's Medical Dictionary)

12 Letters

(15)	*appareillier*		N	Old French, 'to make ready'
	ensimmäisenä		N	Finnish, 'the first (ESSIVE)' as in *John I.*
	metasomatose		N	Synonym of *metasomatism*[15]
	palaeoslopes		N	Geology, directionality of dip of former surface (Singular form in OED)
	transeastern		N	Appears in names, e.g. *Texas TransEastern*

10 Letters

(16)	*Aphrophora*		N	A genus of froghoppers
	fangufangu	*t*	N	A type of Tongan flute
	Kharkharee		N	Variant spelling of *Kharkhari*, a city in India

nagbubunga	N		Tagalog, 'flowering plant'
non-ordered	N		
Nyaminyami	*t*	N	A God living in the Zambezi River
palaeopole	N		(Listed under *paleo-* in OED)
Peddapalle	N		V. of *Peddapalli*, a place in Telangana, India
polypyrrol	N		Variant spelling of *polypyrrole*
Pungapunga	*t*	N	A river in New Zealand
Rengarenga	*t*	N	A type of lilly
rewharewha	*t*	N	Maori, 'influenza'
Roccagorga	N		A municipality in Latina, Italy
shabu-shabu	*t*	N	Japanese dish of boiled beef
Tingatinga	*t*	N	A style of painting from Tanzania
Wellawatte	N		A neighbourhood of Colombo, Sri Lanka

3-isograms

There are only two 3-isograms in the corpus, both of 9 letters length:

(17)	*chachacha*	N	B	(already reported by Gooch 1998)
	naanan	N		Ojibe, 'five'; also a given name
	shshsh	N	B	Iconic interjection, prolonged *sh* (in OED)

4-isograms

There are two 4-isograms in the data, neither of which have been reported as such before. Both are tautonyms and *poop-poop* is also a palindrome:

(18)	*Nangganangga*		*t*	N	Fijian spirit who guards heaven from bachelors
	poop-poop	*p*	*t*	N	B

More Palindromous Isograms

There are only three solid isograms which are both tautonyms and palindromes in the data. These are: *peep-peep*, *poop-poop*, and *toot-toot*.

If isograms which are palindromes but not tautonyms are considered, the number is larger, but still surprisingly small. The ones that can be found in the intersected dataset are: *abba* (Semitic 'father'), *Gwallawg* (given name of 6th century king of Elmet, *Gwallawg mab Llaenawg*), *degged*, *Hannah*, *mallam* (honorific title given to Islamic scholars), *Notton* (village in West Yorkshire, England), *pull-up*, *redder*, *succus* (a fluid secreted by living tissue), *amma* (Hindu 'mother'), *Anna*, *beeb* (nickname for the *BBC*), *boob*, *deed*, *dood* (OED has this as an old form of *to do* and variant of *dude*), *ebbe* (an old spelling variant of *ebb*), *ma'am*, *naan*, *noon*, *Otto* (given name), *peep*, *poop*, *sees*, and *toot*.

Conclusion

This paper presented a novel approach to the search for and study of English language isograms by computational means. In addition to making available a set of tools, which can

be applied to other corpora both of English and other languages, the paper presented the first quantification of isogram distribution across the English language.

The previously intuitively gauged levels of 15 letters for 1-isograms and 10 letters for 2-isograms were shown to indeed coincide with a marked rise in the rarity of isograms of that length. It was shown that the most frequent isograms in English are all short function words and that lower word frequency is also inversely correlated to the density of isograms of that particular frequency. Tautonymy and palindromy were shown not to be a major contributor to isograms overall, but they become more significant the higher the order of isogramy. Lastly, the study produced a significant number of hitherto unnoticed attested isograms; among these are two fourth-order isograms, although both are tautonyms.

Directions for future research include the expansion of the set of corpora which are studied in this way, the application to other languages, and more refined quantitative analysis, for instance pitching isogram distribution directly against other metrics of the English orthographic system and word distribution. With regards to the significance which tautonymy and palindromy play in isogramy, the methods can be adapted to extract a similar list of non-isogram palindromes and tautonyms. The comparative study of two such lists could reveal to which degree tautonymy and palindromy are actually statistically predictive of isogramy. Beyond the bounds of 'logology proper', if coupled with automatic phonemic transcription, the method offers the basic ingredients for a comparative study of isogramy and isophony (words in which each phoneme occurs the same number of times); this may reveal interesting patterns across languages with regard to the closeness of fit a language's orthography provides to phonemes in the language. Lastly, as it has been shown that a sizeable proportion of long-length high-order isograms are noise, future work could usefully investigate in how far isogramy provides a measure for more general noise in a corpus.

Competing Interests
The author declares that they have no competing interests.

Acknowledgments
I am grateful to the two reviewers, Darryl Francis and one remaining anonymous, for very extensive and useful comments. I also thank Nick Neasom, Andrew Nevins, Kureyon Shinchan and Marlies Gabriele Prinzl who have similarly provided useful comments and feedback.

Notes
[1] That is to say, they are due to other factors such as a cross-linguistic dispreference for light monosyllabic content words (cf. e.g. Hayes 1995), and in the case of *mama* and *papa* other constraints of acquisition (cf. Jakobson 1960).
[2] Cadence is the recurrence of letters at a regular interval, e.g. *b* in the string *abcbdbebf*; cf. Eckler (1983).
[3] Note that I mark unattested words with an asterisk (*) and those with a questionable status with a superscript question mark ([?]).

[4] For more on long isogrammic place names, see Tilque (1996).
[5] Though as Darryl Francis (pc) rightly points out, from a linguistic point of view, these forms are often comparable to other compounds written as a single word, e.g. *blackbird*, and their plurals should thus be accepted as valid examples of 2-isograms, and whether they are written with a space, hyphen or as a single word is often only based on their age and convention. The problem for an automated analysis such as the one presented here is that (i) frequency lists don't generally include such orthographically loose compounds as types but rather record their constituents individually, and (ii) there is no established way to distinguish them from other types of noun-noun sequences based on orthography alone. The same problem applies to other types of compounds. Compare the adjective-noun compound *blackboards* with the adjective-noun sequence *black boards*. Identifying these types of compounds in an automated analysis thus presents an interesting problem for further study.
[6] However, Darryl Francis (pc) notes that there are a number of other notable interjections and iconic forms: The OED notes as a variant of *mm* the form *mmmm* as well as the fact that the interjection is occasionally written with five or more m's, and Cassidy and le Page's (1967) Dictionary of Jamaican English has iconic fourth order *kyou-kyou-kyou-kyou* for the sound of destruction and the seventh order interjection *bububububububu* for the sound of flight, though they of course result from reduplication. Note though that neither *kyou-kyou-kyou-kyou* nor *bububububububu* are attested in the combined dataset from Google Ngrams or the BNC.
[7] The Google Ngram 1-gram files are available from http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.
[8] Adam Kilgarriff's BNC frequency lists are available from ftp://ftp.itri.bton.ac.uk/bnc/.
[9] See http://unicode.org/reports/tr15/ for more information about Unicode normalization.
[10] The high co-incidence of non-word isograms at low word length and the relative ease with which such a small subset of the data can be computationally checked against a list of known short 1-isograms would potentially lend itself to investigating the type of noise in large corpora sometimes associated with a more recent increase in (mis-)parses of data, special and foreign characters, etc; cf. e.g. the motivation for 'the-normalization' in Bentley et al. (2012), and see Acerbi (2013) for commentary.
[11] Nick Neasom (pc) points out that the frequency effects here mimic an interesting dichotomy between function and content words.
[12] Thanks to Darryl Francis for pointing out that some of these are found in the Oxford Dictionary of English and Webster's New International Dictionary (2nd Edition), as indicated in brackets (non-comprehensive), as well as pointing out the two isograms *benzhydroxamic*, *Schizotrypanum* and *shshsh,* which are in the data but were missing from an earlier draft. He also points out a number of other isograms, which can be found via Google Search, but which are neither in Ngrams nor the BNC. These are: *amphistrongyle*, *ancylotheriums*, *dichlorbutanes*, *dimethylfurans*, *ethylcoumarins*, *hydrocalumites*, *lycanthropised*, *lycanthropized*, *lycanthropizes*, mynpachtbriefs, *Oldbury-Smethwick*, *phytalbuminose*, *sulphogermanic*, *trichlamydeous*, *troublemakings*, *ultrasymphonic*, *uncompahgrites*, *unfarsightedly*, and *unstylographic*; as well as the unhyphenated phrases and loose compounds (cf. fn. 5) *backing yourself*, *blacking powders*, *blasting powder*, *breakdown lights*, *Buckingham's revolt*, *buckthorn family*, *folding brackets*, *Judgment of Paris*, *McKnight Boulevard*, *thick-warbled song*, *tumbledown shack*, and *white gyrfalcons*.

[13] It is notable that many of these new isograms are hyphenated compound forms. While such forms are not uncommon in the established literature on isograms, their proportion here is clearly much larger than the proportion of hyphenated forms reported in previous articles. Note also that, by definition, all 1-isograms that are compounds do in-turn consist of non-overlapping 1-isogram constituents (the same is not true for higher-order isograms).
[14] One may wonder why there are numerous foreign language examples in English-language corpora. From manual inspection of some of these items in Google Books and the Ngrams viewers, it appears that while a minority of these are original-language citations and non-assimilated loanwords or phrases, such as *per concomitantiam*, a large proportion of these, especially in German and French, seem to be due to mixed language books and journals and volumes erroneously classified as English.
[15] See for instance http://www.bgs.ac.uk/scmr/docs/papers/paper_9.pdf and the Wikipedia entry *metasomatism*.

References

Acerbi, A 2013 *Normalization Biases in Google Ngrams*, 14 April 2013. Available at http://acerbialberto.wordpress.com/2013/04/14/normalisation-biases-in-google-ngram/ [Last accessed 26 October 2015].
Adelman, J S, Brown G D A & Quesada, J F 2006 Contextual Diversity, Not Word Frequency, Determines Word-naming and Lexical Decision Time. *Psychological Science*, 17(9), pp. 814-823. DOI: http://dx.doi.org/10.1111/j.1467-9280.2006.01787.x
Altmann, G 1988 *Wiederholungen in Texten*. Bochum: Brockmeyer.
Bentley, R, Garnett P, O'Brien M & Brock W 2012 Word Diffusion and Climate Science. *PLoS ONE*, 7(11): e47966. DOI: http://dx.doi.org/10.1371/journal.pone.0047966
Borgmann, D 1965 *Language on Vacation: An Olio of Orthographical Oddities*. New York, NY: Scribner.
Borgmann, D 1974 An Overview of Isograms, *Word Ways*, 7(1): pp. 33-36. Available at http://digitalcommons.butler.edu/wordways/vol7/iss1/10
Borgmann, D 1985a Long Isograms (Part 1), *Word Ways*, 18(2): pp. 67-75. Available at http://digitalcommons.butler.edu/wordways/vol18/iss2/2
Borgmann, D 1985b Long Isograms (Part 2, *Word Ways*, 18(3): pp. 140-147. Available at http://digitalcommons.butler.edu/wordways/vol18/iss3/5
Borgmann, D 1985c Long Isograms (Part 2), *Word Ways*, 18(4): pp. 243-246. Available at http://digitalcommons.butler.edu/wordways/vol18/iss4/13
Breit, F 2017 Data and tools for studying isograms. *figshare*. DOI: http://dx.doi.org/10.6084/m9.figshare.5245810.v1
Cassidy, F & le Page, R 1967 [1980] *Dictionary of Jamaican English*. Cambridge: Cambridge University Press
Crystal, D 2007a What's so special about Bricklehampton? The Guardian, 19 May. Available at http://www.theguardian.com/books/2007/may/19/featuresreviews.guardianreview3
Crystal, D 2007b *By Hook or by Crook: A Journey in Search of English*. New York, NY: Overlook
Eckler, R (ed.) 1971 A Fourteen-Letter Pair Isogram, *Word Ways*, 4(3): p. 136. Available at http://digitalcommons.butler.edu/wordways/vol4/iss3/3

Eckler, R 1975 The Twenty-One Words, *Word Ways*, 8(4): pp. 250-254. Available at http://digitalcommons.butler.edu/wordways/vol8/iss4/17
Eckler, R 1983 The Mathematics of Words, *Word Ways*, 16(4): pp. 252-254. Available at http://digitalcommons.butler.edu/wordways/vol16/iss4/16
Eckler, R 1997 *Making the Alphabet Dance: Recreational Wordplay*. New York, NY: St Martin's Press.
Francis, D 2012 New Pair Isograms, *Word Ways*, 45(1), pp. 64-69. Available at http://digitalcommons.butler.edu/wordways/vol45/iss1/17
Gooch, R 1998 Isograms: The Sequel, *Word Ways*, 31(1), pp. 68-70. Available at http://digitalcommons.butler.edu/wordways/vol31/iss1/17
Goldsmith, L 1990 Chaos: To See a World in a Grain of Sand and a Heaven in a Wild Flower. *Archives of Dermatology*, 126(9): pp. 1159-1160. DOI: http://dx.doi.org/10.1001/archderm.1990.01670330039003
Grant, J 1982 Pair and Trio Isograms, *Word Ways*, 15(3): pp. 136-139. Available at http://digitalcommons.butler.edu/wordways/vol15/iss3/3
Grant, J 2012 Long Pair Isograms, *Word Ways*, 45(3). Available at http://digitalcommons.butler.edu/wordways/vol45/iss3/16
Hale, M & Reiss, C 2008 *The Phonological Enterprise*. Oxford: Oxford University Press
Hayes, B 1995 *Metrical Stress Theory: Principles and Case Studies*. Chicago, IL: University of Chicago Press.
Hipp, D R, Kennedy D & Mistachkin J 2010 SQLite, Version 3.7.3. Available at http://www.sqlite.org
Jakobson, R 1960 Why 'mama' and 'papa'? In: Kapaln, B & Wapner, S (eds.) *Perspectives in Psychological Theory: Essays in Honor of Heinz Werner*. New York: International Universities Press, pp. 124-134.
Python Software Foundation 2011 Python, Version 3.2.5. Available at http://www.python.org
R Core Team 2013 R: A language and environment for statistical computing, Version 3.0.2. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.r-project.org
Rothschild 1986 The distribution of English dictionary word lengths. *Journal of Statistical Planning and Inference*, 14(2-3), pp. 311-322. DOI: http://dx.doi.org/10.1016/0378-3758(86)90169-2
Shillock, R C, Kelly, M L & Monaghan, P 1998 Processing of palindromes in neglect dyslexia, *NeuroReport*, 9(13), pp. 3081-3083.
Smith, R 2012 Distinct word length frequencies: distributions and symbol entropies, *Glottometrics*, 23, pp. 7-22. Available at http://arxiv.org/abs/1207.2334v2
Tilque, D 1996 Placename Isogramy, *Word Ways*, 29(4): pp. 211-214. Available at http://digitalcommons.butler.edu/wordways/vol29/iss4/2
Wimmer, G & Altmann, G 1996 The theory of word length: some results and generalizations, *Glottometrika* 15, pp. 112-133.
Wolpow, E 1991 Subdermatoglyphic: A New Isogram, *Word Ways*, 24(1): p. 18. Available at http://digitalcommons.butler.edu/wordways/vol24/iss1/4