UNIVERSITY OF TARTU
DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE

ANTWORDPROFILER ANALYSIS OF THE NOVELS OF KURT VONNEGUT, JR., AS SET IN OPPOSITION TO THE GSL, THE AWL, AND THE BNC/COCA WORD-FAMILY LISTS

MA thesis

EDMUND ALEXANDER DALTON
SUPERVISOR: ASSOC. PROF. REET SOOL

TARTU
2014

ABSTRACT

This thesis was conceived in response to an article titled “How large a vocabulary is needed for reading and listening?” by Paul Nation, in which a selection of written and spoken texts are analyzed with a vocabulary-profiling program called Range. In his article, certain texts that are thereby posited to typify discrete categories, such as the genre of the novel, are measured against the frequency-based lists of 14,000 word families, along with an additional list of proper nouns, compiled from the British National Corpus. The current study, however, takes a revised approach that is encapsulated in a 2013 journal paper titled “Mid-frequency readers”, by Nation and Laurence Anthony. They apply AntWordProfiler for computer-assisted analysis of educational literature by the lists of 25,000 word families, as well as those of marginal words, that have been made from the British National Corpus in conjunction with the Corpus of Contemporary American English. Moreover, their work includes a concise definition of the concepts of high-, mid-, and low-frequency vocabulary. Thus, the aim of this thesis is to adopt and optimize an established methodology applicable to statistical analysis of the active vocabulary of Kurt Vonnegut, Jr. and its evolution over a 45-year period that encompasses his career as a novelist. The purpose of the current study is to determine what vocabulary size is necessary to attain sufficient coverage of the novels of Vonnegut, whether the existing word-family lists are arranged in an order that represents a logical progression of word-frequency levels, and how appreciable is the overall effect of disparate forms of lexical elements in the diametrically opposed corpora.

The introductory chapter to this thesis integrates a contextualization of all relevant options and selections with a literature review of fundamental secondary sources prior to the formulation of a methodological point of departure. The chapter in question constitutes the first theoretical part of the thesis, comprising delineations of concepts, definitions, and postulates that are deemed internally consistent. Additionally, it provides a description of the aims of the two core chapters, “Details pertaining to the corpora” and “Sample group analysis and control group validation”, of their component sections, plus of the appendices.

The first core chapter, “Details pertaining to the corpora”, is the second theoretical part of this thesis. Its opening paragraph describes in outline its symmetrical subdivision into two sections: “Overview of the control group” and “Overview of the sample group”. Each section comprises four paragraphs, which particularize the distinguishing features of the corpora, address the problem of information gaps, consider the applicability of inchoate assessment criteria, and summarize the categories that are factored into the analysis proper.

“Sample group analysis and control group validation” is the second core chapter and the empirical part of this thesis. As in the preceding chapter, the opening exposition establishes the four-paragraph structure of its sections. Their titles and arrangement are an enumeration of the novels of Vonnegut in chronological order of original publication. The component paragraphs of each of these sections discuss a stage of the analysis: first, the computation of the proportion of general and academic words; second, the comparison of a specific volume with the 25,000 word families and four kinds of marginal words; third, the identification of any frequency-related anomalies and the computation of integral running words in percentage terms; fourth, the summation of the underlying trends of the tabulated results, which concludes the analysis of a text and the discussion thereof.

TABLE OF CONTENTS

INTRODUCTION

In an article entitled “How large a vocabulary is needed for reading and listening?”, Paul Nation (2006: 62) describes a computer-assisted analysis performed in the context of a corpus-based study that undertook to answer the following research question: “How big a vocabulary do you need to get adequate coverage of various kinds of texts?” In actuality, according to Nation (2006: 61), his study has two avowed aims, namely “to trial word-family lists recently developed from data from the British National Corpus (BNC)” and “to use these lists to see what vocabulary size may be needed to reach a 98% coverage level of a variety of written and spoken texts”. The same article covers a representative sample of novels in that two sections of the text are devoted to the discussion of the genre, focalizing a juxtaposition of the lexical profiles of “*Lord Jim* by Joseph Conrad, *Lady Chatterley's Lover* by D. H. Lawrence, *The Turn of the Screw* by Henry James, *The Great Gatsby* by F. Scott Fitzgerald, and *Tono-Bungay* by H. G. Wells” (Nation 2006: 70). The relevance of the aforenamed range of volumes to the current study lies in the fact of their analysis forming the rationale for the averment that an “8,000–9,000-word-family vocabulary is needed for dealing with written text” (Nation 2006: 79).

In recent times, however, Nation collaborated with Laurence Anthony (2013: 5) in developing “a new free extensive reading resource for learning the mid-frequency words of English and for reading well known texts with minor vocabulary adaptation”. Even though their report on the systematization of a strategy for lexical simplification, entitled “Mid-frequency readers”, builds on the conclusion of the antecedent article, certain aspects of the former approach were modified as part of this joint effort. The key difference seems to be that Anthony’s (2013) AntWordProfiler has supplanted Alex Heatley et al.’s (2004; 2012a; 2012b) original Range program. For the purposes of this thesis, the basic functionality of

versions 1.32 and 1.32H of Range, both of which are downloadable as freeware from the official project page as of the time of writing, was tested, indicating at least one irreducible incompatibility between profiling algorithms: Range treats hyphenated forms, inclusive of compounds, as indivisible units. Furthermore, the 25,000 word families and four varieties of marginal words that have been compiled into reference lists for coherent research "on the basis of frequency information from the BNC combined with that from the Corpus of Contemporary American English (COCA)" (Nation and Anthony 2013: 8) have replaced the older control group of 14,000 word families plus a single list of marginal words (i.e., proper nouns) from solely the BNC. The creator of COCA, Mark Davies (2012b), declares that the two corpora "complement each other nicely", and yet his comparison stresses that "COCA is much larger and more recent, which has important implications for the quantity and quality of the data overall".

It should be noted early on that the entirety of this thesis is the final product of an exercise in constrained writing. This term was coined to describe a conscious submission of one's text to precise formal and, by extension, thematic "boundaries that explicitly limit the possible realizations of a text in some respects" (de Geest and Goris 2010: 82). Such a technique, Dirk de Geest and An Goris (2010: 82) elaborate, typically governs "creative stimuli for the artistic process" via overriding determinants, and it is also exploited here to enhance the discursive progression. Therefore, although this text strives for accessibility to as wide a range of readers as possible, primacy is given to formal academic rhetoric, with the main topic imposing its share of technical terms. Typographically, the paragraphs factor in line length, whereas the abundance of compound–complex sentences and the scarcity of recurrent explanations stem from the limitation on the number of pages; even so, a prior reading of the primary sources is not a precondition for following the discussion. What this

thesis sets out to do is to guide the reader through the exploration of the complete surface text of one's personal copies of these novels. To this end, mixed numerals, which express exact values, are supplemented with percentages, which are approximations in that they are rounded down to two decimal places. The latter, then, are expected to be less opaque to the reader. All the statistical computations revolve around fractions to prevent round-off errors.

"Mid-frequency readers" serves, first and foremost, as an exemplification of how notable works in the public domain that have been introduced into the Project Gutenberg collection can be properly adapted for intermediate learners of English as a foreign or second language. It is done in an effort to bridge the gap that is perceived to exist between graded readers intended for beginners, which have been heretofore made available, and lexically complex originals, which tend to be fully accessible only to native speakers and advanced learners. For that reason, Nation and Anthony (2013: 7–10) define three cardinal frequencies of vocabulary items in terms of the equivalent word-family levels: (a) high-frequency vocabulary, comprising the first 3,000 most wide-ranging, general-purpose word families, which have the distinction of being essential; (b) the next 6,000 comparatively wide-ranging, general-purpose word families, constituting the mid-frequency vocabulary; (c) the 16,000 "more narrowly focused" low-frequency word families, thus far successfully classified as such, including "some technical vocabulary unique to a particular discipline". Equally germane to this thesis is the theorists' explanation of the three difficulty levels of mid-frequency readers, aimed at learners who know 4,000, 6,000, or 8,000 word families; hence, up to two subsequent word-family levels contain "target word families to support" (Nation and Anthony 2013: 7), while every item of more infrequent occurrence is replaced.

In essence, the principal elements of the theoretical and methodological framework of the empirical analysis of e-books in this thesis rests on a conceptual hybridization of the

approaches illustrated by the theorists in question in the 2006 and the 2013 study, with the aspects specified above having been superseded by the changes implemented in the latter piece. In contrast, the data yielded by these procedures should be effectively original, given that preparatory background research did not produce any evidence of past studies whereby the linguistic composition of a literary work of Kurt Vonnegut, Jr. (1922–2007), let alone an entire genre of his works, was processed with a vocabulary profiling software. Coming back, for the remainder of this paragraph, to the subject of the underlying theory, Nation (2006: 61) claims the following in view of a typical novel: “98% text coverage (1 unknown word in 50) would be needed for most learners to gain adequate comprehension”. This idea derives from Hsueh-chao Marcella Hu and Nation’s (2000: 422) conclusion that “learners need to have around 98% coverage of the words in the text to be able to read for pleasure”, which, in turn, accords with the findings of Ronald P. Carver (1994: 435) in that “students are likely to increase their vocabulary by reading relatively hard materials because 2% or more of the words are likely to be unknown”. Latterly, the requisite percentage has been also discussed by Norbert Schmitt, Xiangying Jiang, and William Grabe (2011: 40), whose “study suggests that readers should control 98%–99% of a text’s vocabulary to be able to read independently for comprehension”.

The fundamental reason for electing to conduct the current study completely within the parameters in the introductory chapter thus set is that the thesis undertakes to refute the validity of neither the confirmed notion of “98% as the ideal coverage” (Nation 2006: 79) nor the resultant estimation of the correlative vocabulary size. Instead, the objective of the ensuing quantitative analysis is to adhere to the methods employed by the aforementioned researchers in order to ensure both the internal consistency and the generalizability of the outcome, with respect to cognate analyses in particular. Additionally, both literary criticism

and applied linguistics are outside the scope of the current study, albeit the descriptive statistics presented herein should be of prime interest and potential benefit to teachers and learners of contemporary literature and/or English as a foreign language. The findings are expected to make explicit the size and complexity of the active vocabulary of Vonnegut, whose works are still protected by copyright, as well as to reveal the extent to which the BNC/COCA lists would have to be updated so as to chart, as meticulously as possible, all relevant developments throughout his career as a novelist, a period of 45 years. Implicit in the analysis is the suitability of his novels for adaptation. Taking account of the foregoing, the research question is threefold: (a) How many word families do you need to know to be familiar with most words in the novels of Vonnegut; (b) are the BNC/COCA word-family lists properly sequenced to reflect consecutive word-frequency levels in the case of each volume; (c) how substantial is the margin of error resulting from discrepancies between the novelist's active English vocabulary and the unmodified contents of the BNC/COCA lists?

For the processing of an independent corpus of volumes, the current study consists in the use of version 1.4.0w of the AntWordProfiler program and its default lists by Nation (2012b; 2013). To elucidate, the analyst elects to use this profiler to compare each primary source with (a) Michael West's General Service List (GSL) of nearly 2,000 word families, which, according to David Hirsh, can cover "around 90% of the running words in fiction" (Nation and Kyongho 1995: 35), in conjunction with (b) the Academic Word List (AWL) of 570 families by Averil Coxhead (2000: 225), which "accounts for approximately 1.4% of the tokens in /…/ fiction texts". All the comparisons lead up to the application of (c) the ready-made lists from both the BNC, which contains "90% written, 10% orthographically transcribed spoken text" (Burnard 2009a), and COCA, which "is evenly divided between the five genres of spoken, fiction, popular magazines, newspapers, and academic journals"

(Davies 2012a). The literary dimension integrates with this corpus-linguistic one to finalize an interdisciplinary approach to the former. Its potential benefits to the reader are designed to be its unprecedented, not abstract or esoteric, insights into the complexity of legitimate representations of Vonnegut's active vocabulary, hence an exhaustive guide to the textual surface of idiosyncratic linguistic entities. Therein lies improved access to the context that permits the readerly construction of meaning, requiring one to recognize that "readers can produce a wide range of variant interpretations of a given literary work depending on their experiences both as readers of literature and as members of the larger human community" (Davis and Womack 2002: 61). In that connection, readerly competence retains its validity.

Now, it could be argued that 1953's GSL is "dated" (Browne 2013: 12; Gardner and Davies 2013: 8, 19–20), and criticisms leveled at "the pitfalls of using the AWL as a filter" (Neufeld et al. 2011: 533), such as "a grave risk of making serious teaching and learning omissions" (Hancioğlu et al. 2008: 471–472), continue to proliferate with the emergence of substitutes; notably, by Ali Billuroğlu et al. (n.d.), Charles Browne et al. (2014a; 2014b), and Davies and Dee Gardner (2013). Regardless, the fact remains that, with the extraction of samples from copyright material digitized and published by a third party entailing many variables, a freeware profiler and its default lists can function as constants. In other words, owing to the magnitude of the research effort, the use of, for instance, WordAndPhrase is counterproductive because of the "limitations on use via the web interface" (Davies 2012c) and the lack of identical freely downloadable resources. Neither can the analyst advise that the words missing from the BNC/COCA subgroup "should be added to the families in the existing lists" (Nation 2012a: 5). Finally, any "still not perfect" (Davies 2012d) automatic ascription of preconceived meaning to words has no purpose if "only readerly performance ultimately shapes the nature of meaning" (Davis and Womack 2002: 84). Here, a dictionary

aids a cursory investigation into discrepancies. The keyword *vocabulary* denotes a subject narrower than lexicology and, therefore, should not be misconstrued to imply semasiology.

AntWordProfiler generates vocabulary statistic and frequency information about a corpus of texts, from which the main categories of data to be collected are the total running words (tokens), different word forms (types), and word families in a given text. The third category, then, is to be "defined as a stem plus all closely related affixed forms" (Coxhead 2000: 218) whose "base form must be recognizable as a freely occurring word" (Bauer and Nation 1993: 254). Further, Nation (2006: 63) clarifies that a single family may consist of multiple lemmas and that "a list of lemmas made from the BNC" formed the basis for the original "range, frequency, and dispersion data that were used for the division of the words into lists". It is held in Nation (2012a: 3), as well as in Nation and Anthony (2013: 13), that the lemma is considered a "sensible unit" of lexical knowledge "for productive purposes". Still, on account of receptive skills, the preference for the alternative concerns two reasons:

> research has shown that word families are psychologically real /…/ when reading, knowing one member of the family and having control of the most common and regular word-building processes makes it possible to work out the meaning of previously unmet members of the family. (Nation and Anthony 2013: 13)

The results of the analysis are organized into fourteen corresponding sets of three tables. Each set presents (a) the number of tokens, types, and families in a novel by the GSL and the AWL; (b) the number of tokens, types, and families in a novel by the BNC/COCA lists; (c) a lexicon of anomalous types in a novel that have been omitted from the latter control subgroup, complemented by an appended lexicon of types that are adequately classifiable with the second edition of the *Oxford English Dictionary* (*OED2*). The rationale behind the compilation of a lexicon to differentiate unclassified types is explanatory supplementation of a discussion of cumulative percentage coverage figures for the tokens in the novels by the BNC/COCA subgroup, yielding an ascertainable margin of error.

In addition to particularizing the categories of data that are consequently factored into the analysis proper, the first core chapter, “Details pertaining to the corpora”, identifies two contextually apposite criteria for interpretive approaches: the sexpartite scale of ‘can do’ descriptors aligned to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) and the system of Vonnegut’s self-assessment in *Palm Sunday: An Autobiographical Collage* (1981). The examination of their applicability occurs in, respectively, the two sections of that chapter: “Overview of the control group” and “Overview of the sample group”. The second core chapter of the thesis, “Sample group analysis and control group validation”, reports on the empirical work. The latter chapter is subdivided into sections that concentrate on the lexical composition of a particular novel, enumerated chronologically as follows: *Player Piano* (1952); *The Sirens of Titan* (1959); *Mother Night* (1961); *Cat’s Cradle* (1963); *God Bless You, Mr. Rosewater, or Pearls Before Swine* (1965); *Slaughterhouse-Five, or The Children’s Crusade: A Duty-Dance with Death* (1969); *Breakfast of Champions, or Goodbye Blue Monday* (1973); *Slapstick, or Lonesome No More!* (1976); *Jailbird* (1979); *Deadeye Dick* (1982); *Galápagos: A Novel* (1985); *Bluebeard, the Autobiography of Rabo Karabekian (1916–1988)* (1987); *Hocus Pocus* (1990); *Timequake* (1997). The analysis and presentation of the data involves their incorporation into the accompanying discussion of pertinent findings in the form of table supplements. The tables in the core chapters serve the manifold purpose of (a) exploration, (b) communication, and (c) storage. The tables appended to the thesis, on the other hand, are intended for self-reliant scrutiny, contributing to the creation of a comprehensive frame of reference for the reader. The concluding chapter will complete the comparative part of the statistical analysis with a recapitulation of the tables’ highlights and inferences drawn from the data, followed by perspectives on directions for future research into these corpora.

DETAILS PERTAINING TO THE CORPORA

This chapter resumes the elucidation of the theoretical part. The binary division of the ensuing paragraphs into sections parallels the intrinsic dichotomy between word-family lists and an independent corpus of texts in an analysis. On that account, the contrast should be analogous to the established approach to generating vocabulary profiles. Both of the two sections, “Overview of the control group” and “Overview of the sample group”, are meant to be identical in terms of structural symmetry, each consisting of four paragraphs, which are informed and unified by an overarching theme. First of all, the sections delimit the size of the corpora and the organization of elements within them, taking into consideration the fact that the variables associated with the actual distribution of reference lists and different texts are likely to have adverse effects on the comparability of the tabulated data. Second, problems of information gaps and miscellaneous unresolved issues concerning the integrity of the corpora are addressed in order to account for possible inaccuracies. Third, relatively inchoate assessment criteria that lie beyond the immediate scope of the current study are, nevertheless, exemplified with the aim of contextualizing the adopted approach. The fourth paragraph in each section concludes a thematic overview with a summary of key factors in the analysis proper, the relevance of which transcends the discussion in the current chapter.

Overview of the control group

In the prototypical study, Nation (2006: 65) itemizes the total number of word types in each of the frequency-related lists of slightly over 14,000 word families from the BNC to verify whether “the data confirm the expected pattern of decrease”. That the BNC has not been expanded “after the completion of the project /…/ in 1994” (Burnard 2009b) may not be an issue, for it comes close to the upper chronological boundary of the novels. What

Table 1.1: Paul Nation's GSL/AWL lists

Filename	Modification date	Number of types	Number of families
1_gsl_1st_1000.txt	5 May 2012	4,114	998
2_gsl_2nd_1000.txt	19 April 2012	3,708	988
3_awl_570.txt	5 May 2012	3,082	569
http://www.antlab.sci.waseda.ac.jp/software/wordlists/gsl_awl_cleaned.zip			

Table 1.2: Paul Nation's BNC/COCA lists

Filename	Modification date	Number of types	Number of families
basewrd1.txt	9 April 2013	6,857	1,000
basewrd2.txt	9 April 2013	6,370	1,000
basewrd3.txt	9 April 2013	5,880	1,000
basewrd4.txt	9 April 2013	4,865	1,000
basewrd5.txt	9 April 2013	4,294	1,000
basewrd6.txt	9 April 2013	4,102	1,000
basewrd7.txt	9 April 2013	3,679	1,000
basewrd8.txt	9 April 2013	3,419	1,000
basewrd9.txt	9 April 2013	3,196	1,000
basewrd10.txt	9 April 2013	2,982	1,000
basewrd11.txt	9 April 2013	2,942	1,000
basewrd12.txt	9 April 2013	2,754	1,000
basewrd13.txt	9 April 2013	2,415	1,000
basewrd14.txt	9 April 2013	2,299	1,000
basewrd15.txt	9 April 2013	2,283	1,000
basewrd16.txt	9 April 2013	2,086	1,000
basewrd17.txt	9 April 2013	2,076	1,000
basewrd18.txt	9 April 2013	1,933	1,000
basewrd19.txt	9 April 2013	1,872	1,000
basewrd20.txt	9 April 2013	1,820	1,000
basewrd21.txt	9 April 2013	1,651	1,000
basewrd22.txt	9 April 2013	1,539	1,000
basewrd23.txt	9 April 2013	1,394	1,000
basewrd24.txt	9 April 2013	1,296	1,000
basewrd25.txt	9 April 2013	1,675	1,000
basewrd31.txt	9 April 2013	22,409	21,662
basewrd32.txt	9 April 2013	196	38
basewrd33.txt	9 April 2013	6,044	3,108
basewrd34.txt	9 April 2013	1,149	1,083
http://www.antlab.sci.waseda.ac.jp/software/wordlists/bnc_coca_cleaned_20130410.zip			

is of import is that the tabulated data in the 2006 journal article can, in effect, adumbrate the alterations made to the BNC lists in the intervening years of their development when measured against the contents of the lists of 14,000 word families that are distributed with the first of the two versions of the Range program as of the time of writing; for example, in the available archive file, five of the lists were last timestamped as recently as on 6 June 2012, whereas "exclamations, hesitations, interjections, etc." (Nation 2006: 56) have since been removed from its 3rd 1,000 level. For reasons of clarity and specificity, as well as to facilitate meta-analyses, similar tabulations of the totality of elements per control-group list are presented in this section of the thesis. Table 1.1 and Table 1.2 focus, respectively, on the GSL/AWL and the BNC/COCA subgroup. In the latter, the last four levels, ranked in descending order, represent (a) proper nouns, (b) multifarious marginal items, "including swear words, exclamations, and letters of the alphabet" (Nation 2012a: 1), (c) semantically transparent solid compounds of diverse parts of speech, and (d) abbreviations. These tables should furnish the reader with details sufficient to obtain the analogous files, to replicate the conditions of the analysis, and to approximate the data.

It is evident from the fourth column in Table 1.1 that the families in the GSL/AWL subgroup do not total a round figure. By comparison, for instance, the original variant of the AWL, which is dated 30 August 2001 and distributed with the first available version of Range in a different archive file, lists the hyphenated compound *so-called* among its 570 word families, instead of allowing *so* and *called* to be counted as separate free forms in the GSL. As regards Table 1.2, while the rest of the frequency-based lists display a decrease in the number of word types per 1,000 families along what is virtually an exponential curve, the 25th 1,000-word-family list interrupts this pattern with an acute increase of 379 types. This may be accounted for by the fact that low-frequency items, in consideration of which

"various estimates put the number at somewhere around 100,000 word families" (Nation and Anthony 2013: 9), "continue to be added to the existing families" (Nation 2012a: 3), as opposed to into the "space for additional lists" (Nation 2012a: 1) between the 26th and the 30th 1,000 level. According to Nation (2006: 65; 2012a: 3) and Nation and Anthony (2013: 13), a recommended method for verifying the adequacy of the sequential arrangement of the published lists is to determine whether a downward trend in word frequency correlates with a gradual decline in the number of tokens, types, and families found at each level in an independent corpus. In addition to the issue of rendering hyphenation invalid, the reader should be aware of two critical deficiencies in this system: as Nation (2006: 66; 2012a: 3) acknowledges, neither vocabulary profiler is programed to differentiate homographs, and none of the input files enable the analyst to maintain the integrity of multiword units.

Taken together, such oversimplifications of intricate relationships also exclude the possibility of applying the common reference levels of the CEFR, ranging from A1 to C2, to the process of generating and analyzing vocabulary profiles in a study of this magnitude. This is to say that the CEFR's criteria for assessing language proficiency would necessitate an in-depth analysis of concordances in a comparative study of individual tokens. Certain scales of illustrative descriptors formulated for the CEFR, however, warrant elucidation and incorporation into the reader's frame of reference, given that their system was devised to be "relatable to or translatable into each and every relevant context" (Council of Europe 2011: 21) while being "oriented to the continuum of real world ability" (Council of Europe 2011: 184). The current study concerns exclusively visual reception, namely the activity of reading, which involves, among other things, "the semantic and cognitive understanding of the text as a linguistic entity" (Council of Europe 2011: 184). The conceptual grid offered by the CEFR forms the basis for two searchable resources: the English Vocabulary Profile

(EVP), available at a website operated by Cambridge University Press (n.d.), and the Word Family Framework (WFF), designed for the British Council (n.d.) by Richard West. Both of them use "general English" (British Council: 2012; Cambridge University Press: 2012) and lemmatization to integrate homonymy and polysemy into the descriptive distribution of words, phrases, and phrasal verbs along the horizontal dimension. These resources tend to differ from each other, as well as from the control group; for example, *bank*, constituting a word family at the 1st 1,000 level in the BNC/COCA lists, corresponds to two headwords in the EVP and the WFF; one of the two, defined as 'financial', has lexical items at each of the six levels in the latter resource plus an extra member of indeterminate level, whereas the EVP lacks the compounds *banknote* (B1), *bank balance*, *bank holiday* (both B2), and *merchant bank* (X), the verbs *bank* (B2) and *bankrupt* (C2), and the noun *bankruptcy* (C2), assigning B2 to *banker* and C1 to the adjective *bankrupt* instead of C1 and C2 respectively.

To summarize the main ideas expounded in this section, in the interest of ensuring adequate generalizability of findings, future comparative studies and meta-analyses should endeavor to address the dissimilarities between parallel control groups that comprise lists at different stages of development. Although the total number of word families and family members in a given list can be regarded as the most exhaustive and readily determinable of its distinctive characteristics, such metadata as the date of creation or modification of a file can substantiate probable analogies. Nation's method of presupposing a decremental effect on the quantity of tokens, types, and families found in the independent corpus at successive word-frequency levels of the BNC/COCA constitutes the process of validation in the next chapter of this thesis. AntWordProfiler has been programmed to recognize neither hyphens nor multiword expressions. Classified homo- and heterophonic homographs are, moreover, subsumed under broad word types, thereby reducing the theoretical compatibility between

the adopted system and the CEFR's proficiency scales. Despite this fact, with respect to the core vocabulary of English, the practicability of the EVP and the WFF can be introduced into diverse methodologies of language acquisition and education, inclusive of autonomous adaptation of texts.

Overview of the sample group

All republished editions of the novels of Vonnegut that serve as the primary sources for the current study appear to be in their original American English. Each of the fourteen volumes, which the corresponding sections of the ensuing chapter analyze in chronological order of initial publication, can be considered to comprise not only the main text that is the narrative core of a given work, but also selected portions of the paratext at either periphery. Unless otherwise noted, the front and back matter of a volume add the following elements to the parallel text in the independent corpus by reason of coherence: (a) the title, including the subtitle, if any, and the name of the author as they appear on the title page, for there are instances of an underlying syntactic structure (*Slaughterhouse-Five*, for example, subjoins a compendious autobiography); (b) the dedication, the epigraph (excepting *Hocus Pocus* and *Timequake*), and the preface (excepting *Slaughterhouse-Five*, *Slapstick*, *Jailbird*, and *Galápagos*), for certain proper nouns contained therein also appear in the body text; (c) the introduction to *Mother Night*, the prologue to *Slapstick*, *Jailbird*, and *Timequake*, and the epilogue to *The Sirens of Titan*, *Breakfast of Champions*, *Slapstick*, *Jailbird*, *Deadeye Dick*, and *Timequake*. Multiple occurrences of the author's name accompanying the novel's title and the remainder of other paratextual elements, such as the colophon, have been excluded.

In the list of references at the end of this thesis, the works of Vonnegut furnish both the original year of publication and the year of the cited republication, separated by a slash

in accordance with the style rules of the American Psychological Association. Aside from reflecting the division of the analysis proper into sections, thus forming the structure of the respective chapter, this allusion to the printed editions is designed to preclude the reader from overestimating the fidelity of their electronic versions to the former. On the technical side, for instance, Portable Document Format, into which these novels had been digitized or subsequently converted, fails in distinguishing hard hyphens falling at a line break from soft hyphens, the deletion of which should occur only after the text is reflowed. An adverse concomitant of converting files to be compatible with AntWordProfiler is the concatenation of a separated item ending one line and an uncapitalized item at the beginning of the next. Since these automatic changes have been undone manually, without the codification of a compatible hyphenation algorithm, the reader should be aware that corruptions may have appeared on this account. Furthermore, the electronic editions contained a virtual profusion of typographical errors that had to be rectified in conformity with excerpts and quotations in various publications on Google Books. As to in-text illustrations, all legible characters were transcribed verbatim. Consequently, the constituents of the sample group may bear a closer resemblance to their printed counterparts. Should the reader be able to acquire the primary sources from different publishers and/or in other e-book formats, this profiler will still require the conversion of one's textual input to be UTF-8-encoded Unicode compliant.

In his 1981 collection of shorter works, *Palm Sunday*, Vonnegut evaluates what he deems the upward and downward trend of his success by grading on a scale of A to D the quality of a total of thirteen of his major literary creations between, inclusively, his 1952 debut novel and the aforementioned collage. *Cat's Cradle* and *Slaughterhouse-Five* rank first with a grade of A-plus; conversely, two works in the second half of this contemporary bibliography receive the lowest grade: the 1970 revision of his 1960 play and its 1971 film

adaptation, entitled *Happy Birthday, Wanda June*, and the 1976 novel, entitled *Slapstick*. Additionally, Vonnegut (1981/2009: 284) sets the parameters of the assessment as follows: "The grades I hand out to myself do not place me in literary history. I am comparing myself with myself. Thus can I give myself an A-plus for *Cat's Cradle*, while knowing that there was a writer named William Shakespeare." The fact that Shakespeare is generally recognized for the evidence of lexical innovation and lexical sophistication, to which the etymological component of the entries in the *OED2* attests, is hereby emphasized. In terms of periodization, Arthur Marwick (2012: 8) postulates a 'long sixties' that encompasses a trifurcation into distinctive subperiods – "1958–63, 64–8/9, and 1969–74" – by two pivotal years: the former of these may be associated with "Philip Larkin's declaration that 'sexual intercourse began in 1963'", and the latter one relates to individuals who, "in the manner of George Melly, claim also that the sixties 'ended' in 1968–69". *Jailbird* excepted, the writing and publication of Vonnegut's 'A' material is entirely consistent with Marwick's notion of there having been a cultural and subcultural renaissance. In contrast, the debut novel, which antedates the era, may have been graded B for its adherence to what Marwick (2012: 4) generalizes as "a strict formalism in language", whereas *Breakfast of Champions*, which is graded C, coincides with the end of Eric Hobsbawm's (1995: 8) seminal concept of "the unprecedented and possibly anomalous Golden Age of 1947–73". Therefore, 1973 can be interpreted as the commencement of the history of "a world which lost its bearings and slid into instability and crisis" (Hobsbawm 1995: 403), eclipsing the age that witnessed the culmination of "a general audacity and frankness in books and in the media" (Marwick 2012: 3) guided by "the postmodernist emphasis on language" (Marwick 2012: 13).

In sum, this section undertook, first of all, clarification that, in the current study, the signified denoted by the title of a novel differs in a subtle but important fashion from every

edition of the published volume, whether in electronic or printed format. In particular, the sample-group texts incorporate a miscellany of paratextual elements that are attributable to the author, rather than to the publisher, discarding the tokens in the rest as superfluous. The comprehensive texts of their respective works underwent manual emendation in advance of the analysis with the intention of remedying defects of optical character recognition arising from the digitization process and conversion of computer files. Neither literary theory nor literary criticism comes within the purview of this thesis, which, in turn, obviates the need for the conceivable plethora of subjective readings. However, the preferred reading of the manifest impact of these works by means of an academic grading system can be construed as facilitation of inductive reasoning. For gaining further insight into the alleged zenith of Vonnegut's achievement, between 1963's *Cat's Cradle* and 1969's *Slaughterhouse-Five*, it may be hypothesized that his open dissatisfaction with several of the more recent writings of his intimates to the reader that whatever issues are inherent in their subject matter and its execution are compounded by prosaic diction.

SAMPLE GROUP ANALYSIS AND CONTROL GROUP VALIDATION

This chapter collates empirical data from sets of vocabulary statistics output by the AntWordProfiler program, and this is done in accordance with the theoretical framework set forth in the preceding chapters. The fourteen sections that follow represent the primary sources for the current study in a sequential manner, one novel at a time, arranged in order of first publication date. Each of the sections centers around a series of three tables, which present (a) the figures for tokens, types, and families per GSL/AWL list; (b) the figures for tokens, types, and families per BNC/COCA list; (c) a lexicon of correctly unclassifiable elements. The second and the third table employ gray shading and boldface to highlight, respectively, the discrepant cells of a given array and incorrectly classifiable homographs. What is more, appended to this thesis is a lexicon of verifiable elements hitherto unknown to the BNC/COCA control subgroup, complementing the third table in the foregoing series.

Every table in these sections is accompanied by an expository paragraph, the first kind of which contrasts the quantity of general English tokens, types, and families in a text with that of the categories of academic variety, in addition to contextualizing the range of types and families by ascertaining the coverage of the GSL/AWL subgroup. The second of such paragraphs provides similar information in terms of high-, mid-, and low-frequency elements in the BNC/COCA lists. The third paragraph of a section augments the discussion of the aggregations that arrest or reverse the systematic decrease across the three categories by instancing proper nouns occurring among the predominant forms outside the respective list. Furthermore, underpinned by Nation's (2006: 70) presupposition that "proper nouns can be counted as having a minimal learning burden", as well as by the fact that Nation and Anthony (2013: 7) omit "the coverage of proper nouns and other marginal words" from the computation of the 98% threshold, those levels are subtracted, along with a margin of error.

While the pragmatic and semantic functions of proper nouns tend to be determined by their contexts of introduction and application, other marginal words can be semantically self-contained. In the current study, a margin of error represents all tokens corresponding to the ideally classifiable types in an appended table, in which only the first category requires the respective bound and free morphemes to accord with lemmas and variant spellings in the *OED2*, inclusive of unbound morphemes comprised by multiword expressions therein. Both the BNC/COCA lists and the *OED2* can be evidenced to contain cases of morphemic importation from French, German, and Latin, among others. The *OED2*, however, verifies historical spellings besides alternatives to modern ones, thus permitting the recognition of, for instance, the forms in Vonnegut's (1959/2007: 223) quotation from Geoffrey Chaucer. For the purpose of attenuating gradations of style as a pragmatic means of expediting one's comparative reading of the numerical data interspersed throughout the subsequent sections, the syntactic structure of these fourteen sequences of paragraphs is rendered formulaic. As in the previous chapter, each section concludes with a summation of the principal findings.

Player Piano

Table 2.1 shows that, of the two GSL levels, the first one receives exactly 13 $^{2041}\!/_{6103}$ times as many tokens, 1 $^{932}\!/_{1575}$ times as many types, and 1 $^{128}\!/_{835}$ times as many families, equating to a difference of roughly 92.5, 37.18, and 13.29 percentage points respectively. Taken conjointly, the general vocabulary exceeds the academic kind by 50 $^{161}\!/_{580}$ (98.01%) of the tokens, 5 $^{677}\!/_{681}$ (83.32%) of the types, and 4 $^{266}\!/_{383}$ (78.7%) of the families. According to Table 1.1, this text covers (a) 60.94% of all types and 96.49% of all families at the first GSL level, (b) 42.48% of all types and 84.51% of all families at the second GSL level, and (c) 22.1% of all types and 67.31% of all families at the AWL level.

Table 2.1: Analysis of *Player Piano* by the GSL/AWL lists

File	Tokens	(%)	Types	(%)	Families
1_gsl_1st_1000.txt	81,380	(78.66)	2,507	(25.53)	963
2_gsl_2nd_1000.txt	6,103	(5.90)	1,575	(16.04)	835
3_awl_570.txt	1,740	(1.68)	681	(6.93)	383
	14,241	(13.76)	5,058	(51.50)	
Total	103,464		9,821		2,181

Table 2.2: Analysis of *Player Piano* by the BNC/COCA lists

Level	Tokens	(%)	Types	(%)	Families
1st 1,000	83,024	(80.24)	2,696	(27.45)	979
2nd 1,000	5,853	(5.66)	1,833	(18.66)	886
3rd 1,000	2,434	(2.35)	1,090	(11.10)	674
4th 1,000	1,798	(1.74)	804	(8.19)	547
5th 1,000	1,297	(1.25)	546	(5.56)	432
6th 1,000	674	(0.65)	388	(3.95)	315
7th 1,000	467	(0.45)	307	(3.13)	247
8th 1,000	353	(0.34)	212	(2.16)	182
9th 1,000	341	(0.33)	190	(1.93)	164
10th 1,000	316	(0.31)	129	(1.31)	115
11th 1,000	168	(0.16)	102	(1.04)	92
12th 1,000	119	(0.12)	83	(0.85)	75
13th 1,000	84	(0.08)	68	(0.69)	65
14th 1,000	69	(0.07)	47	(0.48)	45
15th 1,000	46	(0.04)	35	(0.36)	31
16th 1,000	**273**	**(0.26)**	24	(0.24)	23
17th 1,000	166	(0.16)	19	(0.19)	19
18th 1,000	123	(0.12)	**22**	**(0.22)**	**21**
19th 1,000	22	(0.02)	17	(0.17)	16
20th 1,000	17	(0.02)	13	(0.13)	12
21st 1,000	6	(0.01)	5	(0.05)	5
22nd 1,000	**15**	**(0.01)**	**10**	**(0.10)**	**10**
23rd 1,000	**96**	**(0.09)**	8	(0.08)	8
24th 1,000	2	(0.00)	2	(0.02)	2
25th 1,000	**7**	**(0.01)**	**6**	**(0.06)**	**6**
Proper nouns	2,838	(2.74)	292	(2.97)	285
Exclamations	546	(0.53)	56	(0.57)	28
Transparent compounds	473	(0.46)	214	(2.18)	197
Abbreviations	65	(0.06)	25	(0.25)	25
Not in the lists	1,772	(1.71)	578	(5.89)	
Total	103,464		9,821		5,506

Table 2.3: Lexical anomalies in *Player Piano* pursuant to the *OED2*

nibo (11), *brahouna* (8), *sibi* (7), *khabu* (6), *siki* (6), ***akka* (5)**, *sahn* (5), *dibo* (4), *-wunnnn* (4), *athalete* (3), *brahous* (3), *beeby* (2), *beejee* (2), *beezle* (2), *bouna* (2), *brouha* (2), *'cayful* (2), *dinko* (2), *dollahs* (2), *drahve* (2), *foah* (2), *friggin'* (2), *houna* (2), ***-hov* (2)**, *kuppo* (2), *lakki-* (2), ***noozle* (2)**, *nuttin'* (2), *ourrrrrrrrs* (2), *prakhouls* (2), ***raht* (2)**, *reeble* (2), ***shou-* (2)**, *souri* (2), *sumpin'* (2), ***sy-* (2)**, *takki* (2), *theah* (2), *vagga* (2), ***-yuss* (2)**, ***aki***, *allakahi*, *allasan*, *anotha'*, *ashked*, *assu*, *awri*, *bakula*, *batouli*, *billa*, *bloodyin'*, *bonum*, *carryin'*, *chambah*, *crawlin'*, *crossin'*, *dollah*, ***ebo***, *enj-*, *equipmen'*, *evah*, *evenin'*, *facin'*, *faht*, *fahve*, ***figger***, *fightin'*, *-flectah*, *fryin'*, ***-fut***, *gladja*, ***goura***, *harch*, *inspectin'*, *ippi*, ***-itty***, *khabou*, ***koula***, *koze*, *losht*, *manko*, *matoority*, *mismit*, *mortuis*, *moumi*, *nakka*, *-neee*, *ohdnance*, *openin'*, *ouah*, *ov-*, ***pala***, *pillan*, ***pitty***, ***-ple***, *poopin'*, ***poppin'***, ***powah***, *prakka-*, ***puka***, ***puku***, ***qual-***, *quiverin'*, *-reeee*, *sabotagin'*, *sabotoors*, *salet*, ***sakki***, *screamin'*, *-seein'*, *selano*, *sensin'*, *serani*, ***serin***, *shorry*, *shtuff*, *sihn*, ***simi***, ***'smatter***, *softb-*, *souli*, *speakin'*, *'sposal*, *startin'*, *-stin'*, *sutta*, *tahm*, ***-tcha***, *tilla*, *tippo*, *tooie*, *touri*, ***trippin'***, *-veesh-*, *whadja*, *whatch*, ***worl'***, *-wunnn*, *wuth*, *yamu*, ***-yers***, ***-yut***
Total: 142 types (222 tokens)

In Table 2.2, the three high-frequency levels contain 18 $^{2571}/_{4930}$ times (94.6%) more tokens, 2 $^{725}/_{2447}$ times (56.45%) more types, and 1 $^{652}/_{1887}$ times (25.68%) more families than the six mid-frequency levels, while the former ones contain 59 $^{100}/_{139}$ times (98.33%) more tokens, 9 $^{309}/_{590}$ times (89.5%) more types, and 4 $^{359}/_{545}$ times (78.53%) more families than the sixteen low-frequency levels. In contrast, the mid-frequency levels hold 3 $^{343}/_{1529}$ times (68.99%) more tokens, 4 $^{87}/_{590}$ times (75.89%) more types, and 3 $^{252}/_{545}$ times (71.12%) more families than the low-frequency levels. According to Table 1.2, this text covers (a) 29.41% of all types and 84.63% of all families in the high-frequency vocabulary, (b) 10.39% of all types and 31.45% of all families in the mid-frequency vocabulary, as well as (c) 1.79% of all types and 3.41% of all families in the low-frequency vocabulary of the BNC/COCA subgroup.

Owing to their treatment as proper nouns, the predominant word types extraneous to their respective lists in the BNC/COCA subgroup, hence accounting for a commensurate quantitative superfluity, are *Pond* (L4: 29 tokens), *Shepherd* (L5: 125 tokens), *Hertz* (L8: 13 tokens), *Homestead* (L9: 33 tokens), *Finch* (L11: 17 tokens), *Miasma* (L14: 7 tokens),

Kroner (L16: 244 tokens), *Halyard* (L17: 147 tokens), *Ilium* (L18: 99 tokens), *Frascati* (L22: 4 tokens), *Proteus* (L23: 85 tokens), and *Esperanto* (L24: 1 token). The tokens and types that remain unclassified in Table 2.2 encompass the unclassifiable elements itemized in Table 2.3. In the latter table, a total of 41 tokens of 31 types bear a specious resemblance to various base forms in the *OED2*, including elements in fictional Bratpuhrian: *aki*, *akka*, *ebo*, *-fut*, *goura*, *koula*, *pala*, *pitty*, *puka*, *puku*, *sakki*, *serin*, and *simi*. Conversely, the 1,550 tokens in Appendix A can be argued to be classifiable, resulting in an approximate margin of error of 1.5%. With the addition of the exact figure to the aggregate of proper nouns and marginal words, the inclusion of which becomes wholly implicit, 98.21% coverage of the 97,992 requisite tokens is achieved with 9,000 word families. By comparison, a vocabulary of 8,000 families would limit reading comprehension to 97.87% of the text.

To summarize these data in relation to those collected from the independent corpus in its entirety, *Player Piano* is unique in that it surpasses the other constituents of the group with respect to both the size of the text and the complexity of the vocabulary. Additionally, it has more tokens, types, and families at each level of the GSL/AWL subgroup, as well as in the high-, mid-, and low-frequency supersets of the BNC/COCA lists, than any one of its counterparts. This, in turn, entails the highest coverage of the control group itself in these particular regards. Lastly, this novel reaches the highest proportion of types and families at the second GSL level and the AWL level, as well as the lowest proportion of high- to mid-frequency and high- to low-frequency types and families at the integral BNC/COCA levels.

The Sirens of Titan

Table 3.1 shows that, of the two GSL levels, the first one receives 13 $^{1363}/_{4538}$ times as many tokens, 1 $^{961}/_{1343}$ times as many types, and 1 $^{29}/_{107}$ times as many families, equating

Table 3.1: Analysis of *The Sirens of Titan* by the GSL/AWL lists

File	Tokens	(%)	Types	(%)	Families
1_gsl_1st_1000.txt	60,357	(77.81)	2,304	(28.50)	952
2_gsl_2nd_1000.txt	4,538	(5.85)	1,343	(16.61)	749
3_awl_570.txt	1,621	(2.09)	554	(6.85)	333
	11,054	(14.25)	3,884	(48.04)	
Total	77,570		8,085		2,034

Table 3.2: Analysis of *The Sirens of Titan* by the BNC/COCA lists

Level	Tokens	(%)	Types	(%)	Families
1st 1,000	60,802	(78.38)	2,401	(29.70)	955
2nd 1,000	5,256	(6.78)	1,574	(19.47)	825
3rd 1,000	2,074	(2.67)	909	(11.24)	593
4th 1,000	1,507	(1.94)	663	(8.20)	456
5th 1,000	900	(1.16)	446	(5.52)	355
6th 1,000	870	(1.12)	339	(4.19)	269
7th 1,000	566	(0.73)	240	(2.97)	209
8th 1,000	281	(0.36)	181	(2.24)	159
9th 1,000	240	(0.31)	160	(1.98)	132
10th 1,000	191	(0.25)	101	(1.25)	87
11th 1,000	127	(0.16)	79	(0.98)	74
12th 1,000	120	(0.15)	70	(0.87)	66
13th 1,000	60	(0.08)	34	(0.42)	32
14th 1,000	**82**	**(0.11)**	**43**	**(0.53)**	**40**
15th 1,000	**128**	**(0.17)**	23	(0.28)	19
16th 1,000	77	(0.10)	20	(0.25)	**19**
17th 1,000	18	(0.02)	16	(0.20)	16
18th 1,000	6	(0.01)	5	(0.06)	5
19th 1,000	**19**	**(0.02)**	**12**	**(0.15)**	**12**
20th 1,000	**42**	**(0.05)**	**15**	**(0.19)**	**14**
21st 1,000	17	(0.02)	6	(0.07)	6
22nd 1,000	5	(0.01)	4	(0.05)	4
23rd 1,000	4	(0.01)	**4**	**(0.05)**	**4**
24th 1,000	3	(0.00)	3	(0.04)	3
25th 1,000	**3**	**(0.00)**	**3**	**(0.04)**	**3**
Proper nouns	1,551	(2.00)	268	(3.31)	257
Exclamations	151	(0.19)	32	(0.40)	22
Transparent compounds	263	(0.34)	144	(1.78)	135
Abbreviations	12	(0.02)	10	(0.12)	10
Not in the lists	2,195	(2.83)	280	(3.46)	
Total	77,570		8,085		4,781

Table 3.3: Lexical anomalies in *The Sirens of Titan* pursuant to the *OED2*

progerse (5), *getcher* (3), *wuzza* (3), *doooooooooommmmmmmmmm* (2), *everthing* (2), *genuwine* (2), ***salpa* (2)**, *-stuh* (2), *afo*, *beebees*, *braugh*, *commmmmmmmme*, *dreat*, ***-erse***, *everbody*, *everwhere*, ***-faw***, *floof*, *fraugh*, *kroh-*, *-kup*, *mabba*, *plui*, *printemps*, *skiiiiiiiiiiiiiiiiiip*, *sumpin'*, *-tennnn-*
Total: 27 types (40 tokens)

to a difference of 92.48%, 41.71%, and 21.32% respectively. Relative to the AWL level, the conjoint lists of general words are in excess of 40 $^{55}/_{1621}$ (97.5%) of the tokens, 6 $^{323}/_{554}$ (84.81%) of the types, and 5 $^{4}/_{37}$ (80.42%) of the families. The text covers (a) 56% of every type and 95.39% of every family at the first GSL level, (b) 36.22% of every type and 75.81% of every family at the second GSL level, and (c) 17.98% of every type and 58.52% of every family at the AWL level.

In Table 3.2, the high-frequency superset contains 15 $^{668}/_{1091}$ times (93.59%) more tokens, 2 $^{826}/_{2029}$ times (58.46%) more types, and 1 $^{793}/_{1580}$ times (33.42%) more families than that of the mid-frequency levels, plus 75 $^{241}/_{451}$ times (98.68%) more tokens, 11 $^{11}/_{73}$ times (91.03%) more types, and 5 $^{353}/_{404}$ times (82.98%) more families than the low-frequency superset. The mid-frequency superset holds 4 $^{378}/_{451}$ times (79.33%) more tokens, 4 $^{277}/_{438}$ times (78.41%) more types, and 3 $^{92}/_{101}$ times (74.43%) more families than the low-frequency levels. The text covers (a) 25.56% of every type and 79.1% of every family in the high-frequency vocabulary, (b) 8.61% of every type and 26.33% of every family in the mid-frequency vocabulary, and (c) 1.33% of every type and 2.53% of every family in the low-frequency vocabulary.

The predominant types of proper nouns and derivatives subsumed under specific types extrinsic to their designated BNC/COCA level are *Constant* (L2: 487 tokens), *Skip* (L4: 34 tokens), *Mars* (L6: 149 tokens), *Titan* (L7: 64 tokens), *Galactic* (L9: 11 tokens), *Magnum* (L10: 30 tokens), *Bobby* (L11: 16 tokens), *Opus* (L12: 30 tokens), *Earthling*

(L15: 78 tokens), *Pablum/Psychokinesis* (L23: 1 token each), and *Cro-/Phi* (Abbreviations: 2 tokens each). The unclassifiable elements in Table 3.3 include homographs of base forms in the *OED2*, such as *-erse* and *-faw*, which should be read as nonphonemic transcriptions. The 2,155 tokens in Appendix B account for a margin of error of 2.78%. This figure, along with the proper nouns and marginal words in their respective lists, leaves 73,438 tokens for adjusted analysis. Cumulatively, the distribution of the final total across the fundamental levels of the control subgroup facilitates 98.01% coverage with 7,000 word families.

The Sirens of Titan achieves the lowest proportion of tokens at the conjoint GSL levels to those at the AWL level, for the oft-repeated proper noun *Constant* is counted as an academic word. It differs from *Player Piano* in that the proportion of types and families at the second GSL level and the AWL level has diminished by up to 8.03 and 1.72 percentage points respectively. Aside from a negligible decrease in its high- to mid-frequency tokens, the proportion of their enveloping types and families plus that of high- to low-frequency and mid- to low-frequency elements increased by up to 7.74, 4.44, and 10.35 percentage points respectively, with the percentage of mid- to low-frequency types reaching its apex. Overall, the coverage of the control group was reduced by 5.56% of the types and 4.88% of the families for the GSL, by 4.12% of the types and 8.79% of the families for the AWL, and by 1.72% of the types and 2.46% of the families for the BNC/COCA sequence of lists.

Mother Night

Table 4.1 shows that, of the two GSL levels, the first one receives 16 $^{1343}/_{2492}$ times as many tokens, 1 $^{914}/_{1043}$ times as many types, and 1 $^{251}/_{651}$ times as many families, equating to a difference of 93.95%, 46.7%, and 27.83% respectively. Relative to the AWL level, the conjoint lists of general words are in excess of 63 $^{300}/_{689}$ (98.42%) of the tokens, 8 $^{1}/_{3}$ (88%)

Table 4.1: Analysis of *Mother Night* by the GSL/AWL lists

File	Tokens	(%)	Types	(%)	Families
1_gsl_1st_1000.txt	41,215	(82.06)	1,957	(33.07)	902
2_gsl_2nd_1000.txt	2,492	(4.96)	1,043	(17.62)	651
3_awl_570.txt	689	(1.37)	360	(6.08)	243
	5,830	(11.61)	2,558	(43.22)	
Total	50,226		5,918		1,796

Table 4.2: Analysis of *Mother Night* by the BNC/COCA lists

Level	Tokens	(%)	Types	(%)	Families
1st 1,000	41,398	(82.42)	2,023	(34.18)	925
2nd 1,000	2,588	(5.15)	1,135	(19.18)	713
3rd 1,000	1,190	(2.37)	633	(10.70)	454
4th 1,000	641	(1.28)	394	(6.66)	310
5th 1,000	402	(0.80)	249	(4.21)	218
6th 1,000	**404**	**(0.80)**	209	(3.53)	184
7th 1,000	195	(0.39)	129	(2.18)	118
8th 1,000	162	(0.32)	102	(1.72)	97
9th 1,000	135	(0.27)	75	(1.27)	70
10th 1,000	86	(0.17)	70	(1.18)	68
11th 1,000	85	(0.17)	54	(0.91)	53
12th 1,000	35	(0.07)	30	(0.51)	30
13th 1,000	**35**	**(0.07)**	28	(0.47)	28
14th 1,000	30	(0.06)	18	(0.30)	18
15th 1,000	**38**	**(0.08)**	**23**	**(0.39)**	**22**
16th 1,000	13	(0.03)	9	(0.15)	8
17th 1,000	**14**	**(0.03)**	**12**	**(0.20)**	**12**
18th 1,000	**17**	**(0.03)**	9	(0.15)	8
19th 1,000	9	(0.02)	7	(0.12)	7
20th 1,000	3	(0.01)	3	(0.05)	3
21st 1,000	1	(0.00)	1	(0.02)	1
22nd 1,000	**2**	**(0.00)**	**2**	**(0.03)**	**2**
23rd 1,000	**4**	**(0.01)**	**4**	**(0.07)**	**4**
24th 1,000	2	(0.00)	2	(0.03)	2
25th 1,000	**7**	**(0.01)**	**4**	**(0.07)**	**4**
Proper Nouns	1,642	(3.27)	283	(4.78)	260
Exclamations	178	(0.35)	22	(0.37)	15
Transparent Compounds	211	(0.42)	116	(1.96)	104
Abbreviations	16	(0.03)	8	(0.14)	8
Not in the lists	683	(1.36)	264	(4.46)	
Total	50,226		5,918		3,746

Table 4.3: Lexical anomalies in *Mother Night* pursuant to the *OED2*

olly- (8), *-Bundesfuehrer* (7), *Wiedersehen* (5), *zwei* (3), *Dampfwalze* (2), ***doch* (2)**, *getrennt* (2), *Leib* (2), *Leichenträger* (2), *Lieb* (2), *Reichsleiter* (2), ***sein* (2)**, *vom* (2), ***Wache* (2)**, ***Walze* (2)**, *Zeitgeschehen* (2), *Abstand*, *allerwärts*, *andern*, *beiden*, *Bergeshöhen*, *bleiben*, *Blickes*, *Blutgeschickes*, *Brunnens*, *dafür*, *dahin*, *dieses*, *durchstreift*, *entfliehn*, *Erbleichen*, *erhält*, *Flucht*, *-freeeeeee*, *frei*, *Freiherr*, *geborgen*, *geduckte*, *gehn*, *Geschichte*, *Glocke*, *herab*, *Herz*, ***hier***, *ihr*, *kargen*, *keine*, *keiner*, ***Klang***, *kommt*, *Körpers*, *kühl*, *leerer*, *Leichen*, ***leis***, *liegt*, *Lohn*, *mächtige*, *Mägdlein*, *Menschen*, *Menschheit*, ***naht***, ***nicht***, *Oberdienstleiter*, *Opfer*, *Pfad*, *quälenden*, *Rätsel*, *rollt*, *schaun*, ***schone***, *schreit*, *schwarzen*, *sehn*, *sitzt*, *sollen*, ***Sonne***, *Sonnenaufgang*, *Sorgen*, *sprechen-*, *starren*, *sterben*, *sucht*, *tiefen*, *triffst*, *unseren*, *verfaulte*, *verbrenn*, *verderben*, *vorbei*, ***wo***, *wollen*
Total: 92 types (123 tokens)

of the types, and 6 $^{95}/_{243}$ (84.35%) of the families. The text covers (a) 47.57% of every type and 90.38% of every family at the first GSL level, (b) 28.13% of every type and 65.89% of every family at the second GSL level, and (c) 11.68% of every type and 42.71% of every family at the AWL level.

In Table 4.2, the high-frequency superset contains 23 $^{579}/_{1939}$ times (95.71%) more tokens, 3 $^{317}/_{1158}$ times (69.45%) more types, and 2 $^{98}/_{997}$ times (52.34%) more families than that of the mid-frequency levels, plus 118 $^{218}/_{381}$ times (99.16%) more tokens, 13 $^{203}/_{276}$ times (92.72%) more types, and 7 $^{101}/_{135}$ times (87.09%) more families than the low-frequency superset. The mid-frequency superset holds 5 $^{34}/_{381}$ times (80.35%) more tokens, 4 $^{9}/_{46}$ times (76.17%) more types, and 3 $^{187}/_{270}$ times (72.92%) more families than the low-frequency levels. The text covers (a) 19.84% of every type and 69.73% of every family in the high-frequency vocabulary, (b) 4.92% of every type and 16.62% of every family in the mid-frequency vocabulary, as well as (c) 0.84% of every type and 1.69% of every family in the low-frequency vocabulary.

The word types treated as proper nouns in their context that the analysis revealed to be predominant in other BNC/COCA lists are *-Hare* (L6: 35 tokens), *Grail* (L9: 8 tokens),

Minuteman (L15: 11 tokens), *Luger* (L16: 3 tokens), *Casanova* (L18: 7 tokens), *Klux* (L24: 1 token), *Gingiva-* (L25: 3 tokens), and *O-* (Exclamations: 35 tokens). A total of 16 tokens of 12 different types, all of which have been italicized in the source volume to convey their meaning in German, among those unclassifiable in Table 4.3 are homographs of base forms not defined in the *OED2*. The 560 classifiable tokens in Appendix C account for 1.11% of the total as a margin of error. The addition of this figure to the aggregate of proper nouns and marginal words at their designated levels for implicit inclusion means that the requisite quantity of 47,619 tokens attains 98.32% coverage with 7,000 word families. Alternatively, a vocabulary of 6,000 would provide access to as much as 97.91% of the text.

Mother Night developed the preceding text's upward trend in the proportion of the first GSL level to its second one by 1.47% of the tokens, 4.99% of the types, and 6.5% of the families, as well as in that of the combined GSL levels to the AWL one by 0.92% of the tokens, 3.19% of the types, and 3.93% of the families. As regards the BNC/COCA lists, the proportion of high- to mid-frequency tokens, types, and families rose by 2.11%, 11%, and 18.92% respectively, with the category of types reaching its height, while that of high- to low-frequency elements rose by 0.48% of the tokens, 1.69% of the types, and 4.12% of the families, with each category reaching its height. Although the proportion of mid- to low-frequency tokens rose by 1.02%, that of their types and families fell by 2.25% and 1.51% respectively. The coverage of all types and families fell by 8.27% and 7.45% for the GSL, by 6.29% and 15.82% for the AWL, and by 2.81% and 3.99% for these BNC/COCA levels.

Cat's Cradle

Table 5.1 shows that, of the two GSL levels, the first one receives 14 $^{2217}/_{2954}$ times as many tokens, 1 $^{905}/_{1089}$ times as many types, and 1 $^{53}/_{175}$ times as many families, equating

Table 5.1: Analysis of *Cat's Cradle* by the GSL/AWL lists

File	Tokens	(%)	Types	(%)	Families
1_gsl_1st_1000.txt	43,573	(80.09)	1,994	(31.33)	912
2_gsl_2nd_1000.txt	2,954	(5.43)	1,089	(17.11)	700
3_awl_570.txt	684	(1.26)	387	(6.08)	261
	7,195	(13.22)	2,895	(45.48)	
Total	54,406		6,365		1,873

Table 5.2: Analysis of *Cat's Cradle* by the BNC/COCA lists

Level	Tokens	(%)	Types	(%)	Families
1st 1,000	44,005	(80.88)	2,124	(33.37)	939
2nd 1,000	3,027	(5.56)	1,197	(18.81)	749
3rd 1,000	1,179	(2.17)	631	(9.91)	461
4th 1,000	813	(1.49)	460	(7.23)	370
5th 1,000	659	(1.21)	320	(5.03)	270
6th 1,000	353	(0.65)	231	(3.63)	200
7th 1,000	283	(0.52)	167	(2.62)	146
8th 1,000	**372**	**(0.68)**	125	(1.96)	114
9th 1,000	158	(0.29)	112	(1.76)	107
10th 1,000	113	(0.21)	69	(1.08)	64
11th 1,000	**114**	**(0.21)**	**70**	**(1.10)**	**64**
12th 1,000	56	(0.10)	40	(0.63)	40
13th 1,000	**75**	**(0.14)**	**44**	**(0.69)**	**41**
14th 1,000	32	(0.06)	25	(0.39)	24
15th 1,000	**37**	**(0.07)**	20	(0.31)	20
16th 1,000	25	(0.05)	14	(0.22)	13
17th 1,000	12	(0.02)	10	(0.16)	10
18th 1,000	**43**	**(0.08)**	**11**	**(0.17)**	**11**
19th 1,000	18	(0.03)	7	(0.11)	7
20th 1,000	16	(0.03)	**9**	**(0.14)**	**7**
21st 1,000	**16**	**(0.03)**	5	(0.08)	5
22nd 1,000	8	(0.01)	3	(0.05)	3
23rd 1,000	**17**	**(0.03)**	2	(0.03)	2
24th 1,000	2	(0.00)	**2**	**(0.03)**	**2**
25th 1,000	**2**	**(0.00)**	**2**	**(0.03)**	**2**
Proper nouns	1,597	(2.94)	278	(4.37)	252
Exclamations	173	(0.32)	28	(0.44)	19
Transparent compounds	231	(0.42)	125	(1.96)	115
Abbreviations	19	(0.03)	8	(0.13)	8
Not in the lists	951	(1.75)	226	(3.55)	
Total	54,406		6,365		4,065

Table 5.3: Lexical anomalies in *Cat's Cradle* pursuant to the *OED2*

karass (40), ***boko-* (15)**, *-maru* (14), *wampeter* (11), *duprass* (7), *foma* (6), *-kuh* (6), ***-ook-* (6)**, ***wrang* (6)**, ***hoon-* (5)**, *-toorz* (5), *-yera* (5), *zah-* (5), *-cratz-* (4), *granfalloon* (4), *jy* (4), *-kiul* (4), *tsvent-* (4), *zamoo-* (4), ***voo* (3)**, *wampeters* (3), ***dyot* (2)**, *-erlong* (2), *granfalloons* (2), ***lett-* (2)**, ***oon* (2)**, *sinookas* (2), ***stuppa* (2)**, *tsvantoor* (2), ***brath***, ***-dise***, *granfallooner*, *granfalloonery*, ***hap***, *hooner*, ***kon-***, *-maruing*, *peddiwinkus*, *sarooned*, *-shinik*, *soullllls*, *soullllls*, *-steez-*, *stopf*, *teetron*, *tz-*, *veglia*, *vorry*, *yeeara*, ***yeff***, *-yenk*, *yoze*
Total: 52 types (200 tokens)

to a difference of 93.22%, 45.39%, and 23.25% respectively. Relative to the AWL level, the conjoint lists of general words are in excess of 68 $^{5}/_{228}$ (98.53%) of the tokens, 7 $^{374}/_{387}$ (87.45%) of the types, and 6 $^{46}/_{261}$ (83.81%) of the families. The text covers (a) 48.47% of every type and 91.38% of every family at the first GSL level, (b) 29.37% of every type and 70.85% of every family at the second GSL level, and (c) 12.56% of every type and 45.87% of every family at the AWL level.

In Table 5.2, the high-frequency superset contains 18 $^{727}/_{2638}$ times (94.53%) more tokens, 2 $^{1122}/_{1415}$ times (64.2%) more types, and 1 $^{942}/_{1207}$ times (43.83%) more families than that of the mid-frequency levels, plus 82 $^{159}/_{586}$ times (98.78%) more tokens, 11 $^{289}/_{333}$ times (91.57%) more types, and 6 $^{37}/_{45}$ times (85.34%) more families than the low-frequency superset. The mid-frequency superset holds 4 $^{147}/_{293}$ times (77.79%) more tokens, 4 $^{83}/_{333}$ times (76.47%) more types, and 3 $^{262}/_{315}$ times (73.9%) more families than the low-frequency levels. The text covers (a) 20.68% of every type and 71.63% of every family in the high-frequency vocabulary, (b) 6.01% of every type and 20.12% of every family in the mid-frequency vocabulary, as well as (c) 1.01% of every type and 1.97% of every family in the low-frequency vocabulary.

The leading types of proper nouns and derivatives thereof in different BNC/COCA lists are *Castle* (L2: 106 tokens), *Papa* (L5: 123 tokens), *Newt* (L8: 153 tokens), *Foundry*

(L10: 9 tokens), *Calypso* (L13: 7 tokens), *Enders* (L17: 2 tokens), *Ilium* (L18: 33 tokens), *Humana* (L19: 6 tokens), and *Houdini* (L25: 1 token). A total of 48 tokens of 14 types in Table 5.3 are found to be homographs of base forms in the *OED2*, albeit unclassifiable for reasons of semantic incompatibility. The idiolectal aspect of their usage reveals that, with the exception of the syllable *-dise*, all elements highlighted in the aforesaid table occur in a fictional "English dialect /…/ of San Lorenzo" (Vonnegut 1963/2009: 108) in their context. In Appendix D, the 751 classifiable tokens represent a 1.38% margin of error. The implicit inclusion of both this percentage and the aggregate quantity of proper nouns and marginal words retains an adjusted figure of 51,635 tokens. As a result, the cumulative coverage of this text reaches 98.17% with 8,000 word families.

Cat's Cradle raised the proportion of tokens at the two GSL levels to the AWL one by another 0.11%, plus that of mid- to low-frequency types and families at the BNC/COCA levels by 0.3% and 0.98% respectively. Nevertheless, for the most part, it has reversed the general trend, raising the proportion of tokens, types, and families at the second GSL level between 0.73 and 4.58 percentage points, as well as that of types and families in the AWL between 0.54 and 0.55 percentage points. Furthermore, it lowered the proportion of high- to mid-frequency elements, high- to low-frequency elements, and mid- to low-frequency tokens, in percentage terms, between 1.18 and 8.51, between 0.37 and 1.75, and by 2.56 respectively. The coverage of types and families went up by 1.06% and 2.97% for the GSL, by 0.88% and 3.16% for the AWL, and by 0.63% and 1.25% for the BNC/COCA supersets.

God Bless You, Mr. Rosewater, or Pearls Before Swine

Table 6.1 shows that, of the two GSL levels, the first one receives 14 $^{521}/_{2793}$ times as many tokens, 1 $^{912}/_{1151}$ times as many types, and 1 $^{193}/_{726}$ times as many families, equating to

Table 6.1: Analysis of *God Bless You, Mr. Rosewater* by the GSL/AWL lists

File	Tokens	(%)	Types	(%)	Families
1_gsl_1st_1000.txt	39,623	(79.28)	2,063	(30.42)	919
2_gsl_2nd_1000.txt	2,793	(5.59)	1,151	(16.97)	726
3_awl_570.txt	781	(1.56)	427	(6.30)	287
	6,784	(13.57)	3,141	(46.31)	
Total	49,981		6,782		1,932

Table 6.2: Analysis of *God Bless You, Mr. Rosewater* by the BNC/COCA lists

Level	Tokens	(%)	Types	(%)	Families
1st 1,000	40,090	(80.21)	2,167	(31.95)	950
2nd 1,000	2,851	(5.70)	1,298	(19.14)	768
3rd 1,000	1,218	(2.44)	684	(10.09)	511
4th 1,000	952	(1.90)	513	(7.56)	402
5th 1,000	569	(1.14)	328	(4.84)	280
6th 1,000	352	(0.70)	215	(3.17)	189
7th 1,000	263	(0.53)	188	(2.77)	173
8th 1,000	188	(0.38)	144	(2.12)	135
9th 1,000	128	(0.26)	112	(1.65)	106
10th 1,000	125	(0.25)	92	(1.36)	85
11th 1,000	98	(0.20)	73	(1.08)	67
12th 1,000	79	(0.16)	64	(0.94)	60
13th 1,000	71	(0.14)	51	(0.75)	48
14th 1,000	44	(0.09)	30	(0.44)	28
15th 1,000	21	(0.04)	15	(0.22)	14
16th 1,000	14	(0.03)	13	(0.19)	13
17th 1,000	**15**	**(0.03)**	7	(0.10)	7
18th 1,000	9	(0.02)	**9**	**(0.13)**	**9**
19th 1,000	**11**	**(0.02)**	**10**	**(0.15)**	**10**
20th 1,000	**13**	**(0.03)**	**12**	**(0.18)**	**12**
21st 1,000	**13**	**(0.03)**	9	(0.13)	9
22nd 1,000	4	(0.01)	3	(0.04)	3
23rd 1,000	**5**	**(0.01)**	**4**	**(0.06)**	**4**
24th 1,000	**38**	**(0.08)**	1	(0.01)	1
25th 1,000	4	(0.01)	**4**	**(0.06)**	**4**
Proper nouns	1,705	(3.41)	334	(4.92)	321
Exclamations	112	(0.22)	30	(0.44)	21
Transparent compounds	257	(0.51)	143	(2.11)	136
Abbreviations	11	(0.02)	7	(0.10)	7
Not in the lists	721	(1.44)	222	(3.27)	
Total	49,981		6,782		4,373

Table 6.3: Lexical anomalies in *God Bless You, Mr. Rosewater* pursuant to the *OED2*

samaritrophia (5), *kiddleys* (3), *-blacka-* (2), *frusha-* (2), ***kiddley* (2)**, *anthelminica*, *eclipta*, *-erthrown*, *jambolina*, *-juihirka*, *lohopa-*, *muckety-*, *pluribus*, *prostata*, *samaritrophic*, ***swole***, *uranimum*, *veronia*
Total: 18 types (27 tokens)

a difference of 92.95%, 44.21%, and 21% respectively. Relative to the AWL level, the conjoint lists of general words are in excess of 54 $^{22}/_{71}$ (98.16%) of the tokens, 7 $^{225}/_{427}$ (86.71%) of the types, and 5 $^{30}/_{41}$ (82.55%) of the families. The text covers (a) 50.15% of every type and 92.08% of every family at the first GSL level, (b) 31.04% of every type and 73.48% of every family at the second GSL level, and (c) 13.85% of every type and 50.44% of every family at the AWL level.

In Table 6.2, the high-frequency superset contains 18 $^{23}/_{2452}$ times (94.45%) more tokens, 2 $^{383}/_{500}$ times (63.85%) more types, and 1 $^{944}/_{1285}$ times (42.35%) more families than that of the mid-frequency levels, plus 78 $^{167}/_{564}$ times (98.72%) more tokens, 10 $^{179}/_{397}$ times (90.43%) more types, and 5 $^{359}/_{374}$ times (83.22%) more families than the low-frequency superset. The mid-frequency superset holds 4 $^{49}/_{141}$ times (77%) more tokens, 3 $^{309}/_{397}$ times (73.53%) more types, and 3 $^{163}/_{374}$ times (70.89%) more families than the low-frequency levels. The text covers (a) 21.71% of every type and 74.3% of every family in the high-frequency vocabulary, (b) 6.37% of every type and 21.42% of every family in the mid-frequency vocabulary, as well as (c) 1.2% of every type and 2.34% of every family in the low-frequency vocabulary.

The leading types of proper nouns and derivatives thereof outside the designated BNC/COCA list are *Trout* (L5: 47 tokens), *Bunny* (L6: 34 tokens), *Randy* (L12: 6 tokens), *Weir* (L13: 5 tokens), *Parthenon* (L14: 8 tokens), *Earthling* (L15: 5 tokens), *Ambrosia* (L17: 9 tokens), *Palindrome* (L18: 1 token), *Sutra* (L20: 2 tokens), *Kama* (L23: 2 tokens),

Amanita (L24: 38 tokens), *Greco* (L25: 1 token), and *Phi* (Abbreviations: 3 tokens). Only two word types among those unclassifiable in Table 6.3 are homographs of unrelated base forms in the *OED2*: *kiddley* and *swole*, the context of which seems to make it obvious that the signifier is a mispronunciation. The 694 ideally classifiable tokens in Appendix E yield a 1.39% margin of error, relative to the overall size of the text. Adding this quantity to the proper nouns and marginal words at the levels of the control subgroup that assume lexical simplicity leaves an adjusted figure of 47,202 tokens. Therefore, a 98.08% coverage level of this text is attainable with 7,000 word families.

The preceding text's reversal of the decline in the occurrence of exceptional lexical elements has gained ubiquity in *God Bless You, Mr. Rosewater*, with the relationship of the first GSL level to its second one marked by a decrease of up to 2.24 percentage points and that of the GSL to the AWL by up to 1.26 percentage points. Likewise, in the case of the supersets of the BNC/COCA lists, the proportion of high- to mid-frequency, high- to low-frequency, and mid- to low-frequency elements was reduced by up to 1.48, 2.12, and 3.01 percentage points respectively. Concomitant to this was an increase in the coverage of the control group, specifically by 1.67% of the types and 1.66% of the families for the GSL, by 1.3% of the types and 4.57% of the families for the AWL, as well as by 0.46% of the types and 0.87% of the families for the BNC/COCA lists graded by frequency.

Slaughterhouse-Five, or The Children's Crusade: A Duty-Dance with Death

Table 7.1 shows that, of the two GSL levels, the first one receives 12 $^{2944}/_{3103}$ times as many tokens, 1 $^{791}/_{1150}$ times as many types, and 1 $^{204}/_{713}$ times as many families, equating to a difference of 92.28%, 40.75%, and 22.25% respectively. Relative to the AWL level, the conjoint lists of general words are in excess of 84 $^{107}/_{514}$ (98.81%) of the tokens, 9 $^{301}/_{310}$

Table 7.1: Analysis of *Slaughterhouse-Five* by the GSL/AWL lists

File	Tokens	(%)	Types	(%)	Families
1_gsl_1st_1000.txt	40,180	(79.02)	1,941	(31.24)	917
2_gsl_2nd_1000.txt	3,103	(6.10)	1,150	(18.51)	713
3_awl_570.txt	514	(1.01)	310	(4.99)	230
	7,054	(13.87)	2,812	(45.26)	
Total	50,851		6,213		1,860

Table 7.2: Analysis of *Slaughterhouse-Five* by the BNC/COCA lists

Level	Tokens	(%)	Types	(%)	Families
1st 1,000	40,898	(80.43)	2,096	(33.74)	951
2nd 1,000	2,716	(5.34)	1,184	(19.06)	750
3rd 1,000	1,029	(2.02)	565	(9.09)	434
4th 1,000	**1,121**	**(2.20)**	475	(7.65)	363
5th 1,000	644	(1.27)	320	(5.15)	262
6th 1,000	439	(0.86)	232	(3.73)	197
7th 1,000	245	(0.48)	164	(2.64)	140
8th 1,000	200	(0.39)	103	(1.66)	96
9th 1,000	139	(0.27)	92	(1.48)	83
10th 1,000	105	(0.21)	72	(1.16)	64
11th 1,000	**106**	**(0.21)**	68	(1.09)	62
12th 1,000	55	(0.11)	41	(0.66)	40
13th 1,000	49	(0.10)	35	(0.56)	34
14th 1,000	**65**	**(0.13)**	26	(0.42)	23
15th 1,000	**89**	**(0.18)**	**26**	**(0.42)**	**24**
16th 1,000	15	(0.03)	10	(0.16)	9
17th 1,000	5	(0.01)	5	(0.08)	5
18th 1,000	**49**	**(0.10)**	**7**	**(0.11)**	**6**
19th 1,000	7	(0.01)	**7**	**(0.11)**	**6**
20th 1,000	4	(0.01)	4	(0.06)	4
21st 1,000	**8**	**(0.02)**	**5**	**(0.08)**	**4**
22nd 1,000	5	(0.01)	2	(0.03)	2
23rd 1,000	1	(0.00)	1	(0.02)	1
24th 1,000	**4**	**(0.01)**	**4**	**(0.06)**	**4**
25th 1,000	2	(0.00)	2	(0.03)	2
Proper nouns	1,892	(3.72)	270	(4.35)	251
Exclamations	143	(0.28)	32	(0.52)	20
Transparent compounds	331	(0.65)	158	(2.54)	147
Abbreviations	9	(0.02)	5	(0.08)	5
Not in the lists	476	(0.94)	202	(3.25)	
Total	50,851		6,213		3,989

Table 7.3: Lexical anomalies in *Slaughterhouse-Five* pursuant to the *OED2*

Schlachthof- (3), *deedlee-* (2), *-fünf* (2), *Kuppel* (2), *vork* (2), ***vy* (2)**, *alsdann*, *Baumeisters*, *bedenklich*, *bombenfest*, *deutete*, *diese*, *drivin'*, *eheu*, *einen*, *eingerichtet*, *erbaut*, ***Feind***, *fourragère*, *fugaces*, *gethan*, ***gute***, ***hatte***, *hineingesät*, *Kirche*, *Küster*, *labuntur*, *lakonisch*, *leidigen*, *lightnin'*, ***nach***, *Ordnung*, *rühmte*, *Ruinene*, *sagte*, ***sah***, *Sakristan*, *schon*, *schöne*, *Seiten*, *smashin'*, *städtische*, *Trümmer*, *unerwünschten*, *veek*, *vunce*, *welcher*, *zwischen*
Total: 48 types (55 tokens)

(89.97%) of the types, and 7 $^{2}/_{23}$ (85.89%) of the families. The text covers (a) 47.18% of every type and 91.88% of every family at the first GSL level, (b) 31.01% of every type and 72.17% of every family at the second GSL level, and (c) 10.06% of every type and 40.42% of every family at the AWL level.

In Table 7.2, the high-frequency superset contains 16 $^{35}/_{2788}$ times (93.75%) more tokens, 2 $^{1073}/_{1386}$ times (63.95%) more types, and 1 $^{142}/_{163}$ times (46.56%) more families than that of the mid-frequency levels, plus 78 $^{261}/_{569}$ times (98.73%) more tokens, 12 $^{13}/_{63}$ times (91.81%) more types, and 7 $^{21}/_{58}$ times (86.42%) more families than the low-frequency superset. The mid-frequency superset holds 4 $^{512}/_{569}$ times (79.59%) more tokens, 4 $^{2}/_{5}$ times (77.27%) more types, and 3 $^{271}/_{290}$ times (74.58%) more families than the low-frequency levels. The text covers (a) 20.12% of every type and 71.17% of every family in the high-frequency vocabulary, (b) 5.88% of every type and 19.02% of every family in the mid-frequency vocabulary, as well as (c) 0.95% of every type and 1.81% of every family in the low-frequency vocabulary.

The leading types of proper nouns and derivatives thereof in other BNC/COCA lists are *Pilgrim* (L4: 122 tokens), *Trout* (L5: 72 tokens), *-Hare* (L6: 29 tokens), *Derby* (L8: 54 tokens), *Musketeers* (L9: 14 tokens), *Earthling* (L15: 21 tokens), *Ilium* (L18: 42 tokens), *Tweedledee* (L23: 1 token), *Derringer/Golgotha/Tweedledum* (L24: 1 token each), *Sodom* (L25: 1 token), *O-* (Exclamations: 29 tokens), and *Englishman* (Transparent compounds:

22 tokens). 7 tokens of 6 types among those unclassifiable in Table 7.3 resemble distinct *OED2* base forms, being distinguished in meaning from their counterparts. To be precise, in that table, *vy* conveys a German accent, whereas the rest of the types highlighted have been italicized as foreign words in the source volume. The 421 ideally classifiable tokens in Appendix F constitute a 0.83% margin of error, which, added to the quantity of proper nouns and marginal words at their levels, retains a total of 48,055 tokens for an adjusted analysis of this text. Exactly 98% of these tokens would be rendered familiar by a 7,000-word-family vocabulary, with its contiguous levels effecting 97.49% and 98.41% coverage.

On the one hand, *Slaughterhouse-Five* increased the proportion of the second GSL level by 0.67% of the tokens and 3.46% of the types, with the former category reaching its height. On the other, it increased the proportion of families at the first GSL level by 1.25%, as well as that of the entirety of the GSL to the AWL by up to 3.34%, with every category in the latter set reaching its height. In the case of the BNC/COCA levels, the relationship of high- to mid-frequency tokens was marked by a decrease of 0.69%, whereas the percentage increased, respectively, by 0.11 and 4.21 points for their enveloping types and families, by up to 3.2 points for high- to low-frequency elements, plus by up to 3.74 points for mid- to low-frequency elements. The coverage of all types and families was reduced, in percentage terms, by 1.57 and 0.76 points for the GSL, by 3.8 and 10.02 points to an all-time low for the AWL, as well as by 0.66 and 1.29 points for the ternary system of the BNC/COCA lists.

Breakfast of Champions, or Goodbye Blue Monday

Table 8.1 shows that, of the two GSL levels, the first one receives 13 $^{653}/_{1749}$ times as many tokens, 1 $^{893}/_{1211}$ times as many types, and 1 $^{103}/_{360}$ times as many families, equating to a difference of 92.52%, 42.44%, and 22.25% respectively. Relative to the AWL level, the

Table 8.1: Analysis of *Breakfast of Champions* by the GSL/AWL lists

File	Tokens	(%)	Types	(%)	Families
1_gsl_1st_1000.txt	46,780	(78.71)	2,104	(31.45)	926
2_gsl_2nd_1000.txt	3,498	(5.89)	1,211	(18.10)	720
3_awl_570.txt	949	(1.60)	413	(6.17)	272
	8,209	(13.81)	2,961	(44.27)	
Total	59,436		6,689		1,918

Table 8.2: Analysis of *Breakfast of Champions* by the BNC/COCA lists

Level	Tokens	(%)	Types	(%)	Families
1st 1,000	47,315	(79.61)	2,219	(33.17)	951
2nd 1,000	3,618	(6.09)	1,325	(19.81)	779
3rd 1,000	1,434	(2.41)	664	(9.93)	479
4th 1,000	1,102	(1.85)	478	(7.15)	368
5th 1,000	1,073	(1.81)	345	(5.16)	282
6th 1,000	425	(0.72)	229	(3.42)	199
7th 1,000	382	(0.64)	162	(2.42)	145
8th 1,000	205	(0.34)	115	(1.72)	104
9th 1,000	181	(0.30)	98	(1.47)	93
10th 1,000	104	(0.17)	68	(1.02)	61
11th 1,000	93	(0.16)	60	(0.90)	56
12th 1,000	**93**	**(0.16)**	59	(0.88)	51
13th 1,000	56	(0.09)	38	(0.57)	35
14th 1,000	55	(0.09)	29	(0.43)	29
15th 1,000	47	(0.08)	21	(0.31)	20
16th 1,000	12	(0.02)	12	(0.18)	12
17th 1,000	11	(0.02)	9	(0.13)	8
18th 1,000	4	(0.01)	4	(0.06)	4
19th 1,000	**13**	**(0.02)**	**7**	**(0.10)**	**7**
20th 1,000	4	(0.01)	3	(0.04)	3
21st 1,000	**15**	**(0.03)**	**10**	**(0.15)**	**10**
22nd 1,000	9	(0.02)	3	(0.04)	3
23rd 1,000	**24**	**(0.04)**	**4**	**(0.06)**	**4**
24th 1,000	1	(0.00)	1	(0.01)	1
25th 1,000	**4**	**(0.01)**	**3**	**(0.04)**	**3**
Proper nouns	1,910	(3.21)	304	(4.54)	286
Exclamations	138	(0.23)	31	(0.46)	20
Transparent compounds	337	(0.57)	179	(2.68)	162
Abbreviations	4	(0.01)	4	(0.06)	4
Not in the lists	767	(1.29)	205	(3.06)	
Total	59,436		6,689		4,179

Table 8.3: Lexical anomalies in *Breakfast of Champions* pursuant to the *OED2*

-gaffner (7), *-eeeeem* (5), *gilgongo* (4), *shazzbutter* (4), *olly-* (2), ***'roun'* (2)**, *whuffo* (2), ***demain***, *espérons*, *feets*, *-freeeeeeeeeeeeeeeeeeeeeeeeeee*, *glurp*, ***-io***, *lambos*, *longtemps*, *morepark*, *movin'*, *needin'*, *pluribus*, *vivra*, *wahee-*, *watchin'*, *wavin'*, ***woy***
Total: 24 types (43 tokens)

conjoint lists of general words are in excess of 52 $^{930}/_{949}$ (98.11%) of the tokens, 8 $^{11}/_{413}$ (87.54%) of the types, and 6 $^{7}/_{136}$ (83.48%) of the families. The text covers (a) 51.14% of every type and 92.79% of every family at the first GSL level, (b) 32.66% of every type and 72.87% of every family at the second GSL level, and (c) 13.4% of every type and 47.8% of every family at the AWL level.

In Table 8.2, the high-frequency superset contains 15 $^{1847}/_{3368}$ times (93.57%) more tokens, 2 $^{1354}/_{1427}$ times (66.09%) more types, and 1 $^{1018}/_{1191}$ times (46.08%) more families than that of the mid-frequency levels, plus 96 $^{47}/_{545}$ times (98.96%) more tokens, 12 $^{236}/_{331}$ times (92.13%) more types, and 7 $^{60}/_{307}$ times (86.1%) more families than the low-frequency superset. The mid-frequency superset holds 6 $^{98}/_{545}$ times (83.82%) more tokens, 4 $^{103}/_{331}$ times (76.8%) more types, and 3 $^{270}/_{307}$ times (74.22%) more families than the low-frequency levels. The text covers (a) 22.02% of every type and 73.63% of every family in the high-frequency vocabulary, (b) 6.06% of every type and 19.85% of every family in the mid-frequency vocabulary, as well as (c) 1% of every type and 1.92% of every family in the low-frequency vocabulary.

The predominant types of proper nouns and subsumed derivatives at different levels of the BNC/COCA subgroup are *Trout* (L5: 396 tokens), *Bunny* (L6: 49 tokens), *Hoover* (L7: 117), *Nigger* (L8: 22 tokens), *Bannister* (L9: 20 tokens), *Armistice* (L12: 5 tokens), *Earthlings* (L15: 11 tokens), *Astro-* (L16: 1 token), *Milo* (L23: 21 tokens), *Greco* (L25: 2 tokens), and *Al* (Abbreviations: 1 token). 5 tokens of 4 distinct types in Table 8.3 resemble

certain base forms in the *OED2* exclusively in terms of spelling; for example, *-io* and *woy* involve elements known to this control subgroup in relation to an attempt "to decode the mysterious words phonetically" (Vonnegut 1973/2009: 195). The 724 classifiable tokens in Appendix G account for a 1.22% margin of error. This, coupled with the aggregate quantity of proper nouns and marginal words at their designated levels, contrasts with an adjusted total of 56,323 tokens. As a result, the text reaches 98.27% coverage with a vocabulary of 7,000 word families.

In the case of *Breakfast of Champions*, the alternation of the two trends has become quite erratic. One example of this is that the percentage of tokens and types at the second GSL level fell between 0.25 and 1.69 points, whereas their encompassing families rose by a minute fraction of a point, in addition to the GSL falling in relation to the AWL by up to 2.43 points. Similarly, in the BNC/COCA subgroup, the proportion fell by up to 0.47% for high- to mid-frequency tokens and families, by 0.31% for high- to low-frequency families, plus by up to 0.47% for mid- to low-frequency types and families, but rose by 2.14% for high- to mid-frequency types, by up to 0.33% for high- to low-frequency tokens and types, plus by up to 4.23% for mid- to low-frequency tokens. On the whole, the coverage of types and families was up by 2.86% and 0.81% for the GSL, by 3.34% and 7.38% for the AWL, as well as by 0.55% and 0.56% for the BNC/COCA lists graded by frequency.

Slapstick, or Lonesome No More!

Table 9.1 shows that, of the two GSL levels, the first one receives 14 $^{895}\!/_{2176}$ times as many tokens, 1 $^{887}\!/_{964}$ times as many types, and 1 $^{84}\!/_{211}$ times as many families, equating to a difference of 93.06%, 47.92%, and 28.47% respectively. Relative to the AWL level, the conjoint lists of general words are in excess of 59 $^{82}\!/_{567}$ (98.31%) of the tokens, 8 $^{31}\!/_{348}$

Table 9.1: Analysis of *Slapstick* by the GSL/AWL list

File	Tokens	(%)	Types	(%)	Families
1_gsl_1st_1000.txt	31,359	(80.92)	1,851	(33.94)	885
2_gsl_2nd_1000.txt	2,176	(5.62)	964	(17.68)	633
3_awl_570.txt	567	(1.46)	348	(6.38)	239
	4,650	(12.00)	2,290	(42.00)	
Total	38,752		5,453		1,757

Table 9.2: Analysis of *Slapstick* by the BNC/COCA lists

Level	Tokens	(%)	Types	(%)	Families
1st 1,000	31,553	(81.42)	1,919	(35.19)	912
2nd 1,000	2,169	(5.60)	1,042	(19.11)	658
3rd 1,000	1,023	(2.64)	585	(10.73)	433
4th 1,000	639	(1.65)	368	(6.75)	297
5th 1,000	435	(1.12)	262	(4.80)	226
6th 1,000	242	(0.62)	168	(3.08)	150
7th 1,000	183	(0.47)	126	(2.31)	117
8th 1,000	137	(0.35)	94	(1.72)	87
9th 1,000	**169**	**(0.44)**	82	(1.50)	76
10th 1,000	85	(0.22)	62	(1.14)	56
11th 1,000	**87**	**(0.22)**	55	(1.01)	51
12th 1,000	42	(0.11)	30	(0.55)	29
13th 1,000	41	(0.11)	24	(0.44)	22
14th 1,000	**45**	**(0.12)**	**26**	**(0.48)**	**26**
15th 1,000	21	(0.05)	19	(0.35)	18
16th 1,000	**23**	**(0.06)**	17	(0.31)	16
17th 1,000	10	(0.03)	7	(0.13)	7
18th 1,000	**12**	**(0.03)**	**9**	**(0.17)**	**8**
19th 1,000	5	(0.01)	5	(0.09)	5
20th 1,000	**11**	**(0.03)**	**9**	**(0.17)**	**9**
21st 1,000	5	(0.01)	5	(0.09)	5
22nd 1,000	**5**	**(0.01)**	4	(0.07)	4
23rd 1,000	**5**	**(0.01)**	2	(0.04)	2
24th 1,000	1	(0.00)	1	(0.02)	1
25th 1,000	**4**	**(0.01)**	**3**	**(0.06)**	**3**
Proper nouns	1,099	(2.84)	245	(4.49)	229
Exclamations	150	(0.39)	21	(0.39)	17
Transparent compounds	213	(0.55)	121	(2.22)	109
Abbreviations	3	(0.01)	2	(0.04)	2
Not in the lists	335	(0.86)	140	(2.57)	
Total	38,752		5,453		3,575

Table 9.3: Lexical anomalies in *Slapstick* pursuant to the *OED2*

bluth- (5), *-luh* (5), ***buh* (3)**, *flocka* (3), *-lub* (3), *mub-* (3), ***fuff-* (2)**, *mooooooooooooon* (2), *-nition* (2), *baptiz-*, *bluh*, *brudder*, ***Ende***, *deserv-*, *meester*, *mooooooooooooon*
Total: 16 types (35 tokens)

(87.64%) of the types, and 6 $^{84}/_{239}$ (84.26%) of the families. The text covers (a) 44.99% of every type and 88.68% of every family at the first GSL level, (b) 26% of every type and 64.07% of every family at the second GSL level, and (c) 11.29% of every type and 42% of every family at the AWL level.

In Table 9.2, the high-frequency superset contains 19 $^{90}/_{361}$ times (94.81%) more tokens, 3 $^{123}/_{550}$ times (68.98%) more types, and 2 $^{97}/_{953}$ times (52.42%) more families than that of the mid-frequency levels, plus 86 $^{173}/_{402}$ times (98.84%) more tokens, 12 $^{105}/_{139}$ times (92.16%) more types, and 7 $^{169}/_{262}$ times (86.92%) more families than the low-frequency superset. The mid-frequency superset holds 4 $^{197}/_{402}$ times (77.73%) more tokens, 3 $^{133}/_{139}$ times (74.73%) more types, and 3 $^{167}/_{262}$ times (72.51%) more families than the low-frequency levels. The text covers (a) 18.56% of every type and 66.77% of every family in the high-frequency vocabulary, (b) 4.67% of every type and 15.88% of every family in the mid-frequency vocabulary, as well as (c) 0.84% of every type and 1.64% of every family in the low-frequency vocabulary.

The leading types of proper nouns and derivatives thereof at different levels of the BNC/COCA subgroup are *Melody* (L4: 40 tokens), *-Hare* (L6: 9 tokens), *Raspberries* (L7: 11 tokens), *Daffodil* (L9: 39 tokens), *Bobby* (L11: 12 tokens), *Fu* (L12: 9 tokens), *Oriole-* (L13: 7 tokens), *Chipmunk* (L14: 16 tokens), *Goober* (L20: 2 tokens), *Tarantella* (L21: 1 token), and *Pachysandra* (L23: 4 tokens). The unclassifiable elements in Table 9.3 include the German word *Ende* and two varieties of an “idiot word” (Vonnegut 1976/2010: 66) as

homographs of *OED2* base forms. The 300 tokens in Appendix H, which are adjudged to be classifiable, account for a text-specific margin of error of 0.77%. With its inclusion in conjunction with proper nouns and marginal words, the final total comes to 36,987 tokens. Therefore, one can exceed the target coverage by 0.36 percentage points with a receptive vocabulary knowledge of 8,000 word families, although even 7,000 families would suffice to enable familiarity with up to 97.99% of the tokens.

Compared to the previous work, *Slapstick* reduced the respective proportion of both the second GSL level and the AWL, with the percentage of each category increasing by up to 6.23 points for the first GSL level and by up to 0.78 points for the entire GSL. As to the BNC/COCA lists, the proportion rose by up to 6.34% for high- to mid-frequency elements, as well as by 0.03% of the types and 0.82% of the families for the high-frequency levels in relation to the low-frequency ones. It fell by 0.12% for the tokens thereat, however, and by up to 6.09% for mid- to low-frequency elements. The percentage of types and families at the first GSL level reached its apex, as did high-frequency BNC/COCA families in relation to mid-frequency ones. The overall coverage of types and families was down by 6.39% and 6.45% to an all-time low for the GSL, by 2.11% and 5.8% for the AWL, and by 1.38% and 1.96% to an all-time low for the graded BNC/COCA levels, implicating the size of the text.

Jailbird

Table 10.1 shows that, of the two GSL levels, the first one receives 16 $^{97}\!/_{3818}$ times as many tokens, 1 $^{531}\!/_{622}$ times as many types, and 1 $^{71}\!/_{248}$ times as many families, equating to a difference of 93.76%, 46.05%, and 22.26% respectively. Relative to the AWL level, the conjoint lists of general words are in excess of 59 $^{929}\!/_{1086}$ (98.33%) of the tokens, 7 $^{12}\!/_{49}$ (86.2%) of the types, and 5 $^{23}\!/_{44}$ (81.89%) of the families. The text covers (a) 56.05% of

Table 10.1: Analysis of *Jailbird* by the GSL/AWL lists

File	Tokens	(%)	Types	(%)	Families
1_gsl_1st_1000.txt	61,185	(81.92)	2,306	(29.28)	957
2_gsl_2nd_1000.txt	3,818	(5.11)	1,244	(15.79)	744
3_awl_570.txt	1,086	(1.45)	490	(6.22)	308
	8,596	(11.51)	3,836	(48.70)	
Total	74,685		7,876		2,009

Table 10.2: Analysis of *Jailbird* by the BNC/COCA lists

Level	Tokens	(%)	Types	(%)	Families
1st 1,000	61,489	(82.33)	2,424	(30.78)	964
2nd 1,000	3,982	(5.33)	1,397	(17.74)	818
3rd 1,000	1,908	(2.55)	871	(11.06)	578
4th 1,000	969	(1.30)	569	(7.22)	429
5th 1,000	706	(0.95)	397	(5.04)	329
6th 1,000	461	(0.62)	290	(3.68)	245
7th 1,000	305	(0.41)	197	(2.50)	173
8th 1,000	251	(0.34)	151	(1.92)	131
9th 1,000	166	(0.22)	114	(1.45)	108
10th 1,000	147	(0.20)	91	(1.16)	85
11th 1,000	90	(0.12)	69	(0.88)	66
12th 1,000	71	(0.10)	53	(0.67)	52
13th 1,000	53	(0.07)	37	(0.47)	36
14th 1,000	**55**	**(0.07)**	**39**	**(0.50)**	**38**
15th 1,000	39	(0.05)	32	(0.41)	31
16th 1,000	29	(0.04)	19	(0.24)	19
17th 1,000	27	(0.04)	**22**	**(0.28)**	**21**
18th 1,000	11	(0.01)	7	(0.09)	7
19th 1,000	**19**	**(0.03)**	**15**	**(0.19)**	**14**
20th 1,000	10	(0.01)	9	(0.11)	9
21st 1,000	**22**	**(0.03)**	**9**	**(0.11)**	**9**
22nd 1,000	5	(0.01)	5	(0.06)	5
23rd 1,000	4	(0.01)	3	(0.04)	3
24th 1,000	**5**	**(0.01)**	**5**	**(0.06)**	**5**
25th 1,000	3	(0.00)	3	(0.04)	3
Proper nouns	2,432	(3.26)	519	(6.59)	486
Exclamations	159	(0.21)	29	(0.37)	24
Transparent compounds	349	(0.47)	210	(2.67)	186
Abbreviations	21	(0.03)	8	(0.10)	8
Not in the lists	897	(1.20)	282	(3.58)	
Total	74,685		7,876		4,882

Table 10.3: Lexical anomalies in *Jailbird* pursuant to the *OED2*

bup- (5), *luh-* (5), *muh-* (5), *-honneur* (3), *booping* (2), ***puh-* (2)**, *regrette* (2), *squh-* (2), *wuh-* (2), *bluh*, *boops*, *calabozo*, *delecti*, *fluh*, *miiiiiiiiiiiiiiiillionnnnnnnnn*, *mooooooooooooooooon*, *permettez-*, ***quod***, ***seepy-***
Total: 19 types (38 tokens)

every type and 95.89% of every family at the first GSL level, (b) 33.55% of every type and 75.3% of every family at the second GSL level, and (c) 15.9% of every type and 54.13% of every family at the AWL level.

In Table 10.2, the high-frequency superset contains 23 $^{1645}/_{2858}$ times (95.76%) more tokens, 2 $^{628}/_{859}$ times (63.38%) more types, and 1 $^{189}/_{283}$ times (40.04%) more families than that of the mid-frequency levels, plus 114 $^{119}/_{590}$ times (99.12%) more tokens, 11 $^{47}/_{209}$ times (91.09%) more types, and 5 $^{345}/_{403}$ times (82.92%) more families than the low-frequency superset. The mid-frequency superset holds 4 $^{249}/_{295}$ times (79.36%) more tokens, 4 $^{23}/_{209}$ times (75.67%) more types, and 3 $^{206}/_{403}$ times (71.52%) more families than the low-frequency levels. The text covers (a) 24.56% of every type and 78.67% of every family in the high-frequency vocabulary, (b) 7.29% of every type and 23.58% of every family in the mid-frequency vocabulary, as well as (c) 1.27% of every type and 2.52% of every family in the low-frequency vocabulary.

The predominant types of proper nouns and subsumed derivatives at different levels of the BNC/COCA subgroup are *Fender* (L5: 40 tokens), *Harp* (L7: 14 tokens), *Mormons* (L9: 8 tokens), *-Looney* (L10: 25 tokens), *Vicuna* (L21: 14 tokens), *Pekingese* (L22: 1 token) *Brattle/Golgotha/Guardia/Klux* (L24: 1 token each), *Houdini* (L25: 1 token), *O-* (Excalamations: 26 tokens), and *Phi* (Abbreviations: 4 tokens). Only 4 tokens of 3 separate types in Table 10.3 are homographs of base forms not defined in the *OED2*, although these seem to be otherwise unrelated. The proportion of ideally classifiable elements constituting

Appendix I, on the other hand, consist of 859 tokens, which translates into a 1.15% margin of error. With its inclusion in conjunction with all proper nouns and marginal words at their specific levels, the final total comes to 70,865 tokens. This text reaches a 98.09% coverage level with a vocabulary of no more than 6,000 word families.

The relative lexical simplicity of *Jailbird* is evidently complicated by the repetition of objectively frequent elements therein. Indicative of this is the fact that the proportion of the first GSL level to its second one rose by 0.7% of the tokens, accompanied by a fall of up to 6.22% in types and families, whereas that of the GSL to the AWL rose by 0.02% of the tokens, accompanied by a fall of up to 2.36% in types and families. In the BNC/COCA subgroup, the proportion of high- to mid-frequency tokens, high- to low-frequency tokens, and mid- to low-frequency tokens and types rose by 0.95%, by 0.28%, and by up to 1.63% respectively, falling by up to 12.38% of the types and families in the first case, by up to 4% of the types and families in the second case, and by 0.99% of the families in the third case. In general, the coverage of the control group increased by 9.4% of the types and 9.21% of the families for the GSL, by 4.61% of the types and 12.13% of the families for the AWL, plus by 2.52% of the types and 3.84% of the families for the respective BNC/COCA levels.

Deadeye Dick

Table 11.1 shows that, of the two GSL levels, the first one receives 14 $^{3019}/_{3268}$ times as many tokens, 1 $^{971}/_{1091}$ times as many types, and 1 $^{238}/_{683}$ times as many families, equating to a difference of 93.3%, 47.09%, and 25.84% respectively. Relative to the AWL level, the conjoint lists of general words are in excess of 76 $^{587}/_{677}$ (98.7%) of the tokens, 8 $^{91}/_{120}$ (88.58%) of the types, and 6 $^{74}/_{255}$ (84.1%) of the families. The text covers (a) 50.12% of every type and 92.28% of every family at the first GSL level, (b) 29.42% of every type and

Table 11.1: Analysis of *Deadeye Dick* by the GSL/AWL lists

File	Tokens	(%)	Types	(%)	Families
1_gsl_1st_1000.txt	48,771	(81.89)	2,062	(32.55)	921
2_gsl_2nd_1000.txt	3,268	(5.49)	1,091	(17.22)	683
3_awl_570.txt	677	(1.14)	360	(5.68)	255
	6,843	(11.49)	2,822	(44.55)	
Total	59,559		6,335		1,859

Table 11.2: Analysis of *Deadeye Dick* by the BNC/COCA lists

Level	Tokens	(%)	Types	(%)	Families
1st 1,000	49,505	(83.12)	2,170	(34.25)	954
2nd 1,000	2,966	(4.98)	1,180	(18.63)	739
3rd 1,000	1,160	(1.95)	665	(10.50)	474
4th 1,000	812	(1.36)	441	(6.96)	371
5th 1,000	501	(0.84)	294	(4.64)	256
6th 1,000	372	(0.62)	218	(3.44)	195
7th 1,000	218	(0.37)	142	(2.24)	130
8th 1,000	202	(0.34)	94	(1.48)	86
9th 1,000	117	(0.20)	81	(1.28)	79
10th 1,000	100	(0.17)	65	(1.03)	59
11th 1,000	87	(0.15)	55	(0.87)	49
12th 1,000	**102**	**(0.17)**	53	(0.84)	**52**
13th 1,000	35	(0.06)	28	(0.44)	28
14th 1,000	18	(0.03)	14	(0.22)	14
15th 1,000	**23**	**(0.04)**	**19**	**(0.30)**	**18**
16th 1,000	17	(0.03)	11	(0.17)	10
17th 1,000	**34**	**(0.06)**	**12**	**(0.19)**	9
18th 1,000	7	(0.01)	7	(0.11)	7
19th 1,000	**11**	**(0.02)**	**9**	**(0.14)**	**9**
20th 1,000	5	(0.01)	5	(0.08)	4
21st 1,000	**10**	**(0.02)**	**9**	**(0.14)**	**9**
22nd 1,000	5	(0.01)	4	(0.06)	4
23rd 1,000	**12**	**(0.02)**	**8**	**(0.13)**	**8**
24th 1,000	5	(0.01)	3	(0.05)	3
25th 1,000	**30**	**(0.05)**	**6**	**(0.09)**	**6**
Proper nouns	2,152	(3.61)	344	(5.43)	320
Exclamations	71	(0.12)	23	(0.36)	18
Transparent compounds	450	(0.76)	183	(2.89)	157
Abbreviations	11	(0.02)	8	(0.13)	8
Not in the lists	521	(0.87)	184	(2.90)	
Total	59,559		6,335		4,076

Table 11.3: Lexical anomalies in *Deadeye Dick* pursuant to the *OED2*

cioccolata (3), *skeedee* (3), *beedy* (2), *deedly* (2), *dooby* (2), *foodily* (2), *foodly* (2), *spuma* (2), ***beebee***, ***bodey***, *dohs*, ***dop***, *faaaaaaaaaaaaaaaaa*, *feedily*, *gazoolian*, *reepa*, *skaddy*, *skeedy*, *wahs*, *zang*
Total: 20 types (30 tokens)

69.13% of every family at the second GSL level, and (c) 11.68% of every type and 44.82% of every family at the AWL level.

In Table 11.2, the high-frequency superset contains 24 $^{3}/_{22}$ times (95.86%) more tokens, 3 $^{41}/_{254}$ times (68.37%) more types, and 1 $^{1050}/_{1117}$ times (48.45%) more families than that of the mid-frequency levels, plus 107 $^{8}/_{167}$ times (99.07%) more tokens, 13 $^{1}/_{28}$ times (92.33%) more types, and 7 $^{144}/_{289}$ times (86.66%) more families than the low-frequency superset. The mid-frequency superset holds 4 $^{218}/_{501}$ times (77.45%) more tokens, 4 $^{19}/_{154}$ times (75.75%) more types, and 3 $^{250}/_{289}$ times (74.13%) more families than the low-frequency levels. The text covers (a) 21.01% of every type and 72.23% of every family in the high-frequency vocabulary, (b) 5.39% of every type and 18.62% of every family in the mid-frequency vocabulary, as well as (c) 0.93% of every type and 1.81% of every family in the low-frequency vocabulary.

The predominant types of proper nouns and subsumed derivatives extrinsic to their designated level of the BNC/COCA subgroup are *Fortune* (L2: 37 tokens), *Hoover* (L7: 35 tokens), *Waltz* (L8: 46 tokens), *Lys* (L22: 2 tokens), *Geiger* (L23: 3 tokens), *Austro-* (L24: 3 tokens), and *Deadeye* (L25: 20 tokens). Among the elements unclassifiable in Table 11.3, those highlighted as homographs of certain base forms in the *OED2* include two instances of what the text identifies as scat-singing: *bodey* and *dop*. Conversely, the lexical elements adjudged to be classifiable amount to 491 tokens, as illustrated in Appendix J, translating into a margin of error of 0.82%. The addition of the exact figure to the combined number

of proper nouns and marginal words in the lists that are assumed to be easily understood yields an adjusted total of 56,384 tokens. Consequently, this text reaches 98.11% coverage with a vocabulary of merely 6,000 word families.

The preceding text's predilection for objectively frequent types does not recrudesce in *Deadeye Dick*. The proportion of the second GSL level exhibited a 0.46% rise in tokens coinciding with a fall of up to 3.58% in the rest, while the AWL fell by as much as 2.39%. The relationship of high- to mid-frequency elements was marked by a rise of up to 8.41%, although a mere 0.1% in tokens. The proportion of high- to low-frequency elements fell by 0.06% in tokens while rising by up to 3.74% in types and families, whereas that of mid- to low-frequency elements fell by 1.9% in tokens while rising by up to 2.61% in types and families. The coverage of types and families went down by 5.08% and 4.88% for the GSL, by 4.22% and 9.31% for the AWL, and by 1.63% and 2.42% for the BNC/COCA subgroup.

Galápagos: A Novel

Table 12.1 shows that, of the two GSL levels, the first one receives 13 $^{604}/_{801}$ times as many tokens, 1 $^{911}/_{1251}$ times as many types, and 1 $^{68}/_{245}$ times as many families, equating to a difference of 92.73%, 42.14%, and 21.73% respectively. Relative to the AWL level, the conjoint lists of general words are in excess of 56 $^{201}/_{524}$ (98.23%) of the tokens, 6 $^{455}/_{493}$ (85.56%) of the types, and 5 $^{43}/_{103}$ (81.54%) of the families. The text covers (a) 52.55% of every type and 94.09% of every family at the first GSL level, (b) 33.74% of every type and 74.39% of every family at the second GSL level, and (c) 16% of every type and 54.31% of every family at the AWL level.

In Table 12.2, the high-frequency superset contains 22 $^{328}/_{917}$ times (95.53%) more tokens, 2 $^{454}/_{529}$ times (65.01%) more types, and 1 $^{1026}/_{1301}$ times (44.09%) more families than

Table 12.1: Analysis of *Galápagos* by the GSL/AWL lists

File	Tokens	(%)	Types	(%)	Families
1_gsl_1st_1000.txt	55,085	(79.98)	2,162	(29.98)	939
2_gsl_2nd_1000.txt	4,005	(5.82)	1,251	(17.35)	735
3_awl_570.txt	1,048	(1.52)	493	(6.84)	309
	8,733	(12.68)	3,306	(45.84)	
Total	68,871		7,212		1,983

Table 12.2: Analysis of *Galápagos* by the BNC/COCA lists

Level	Tokens	(%)	Types	(%)	Families
1st 1,000	55,587	(80.71)	2,286	(31.70)	955
2nd 1,000	4,209	(6.11)	1,415	(19.62)	809
3rd 1,000	1,710	(2.48)	835	(11.58)	563
4th 1,000	1,002	(1.45)	538	(7.46)	394
5th 1,000	644	(0.94)	371	(5.14)	316
6th 1,000	431	(0.63)	237	(3.29)	200
7th 1,000	273	(0.40)	181	(2.51)	158
8th 1,000	249	(0.36)	152	(2.11)	134
9th 1,000	152	(0.22)	108	(1.50)	99
10th 1,000	149	(0.22)	98	(1.36)	89
11th 1,000	137	(0.20)	74	(1.03)	66
12th 1,000	131	(0.19)	68	(0.94)	60
13th 1,000	78	(0.11)	36	(0.50)	31
14th 1,000	28	(0.04)	20	(0.28)	20
15th 1,000	**30**	**(0.04)**	**21**	**(0.29)**	19
16th 1,000	26	(0.04)	18	(0.25)	16
17th 1,000	18	(0.03)	14	(0.19)	14
18th 1,000	**30**	**(0.04)**	5	(0.07)	5
19th 1,000	11	(0.02)	**10**	**(0.14)**	**10**
20th 1,000	**37**	**(0.05)**	6	(0.08)	5
21st 1,000	5	(0.01)	4	(0.06)	4
22nd 1,000	**14**	**(0.02)**	**5**	**(0.07)**	**5**
23rd 1,000	2	(0.00)	2	(0.03)	2
24th 1,000	**2**	**(0.00)**	**2**	**(0.03)**	**2**
25th 1,000	1	(0.00)	1	(0.01)	1
Proper nouns	2,462	(3.57)	335	(4.65)	304
Exclamations	57	(0.08)	18	(0.25)	14
Transparent compounds	364	(0.53)	163	(2.26)	155
Abbreviations	5	(0.01)	3	(0.04)	3
Not in the lists	1,027	(1.49)	186	(2.58)	
Total	68,871		7,212		4,453

Table 12.3: Lexical anomalies in *Galápagos* pursuant to the *OED2*

dagonite (6), *difficilis* (5), *geospiza* (5), *glacco* (3), *-zakh* (2), *bustin'*, *desmodontidae*, *geht*, *Ihnen*, *Kaaaaaaa-*, *Kaaaaaaaa-*, *transylvaniensis*
Total: 12 types (28 tokens)

that of the mid-frequency levels, plus 87 $^{231}/_{233}$ times (98.86%) more tokens, 11 $^{13}/_{16}$ times (91.53%) more types, and 6 $^{233}/_{349}$ times (85%) more families than the low-frequency superset. The mid-frequency superset holds 3 $^{218}/_{233}$ times (74.59%) more tokens, 4 $^{17}/_{128}$ times (75.8%) more types, and 3 $^{254}/_{349}$ times (73.17%) more families than the low-frequency levels. The text covers (a) 23.74% of every type and 77.57% of every family in the high-frequency vocabulary, (b) 6.74% of every type and 21.68% of every family in the mid-frequency vocabulary, as well as (c) 1.16% of every type and 2.18% of every family in the low-frequency vocabulary.

The leading types treated as proper nouns in their context despite their occurrence at other levels of the BNC/COCA subgroup are *Bobby* (L11: 15 tokens), *Kamikaze* (L13: 15 tokens), *Ilium* (L18: 23 tokens), *Bouvier* (L24: 1 token), and *Methuselah* (L25: 1 token). As regards the unclassifiable elements shown in Table 12.3, the consultation of the tertiary source for this purpose, as a follow-up to the computer-assisted analysis, did not encounter widely divergent meanings of homographs or ostensible paronyms of bound and free forms contained in the dictionary entries. The 999 additional tokens comprised by the types that permit successful identification are presented in Appendix K. These represent a margin of error of 1.45%, which, in combination with the proper nouns and marginal words found at their respective levels, leaves 64,984 tokens for adjusted analysis. The resultant size of the text attains 98.26% coverage with 7,000 word families in view of the fact that a vocabulary of 6,000 would be limited to 97.84%.

In general, *Galápagos* increased the proportion of objectively infrequent elements and, by extension, the complexity of the vocabulary, with the percentage up between 0.57 and 4.95 points for the second GSL level in relation to its first one, plus between 0.47 and 3.03 points for the AWL in relation to the GSL. At the BNC/COCA levels, the proportion was reduced between 0.33 and 4.36 percentage points for high- to mid-frequency elements, as well as between 0.2 and 1.66 percentage points for high- to low-frequency elements. On the other hand, the percentage of mid- to low-frequency elements was raised by 0.06 points in types despite a reduction of 2.86 points in tokens and 0.95 points in families. The overall coverage of types and families increased by 3.32% and 3.52% for the GSL, by 4.32% and 9.49% for the AWL, and by 1.21% and 1.62% for the foregoing lists from the BNC/COCA.

Bluebeard, the Autobiography of Rabo Karabekian (1916–1988)

Table 13.1 shows that, of the two GSL levels, the first one receives 17 $^{2643}/_{3289}$ times as many tokens, 1 $^{1057}/_{1153}$ times as many types, and 1 $^{77}/_{234}$ times as many families, equating to a difference of 94.38%, 47.83%, and 24.76% respectively. Relative to the AWL level, the conjoint lists of general words are in excess of 63 $^{3}/_{28}$ (98.42%) of the tokens, 7 $^{143}/_{460}$ (86.32%) of the types, and 5 $^{40}/_{101}$ (81.47%) of the families. The text covers (a) 53.72% of every type and 93.49% of every family at the first GSL level, (b) 31.09% of every type and 71.05% of every family at the second GSL level, and (c) 14.93% of every type and 53.25% of every family at the AWL level.

In Table 13.2, the high-frequency superset contains 25 $^{2397}/_{2461}$ times (96.15%) more tokens, 3 $^{185}/_{1394}$ times (68.08%) more types, and 1 $^{1072}/_{1181}$ times (47.58%) more families than that of the mid-frequency levels, plus 113 $^{208}/_{281}$ times (99.12%) more tokens, 12 $^{95}/_{356}$ times (91.85%) more types, and 6 $^{303}/_{325}$ times (85.57%) more families than the low-

Table 13.1: Analysis of *Bluebeard* by the GSL/AWL lists

File	Tokens	(%)	Types	(%)	Families
1_gsl_1st_1000.txt	58,556	(83.39)	2,210	(31.63)	933
2_gsl_2nd_1000.txt	3,289	(4.68)	1,153	(16.50)	702
3_awl_570.txt	980	(1.40)	460	(6.58)	303
	7,393	(10.53)	3,164	(45.28)	
Total	70,218		6,987		1,938

Table 13.2: Analysis of *Bluebeard* by the BNC/COCA lists

Level	Tokens	(%)	Types	(%)	Families
1st 1,000	59,151	(84.24)	2,272	(32.52)	951
2nd 1,000	3,234	(4.61)	1,313	(18.79)	761
3rd 1,000	1,537	(2.19)	782	(11.19)	541
4th 1,000	920	(1.31)	444	(6.35)	357
5th 1,000	617	(0.88)	334	(4.78)	277
6th 1,000	324	(0.46)	218	(3.12)	188
7th 1,000	269	(0.38)	167	(2.39)	151
8th 1,000	178	(0.25)	123	(1.76)	111
9th 1,000	153	(0.22)	108	(1.55)	97
10th 1,000	114	(0.16)	83	(1.19)	76
11th 1,000	94	(0.13)	72	(1.03)	65
12th 1,000	58	(0.08)	47	(0.67)	45
13th 1,000	56	(0.08)	34	(0.49)	31
14th 1,000	25	(0.04)	22	(0.31)	18
15th 1,000	**36**	**(0.05)**	**29**	**(0.42)**	**27**
16th 1,000	**53**	**(0.08)**	19	(0.27)	17
17th 1,000	14	(0.02)	6	(0.09)	6
18th 1,000	**31**	**(0.04)**	**7**	**(0.10)**	**7**
19th 1,000	**36**	**(0.05)**	**11**	**(0.16)**	**9**
20th 1,000	9	(0.01)	5	(0.07)	5
21st 1,000	**12**	**(0.02)**	**6**	**(0.09)**	**5**
22nd 1,000	8	(0.01)	5	(0.07)	**5**
23rd 1,000	5	(0.01)	4	(0.06)	4
24th 1,000	1	(0.00)	1	(0.01)	1
25th 1,000	**10**	**(0.01)**	**5**	**(0.07)**	**4**
Proper nouns	2,180	(3.10)	410	(5.87)	369
Exclamations	119	(0.17)	31	(0.44)	24
Transparent compounds	349	(0.50)	198	(2.83)	179
Abbreviations	6	(0.01)	3	(0.04)	3
Not in the lists	619	(0.88)	228	(3.26)	
Total	70,218		6,987		4,334

Table 13.3: Lexical anomalies in *Bluebeard* pursuant to the *OED2*

***-bek-* (2)**, *cie* (2), *floparroo* (2), *frères* (2), ***kar-* (2)**, *chilluns*, *chlinel*, *linel*
Total: 8 types (13 tokens)

frequency superset. The mid-frequency superset holds 4 $^{213}/_{562}$ times (77.16%) more tokens, 3 $^{163}/_{178}$ times (74.46%) more types, and 3 $^{206}/_{325}$ times (72.48%) more families than the low-frequency levels. The text covers (a) 22.86% of every type and 75.1% of every family in the high-frequency vocabulary, (b) 5.92% of every type and 19.68% of every family in the mid-frequency vocabulary, as well as (c) 1.08% of every type and 2.03% of every family in the low-frequency vocabulary.

The leading types of proper nouns and subsumed derivatives outside the designated BNC/COCA list are *Gypsies* (L6: 11 tokens), *Expressionist* (L9: 12 tokens), *-Luxe* (L16: 21 tokens), *Dura-* (L18: 22 tokens), *Sateen* (L19: 22 tokens), *Coulomb* (L21: 6 tokens), and *Al* (Abbreviations: 4 tokens). The only types among those deemed unclassifiable in Table 13.3 that should not be erroneously conflated with their homographs in the *OED2* pertain to the syllabification of the Armenian last name of Vonnegut's protagonist and its nonphonemic transcription. The 606 tokens in Appendix L are ascertained to warrant classification, hence resulting in a margin of error of 0.86%. Their incorporation into the aggregate of what are supposed to be readily cognizant proper nouns and marginal words brings the final total to 66,958 tokens. The cumulative total for the levels of this control subgroup satisfies 98.25% coverage with 6,000 word families.

In respect of tokens, the proportion of the first GSL level to its second one and the high- to the low-frequency levels of the BNC/COCA subgroup reached an all-time high in *Bluebeard*. There was a net loss of between 1.65 and 5.69 points for the second GSL level,

a 0.07-point gain for families in the AWL, but a loss of up to 0.77 points for the rest. The gain in percentage was between 0.62 and 3.49 points for high- to mid-frequency elements, between 0.26 and 0.57 points for high- to low-frequency elements, and 2.57 points for mid- to low-frequency tokens, with types and families losing 1.34 and 0.69 points respectively. A rise of 1.17% in the coverage of types for the first GSL level was outweighed by a fall of 0.64% in types and 1.96% in families for the GSL, of 1.07% in types and 1.05% in families for the AWL, and of 0.52% in types and 0.87% in families for the graded BNC/COCA lists.

Hocus Pocus

Table 14.1 shows that, of the two GSL levels, the first one receives 16 $^{1707}/_{2011}$ times as many tokens, 1 $^{1055}/_{1203}$ times as many types, and 1 $^{214}/_{719}$ times as many families, equating to a difference of 94.06%, 46.72%, and 22.94% respectively. Relative to the AWL level, the conjoint lists of general words are in excess of 52 $^{7}/_{345}$ (98.08%) of the tokens, 6 $^{209}/_{542}$ (84.34%) of the types, and 4 $^{150}/_{169}$ (79.54%) of the families. The text covers (a) 54.89% of every type and 93.49% of every family at the first GSL level, (b) 32.44% of every type and 72.77% of every family at the second GSL level, and (c) 17.59% of every type and 59.4% of every family at the AWL level.

In Table 14.2, the high-frequency superset contains 22 $^{2533}/_{3289}$ times (95.61%) more tokens, 2 $^{1273}/_{1700}$ times (63.62%) more types, and 1 $^{911}/_{1430}$ times (38.91%) more families than that of the mid-frequency levels, plus 110 $^{311}/_{678}$ times (99.09%) more tokens, 10 $^{403}/_{427}$ times (90.86%) more types, and 5 $^{331}/_{402}$ times (82.83%) more families than the low-frequency superset. The mid-frequency superset holds 4 $^{577}/_{678}$ times (79.39%) more tokens, 3 $^{419}/_{427}$ times (74.88%) more types, and 3 $^{112}/_{201}$ times (71.89%) more families than the low-frequency levels. The text covers (a) 24.46% of every type and 78.03% of every family in

Table 14.1: Analysis of *Hocus Pocus* by the GSL/AWL lists

File	Tokens	(%)	Types	(%)	Families
1_gsl_1st_1000.txt	67,766	(82.17)	2,258	(29.00)	933
2_gsl_2nd_1000.txt	4,022	(4.88)	1,203	(15.45)	719
3_awl_570.txt	1,380	(1.67)	542	(6.96)	338
	9,305	(11.28)	3,782	(48.58)	
Total	82,473		7,785		1,990

Table 14.2: Analysis of *Hocus Pocus* by the BNC/COCA lists

Level	Tokens	(%)	Types	(%)	Families
1st 1,000	68,600	(83.18)	2,346	(30.13)	953
2nd 1,000	4,283	(5.19)	1,431	(18.38)	808
3rd 1,000	2,008	(2.43)	896	(11.51)	580
4th 1,000	1,244	(1.51)	579	(7.44)	448
5th 1,000	718	(0.87)	397	(5.10)	335
6th 1,000	545	(0.66)	290	(3.73)	258
7th 1,000	351	(0.43)	192	(2.47)	168
8th 1,000	228	(0.28)	138	(1.77)	125
9th 1,000	203	(0.25)	104	(1.34)	96
10th 1,000	150	(0.18)	102	(1.31)	93
11th 1,000	116	(0.14)	75	(0.96)	72
12th 1,000	95	(0.12)	62	(0.80)	58
13th 1,000	71	(0.09)	45	(0.58)	44
14th 1,000	42	(0.05)	22	(0.28)	21
15th 1,000	29	(0.04)	**25**	**(0.32)**	**24**
16th 1,000	**34**	**(0.04)**	17	(0.22)	17
17th 1,000	27	(0.03)	**19**	**(0.24)**	**17**
18th 1,000	**42**	**(0.05)**	17	(0.22)	15
19th 1,000	21	(0.03)	13	(0.17)	13
20th 1,000	17	(0.02)	12	(0.15)	11
21st 1,000	8	(0.01)	5	(0.06)	5
22nd 1,000	**16**	**(0.02)**	**8**	**(0.10)**	**7**
23rd 1,000	6	(0.01)	2	(0.03)	2
24th 1,000	3	(0.00)	**2**	**(0.03)**	**2**
25th 1,000	1	(0.00)	1	(0.01)	1
Proper nouns	2,546	(3.09)	525	(6.74)	482
Exclamations	65	(0.08)	21	(0.27)	15
Transparent compounds	355	(0.43)	221	(2.84)	202
Abbreviations	13	(0.02)	8	(0.10)	8
Not in the lists	636	(0.77)	210	(2.70)	
Total	82,473		7,785		4,880

Table 14.3: Lexical anomalies in *Hocus Pocus* pursuant to the *OED2*

Sohn (3), ***whay* (2)**, ***brazzle***, *crurifragium*, *higgies*, *mogies*, ***puh-***
Total: 7 types (10 tokens)

the high-frequency vocabulary, (b) 7.22% of every type and 23.83% of every family in the mid-frequency vocabulary, as well as (c) 1.29% of every type and 2.51% of every family in the low-frequency vocabulary.

The predominant types of proper nouns and subsumed derivatives extrinsic to their designated level of the BNC/COCA subgroup are *Musket* (L9: 22 tokens), *Debs* (L12: 15 tokens), *Quadrangle* (L14: 15 tokens), *Earthlings* (L15: 3 tokens), *Appian* (L20: 3 tokens), *Greco* (L25: 1 token), *C.* (Exclamations: 8 tokens), and *PB* (Abbreviations: 4 tokens). The types among those unclassifiable in Table 14.3 highlighted on account of their resemblance to specific base forms in the *OED2* are all instances of pronunciation-related transcriptions in their context. Collectively, the 626 ideally classifiable tokens in Appendix M represent a 0.76% margin of error. With the addition of the proper nouns and marginal words at their levels, 78,868 of the tokens remain indispensable for an adjusted analysis of this text. The cumulative coverage of the final total reveals that 98.14% corresponds to a vocabulary of 6,000 word families.

In contrast to the antecedent work immediately preceding it, *Hocus Pocus* indicated a reduction of between 0.32 and 1.82 percentage points for the first GSL level in relation to its second one, plus of between 0.34 and 1.98 percentage points for the two GSL levels in relation to the AWL one. From the perspective of the ternary chain of the BNC/COCA lists, the percentage fell between 0.54 and 8.67 points for the high- to the mid-frequency levels, between 0.03 and 2.75 points for the high- to the low-frequency levels, and by 0.59 points

for mid- to low-frequency families, with their tokens and types having risen by 2.22 and 0.42 points. The coverage of the control group produced a net gain of 1.25% in types and 0.86% in families for the GSL, of 2.66% in types and 6.15% in families for the AWL, and of 0.9% in types and 1.66% in families for the relevant sequence of the BNC/COCA levels.

Timequake

Table 15.1 shows that, of the two GSL levels, the first one receives 13 $^{2739}/_{2753}$ times as many tokens, 1 $^{192}/_{217}$ times as many types, and 1 $^{256}/_{675}$ times as many families, equating to a difference of 92.85%, 46.94%, and 27.5% respectively. Relative to the AWL level, the conjoint lists of general words are in excess of 41 $^{527}/_{994}$ (97.59%) of the tokens, 6 $^{56}/_{95}$ (84.82%) of the types, and 5 $^{91}/_{303}$ (81.13%) of the families. The text covers (a) 49.71% of every type and 93.29% of every family at the first GSL level, (b) 29.26% of every type and 68.32% of every family at the second GSL level, and (c) 15.41% of every type and 53.25% of every family at the AWL level.

In Table 15.2, the high-frequency superset contains 17 $^{1579}/_{2466}$ times (94.33%) more tokens, 2 $^{1335}/_{1396}$ times (66.17%) more types, and 1 $^{1033}/_{1217}$ times (45.91%) more families than that of the mid-frequency levels, plus 67 $^{621}/_{640}$ times (98.53%) more tokens, 9 $^{401}/_{414}$ times (89.97%) more types, and 5 $^{95}/_{131}$ times (82.53%) more families than the low-frequency superset. The mid-frequency superset holds 3 $^{273}/_{320}$ times (74.05%) more tokens, 3 $^{77}/_{207}$ times (70.34%) more types, and 3 $^{38}/_{393}$ times (67.71%) more families than the low-frequency levels. The text covers (a) 21.6% of every type and 75% of every family in the high-frequency vocabulary, (b) 5.93% of every type and 20.28% of every family in the mid-frequency vocabulary, as well as (c) 1.25% of every type and 2.46% of every family in the low-frequency vocabulary.

Table 15.1: Analysis of *Timequake* by the GSL/AWL lists

File	Tokens	(%)	Types	(%)	Families
1_gsl_1st_1000.txt	38,528	(77.98)	2,045	(29.40)	931
2_gsl_2nd_1000.txt	2,753	(5.57)	1,085	(15.60)	675
3_awl_570.txt	994	(2.01)	475	(6.83)	303
	7,135	(14.44)	3,351	(48.17)	
Total	49,410		6,956		1,909

Table 15.2: Analysis of *Timequake* by the BNC/COCA lists

Level	Tokens	(%)	Types	(%)	Families
1st 1,000	39,119	(79.17)	2,148	(30.88)	961
2nd 1,000	2,834	(5.74)	1,213	(17.44)	739
3rd 1,000	1,548	(3.13)	766	(11.01)	550
4th 1,000	768	(1.55)	474	(6.81)	379
5th 1,000	729	(1.48)	302	(4.34)	263
6th 1,000	365	(0.74)	220	(3.16)	201
7th 1,000	254	(0.51)	163	(2.34)	152
8th 1,000	182	(0.37)	130	(1.87)	121
9th 1,000	168	(0.34)	107	(1.54)	101
10th 1,000	135	(0.27)	96	(1.38)	89
11th 1,000	117	(0.24)	73	(1.05)	68
12th 1,000	78	(0.16)	53	(0.76)	52
13th 1,000	**90**	**(0.18)**	41	(0.59)	38
14th 1,000	41	(0.08)	27	(0.39)	26
15th 1,000	34	(0.07)	25	(0.36)	23
16th 1,000	**36**	**(0.07)**	**26**	**(0.37)**	**25**
17th 1,000	18	(0.04)	15	(0.22)	15
18th 1,000	17	(0.03)	8	(0.12)	8
19th 1,000	**18**	**(0.04)**	**12**	**(0.17)**	**12**
20th 1,000	12	(0.02)	11	(0.16)	11
21st 1,000	**19**	**(0.04)**	10	(0.14)	9
22nd 1,000	9	(0.02)	5	(0.07)	5
23rd 1,000	2	(0.00)	2	(0.03)	2
24th 1,000	**7**	**(0.01)**	**4**	**(0.06)**	**4**
25th 1,000	**7**	**(0.01)**	**6**	**(0.09)**	**6**
Proper nouns	1,776	(3.59)	540	(7.76)	517
Exclamations	137	(0.28)	31	(0.45)	24
Transparent compounds	269	(0.54)	156	(2.24)	148
Abbreviations	24	(0.05)	14	(0.20)	13
Not in the lists	597	(1.21)	278	(4.00)	
Total	49,410		6,956		4,562

Table 15.3: Lexical anomalies in *Timequake* pursuant to the *OED2*

boop- (2), *karass* (2), *-uck* (2), *buckareenies*, *cultiver*, *-doop*, *fiducia*, *glurp*, *glurps*, *hokay*, *plumbum*, *squoozled*, *squoozles*, *Wiedersehen*, ***youp-***
Total: 15 types (18 tokens)

The predominant types of proper nouns in different BNC/COCA lists are *Eve* (L4: 14 tokens), *Trout* (L5: 261 tokens), *Booth* (L6: 16 tokens), *Bingo* (L7: 12 tokens), *Satan* (L8: 9 tokens), *Schadenfreude* (L18: 5 tokens), *Zine* (L19: 4 tokens), *Rube* (L20: 2 tokens), *Mensa/Petro* (L22: 3 tokens each), *Palladio* (L24: 4 tokens), and *Phi* (Abbreviations: 4 tokens). The unclassifiable elements in Table 15.3 include a single occurrence of an *OED2* homograph, which its context indicates to consist in an alteration for poetic effect. The 579 tokens in Appendix N are adjudged to contrast with their counterparts in the former table in orthographical terms, contributing a margin of error of 1.17%. The addition of this figure to the aggregate quantity of proper nouns and marginal words at their designated levels for implicit inclusion translates into the retention of 46,625 of the tokens. These allow 98.23% coverage with 8,000 word families, whereas a vocabulary of 7,000 would arrive at 97.84%.

In comparison to *Hocus Pocus*, this evolution was finalized with a 1.21-percentage-point rise in tokens that coincided with an up to 4.56-percentage-point fall in their types and families for the second GSL level, and with a 0.49-percentage-point rise in tokens that coincided with an up to 1.59-percentage-point fall in the rest for the AWL. The percentage fell by 1.28 points in tokens, but rose by up to 7 points in the rest for the high- to the mid-frequency levels. It fell by up to 0.89 points for high- to low-frequency elements and by up to 5.34 points for mid- to low-frequency ones; in the latter case, to an all-time low in types and families. The coverage of all types and families fell by 4.23% and 2.32% for the GSL, by 2.17% and 6.15% for the AWL, and by 1.14% and 1.25% for the BNC/COCA subgroup.

CONCLUSION

This thesis set out to collect word-frequency data from the novels of Vonnegut in order that their collation would produce a tripartite set of facts: (a) the vocabulary size that is a prerequisite for adequate reading comprehension, as effectuated by the attainment of the 98% lexical threshold; (b) the number of instances whereby the distribution of lexical elements across the contiguous levels of the control group negates the expected pattern of decrease; (c) the margin of error that signifies the extent to which the omission of elements from the word-family lists has been confused with the absence of dictionary definitions. In addition, the introductory chapter delineated the conceptual framework in which the study is embedded. “Details pertaining to the corpora” delved deeper into the theoretical aspects of the approach, whereas “Sample group analysis and control group validation” presented the bulk of empirical evidence, providing a corpus-linguistic angle on the literary creations.

The answers to the first and the third part of the research problem are recapitulated in the form of Table 16.1, while Table 16.2 attempts to address the second part. The former table demonstrates over two-thirds of the texts to be below the estimated level of difficulty. Although there is no straightforward correlation between these figures and Vonnegut’s self-assigned grades, developmental trends, whether positive or negative, may not be devoid of recondite significance in this regard. As to the latter of these tables, since the conflation of common and proper nouns was revealed to account for several of the marked discrepancies in tokens, the conclusion is drawn from the data on types and families alone; for example, the 19th and the 25th 1,000-word-family list either equaled or surpassed its antecedent in the majority of the texts. It is hereby reiterated that the findings accord with Nation’s theories underlying his vocabulary size tests. A logical extension of this line of research would be to refine the scope and specificity of a project, as well as to systematize the appended lexicon.

Table 16.1: Analysis of Vonnegut's novels by the BNC/COCA lists and the *OED2*

Title	Unadjusted text size	Margin of error	Vocabulary size	Adjusted coverage
Player Piano	103,464	1.50%	9,000	98.21%
The Sirens of Titan	77,570	2.78%	7,000	98.01%
Mother Night	50,226	1.11%	7,000	98.32%
Cat's Cradle	54,406	1.38%	8,000	98.17%
God Bless You, Mr. Rosewater	49,981	1.39%	7,000	98.08%
Slaughterhouse-Five	50,851	0.83%	7,000	98.00%
Breakfast of Champions	59,436	1.22%	7,000	98.27%
Slapstick	38,752	0.77%	8,000	98.36%
Jailbird	74,685	1.15%	6,000	98.09%
Deadeye Dick	59,559	0.82%	6,000	98.11%
Galápagos	68,871	1.45%	7,000	98.26%
Bluebeard	70,218	0.86%	6,000	98.25%
Hocus Pocus	82,473	0.76%	6,000	98.14%
Timequake	49,410	1.17%	8,000	98.23%

Table 16.2: Analysis of the BNC/COCA lists by Vonnegut's active vocabulary

Title	Frequency levels with a surfeit of types	Frequency levels with a surfeit of families
Player Piano	18, 22, **25**	18, 22, **25**
The Sirens of Titan	14, **19**, 20, 23, **25**	14, 16, **19**, 20, 23, **25**
Mother Night	15, 17, 22, 23, **25**	15, 17, 22, 23, **25**
Cat's Cradle	11, 13, 18, 20, 24, **25**	11, 13, 18, 20, 24, **25**
God Bless You, Mr. Rosewater	18, **19**, 20, 23, **25**	18, **19**, 20, 23, **25**
Slaughterhouse-Five	15, 18, **19**, 21, 24	15, 18, **19**, 21, 24
Breakfast of Champions	**19**, 21, 23, **25**	**19**, 21, 23, **25**
Slapstick	14, 18, 20, **25**	14, 18, 20, **25**
Jailbird	14, 17, **19**, 21, 24	14, 17, **19**, 21, 24
Deadeye Dick	15, 17, **19**, 21, 23, **25**	12, 15, **19**, 21, 23, **25**
Galápagos	15, **19**, 22, 24	**19**, 22, 24
Bluebeard	15, 18, **19**, 21, **25**	15, 18, **19**, 21, 22, **25**
Hocus Pocus	15, 17, 22, 24	15, 17, 22, 24
Timequake	16, **19**, 24, **25**	16, **19**, 24, **25**

In the manner of reader-response criticism, then, whether or not the outcome of this interdisciplinary research effort is of use to an individual reader will ultimately depend on whether that individual finds it useful to employ a methodology that allows early detection

of every word form in, for instance, a writer-specific genre of works. To put it another way, the analyst has made a case for the practicality of a freeware vocabulary-profiling program with default reference lists and the practicability of integrating vocabulary profiling with literary criticism, emphasizing the centrality of the idiolect of a novelist of one's preference and calling attention to the perceived benefits of gaining total access to the surface text of every piece of writing. The reader is left to infer, from the data output by AntWordProfiler, the desirability of being, at any stage of reading, aware of the form and the frequency of all the words encountered in a text and, by extension, being able to construct their contextual meaning as a direct consequence of knowing the exact location of their occurrence in that text. It is implied that the teaching of literature to adult learners of English can benefit from the acquaintance with the vocabulary essential to one's comprehension of Vonnegut's style.

In furtherance of future research objectives, the linguistic facet of this project has contributed to the value of its literary dimension in connection with specificity by readying an idiosyncratic corpus and an exhaustive lexicon that could conceivably enable the analyst to perform a close reading and an appraisal of the concordance lines of each of the nodes. Accordingly, scope could vary from a single token to all available writings of Vonnegut, as well as extend to other authors. This analyst is of the opinion that the optimum meaning of each word in the author's vocabulary is inextricably linked with discrete linguistic entities. For this reason, it is proposed that AntWordProfiler's capability of divorcing every vestige of out-of-context meaning from word types and families be considered as a step in the right direction, namely toward the process of constructing meanings in concert with the reader and discourses of reading, which no automatic lemmatization could simulate, anyway. The margins of error in Table 16.1, the whole of Table 16.2, and the inevitable obsolescence of the tools call for revised lists, possibly with the relevant pseudowords, word fragments, etc.

REFERENCES

Primary sources

Vonnegut, Kurt. 1952/2009. *Player Piano*. New York: Dial Press.

Vonnegut, Kurt. 1959/2007. *The Sirens of Titan*. New York: Dial Press.

Vonnegut, Kurt. 1961/2009. *Mother Night*. New York: Dial Press.

Vonnegut, Kurt. 1963/2009. *Cat's Cradle*. New York: Dial Press.

Vonnegut, Kurt. 1965/1998. *God Bless You, Mr. Rosewater, or Pearls Before Swine*. New York: Delta.

Vonnegut, Kurt. 1969/2009. *Slaughterhouse-Five, or The Children's Crusade: A Duty-Dance with Death*. New York: Dial Press.

Vonnegut, Kurt. 1973/2009. *Breakfast of Champions, or Goodbye Blue Monday*. New York: Dial Press.

Vonnegut, Kurt. 1976/2010. *Slapstick, or Lonesome No More!* New York: Dial Press.

Vonnegut, Kurt. 1979/2010. *Jailbird*. New York: Dial Press.

Vonnegut, Kurt. 1982/2009. *Deadeye Dick*. New York: Dial Press.

Vonnegut, Kurt. 1985/2009. *Galápagos: A Novel*. New York: Dial Press.

Vonnegut, Kurt. 1987/2009. *Bluebeard, the Autobiography of Rabo Karabekian (1916–1988)*. New York: Dial Press.

Vonnegut, Kurt. 1990/1997. *Hocus Pocus*. New York: Berkley Books.

Vonnegut, Kurt. 1997/1998. *Timequake*. New York: Berkley Books.

Secondary sources

Anthony, Laurence. 2013. AntWordProfiler (Version 1.4.0w) [Computer software]. Available from http://www.antlab.sci.waseda.ac.jp/antwordprofiler_index.html, accessed 1 May 2014.

Bauer, Laurie and Paul Nation. 1993. Word families. *International Journal of Lexicography*, 6: 4, 253–279.

British Council. 2012. TeachingEnglish: Word Family Framework. Available at http://www.teachingenglish.org.uk/article/word-family-framework, accessed 1 May 2014.

Browne, Charles. 2013. A new general service vocabulary for 2nd language learners. *The English Connection*, 17: 3, 12–13.

Burnard, Lou (Ed.). 2009a. About the British National Corpus: BNC Products. Available at http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=products, accessed 1 May 2014.

Burnard, Lou (Ed.). 2009b. About the British National Corpus: What is the BNC? Available at http://www.natcorp.ox.ac.uk/corpus/index.xml, accessed 1 May 2014.

Cambridge University Press. 2012. About the English Vocabulary Profile: How has it been created? Available at http://vocabulary.englishprofile.org/staticfiles/about.html, accessed 1 May 2014.

Carver, Ronald P. 1994. Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction. *Journal of Reading Behavior*, 26: 4, 413–437.

Council of Europe. 2011. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly*, 34: 2, 213–238.

Davies, Mark. 2012a. Corpus of Contemporary American English (COCA): Composition of the corpus. Available at http://corpus.byu.edu/coca/help/texts_e.asp, accessed 1 May 2014.

Davies, Mark. 2012b. The Corpus of Contemporary American English (COCA) and the British National Corpus (BNC). Available at http://corpus.byu.edu/coca/compare-bnc.asp, accessed 1 May 2014.

Davies, Mark. 2012c. WordAndPhrase: Downloading data for offline use. Available at http://www.wordandphrase.info/h_download.asp, accessed 1 May 2014.

Davies, Mark. 2012d. WordAndPhrase: Reporting errors. Available at http://www.wordandphrase.info/h_problems.asp, accessed 1 May 2014.

Davis, Todd F. and Kenneth Womack. 2002. *Formalist Criticism and Reader-Response Theory*. Basingstoke and New York: Palgrave Macmillan.

De Geest, Dirk and An Goris. 2010. Constrained writing, creative writing: The case of handbooks for writing romances. *Poetics Today*, 31: 1, 81–106.

Gardner, Dee and Mark Davies. 2013. A new academic vocabulary list. *Applied Linguistics*, Advance Access, 1–24.

Hancioğlu, Nilgün, Steven Neufeld, and John Eldridge. 2008. Through the looking glass and into the land of lexico-grammar. *English for Specific Purposes*, 27: 4, 459–479.

Hobsbawm, Eric. 1995. *The Age of Extremes: The Short Twentieth Century, 1914–1991*. London: Abacus.

Hu, Hsueh-chao Marcella and Paul Nation. 2000. Vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13: 1, 403–430.

Marwick, Arthur. 2012. *The Sixties: Cultural Revolution in Britain, France, Italy, and the United States, c. 1958 – c. 1974*. London: Bloomsbury Reader.

Nation, I. S. P. 2006. How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review / La Revue canadienne des langues vivantes*, 63: 1, 59–82.

Nation, Paul. 2012a. Information on the BNC/COCA lists [PDF document]. Available from http://www.victoria.ac.nz/lals/about/staff/paul-nation, accessed 1 May 2014.

Nation, Paul and Hwang Kyongho. 1995. Where would general service vocabulary stop and special purposes vocabulary begin? *System*, 23: 1, 35–41.

Nation, Paul and Laurence Anthony. 2013. Mid-frequency readers. *Journal of Extensive Reading*, 1: 1, 5–16.

Neufeld, Steve, Nilgün Hancioğlu, and John Eldridge. 2011. Beware the range in Range, and the academic in AWL. *System*, 39: 4, 533–538.

Schmitt, Norbert, Xiangying Jiang, and William Grabe. 2011. The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95: 1, 26–43.

Vonnegut, Kurt. 1981/2009. *Palm Sunday: An Autobiographical Collage*. New York: Dial Press.

Tertiary sources

Billuroğlu, Ali, Steve Neufeld, and Tom Cobb. n.d. VocabProfile (BNL). Available at http://www.lextutor.ca/vp/bnl/, accessed 1 May 2014.

British Council. n.d. Word Family Framework. Available at http://www.learnenglish.org.uk/wff/, accessed 1 May 2014.

Browne, Charles, Brent Culligan, and Joseph Phillips. 2014a. A New Academic Word List. Available at http://www.newacademicwordlist.org/, accessed 1 May 2014.

Browne, Charles, Brent Culligan, and Joseph Phillips. 2014b. A New General Service List (1.01). Available at http://www.newgeneralservicelist.org/, accessed 1 May 2014.

Cambridge University Press. n.d. English Vocabulary Profile. Available at http://www.englishprofile.org/index.php/wordlists, accessed 1 May 2014.

Davies, Mark and Dee Gardner. 2013. Academic Vocabulary Lists (Corpus-based; 120 million words). Available at http://www.academicvocabulary.info/, accessed 1 May 2014.

Heatley, Alex, Paul Nation, and Averil Coxhead. 2004. Range program with GSL/AWL list (Version 1.32) [Computer software]. Available from http://www.victoria.ac.nz/lals/about/staff/paul-nation, accessed 1 May 2014.

Heatley, Alex, Paul Nation, and Averil Coxhead. 2012a. Range program with British National Corpus list 14,000 (Version 1.32) [Computer software]. Available from http://www.victoria.ac.nz/lals/about/staff/paul-nation, accessed 1 May 2014.

Heatley, Alex, Paul Nation, and Averil Coxhead. 2012b. Range program with British National Corpus list 25,000 (Version 1.32H) [Computer software]. Available from http://www.victoria.ac.nz/lals/about/staff/paul-nation, accessed 1 May 2014.

Nation, Paul. 2012b. GSL (1000/2000) and AWL (570) family lists [Data file]. Available from http://www.antlab.sci.waseda.ac.jp/antwordprofiler_index.html, accessed 1 May 2014.

Nation, Paul. 2013. BNC/COCA family lists 1000–25000 + extras [Data file]. Available from http://www.antlab.sci.waseda.ac.jp/antwordprofiler_index.html, accessed 1 May 2014.

Simpson, John (Ed.). 2009. Oxford English Dictionary Second Edition on CD-ROM (Version 4.0.0.3) [Computer software]. Oxford: Oxford University Press.

APPENDICES

Appendix A: 436 new types (1,550 tokens) found in *Player Piano*

OED2 base forms, with inflected forms and derivatives: 279 types (424 tokens)

yessir (35), *thet* (14), *hisself* (9), *latchstring* (6), *barbering* (5), *nossir* (5), *sub-* (5), *antisabotage* (4), *billfold* (4), *dubonnet* (4), *loudmouth* (4), *shoeshine* (4), *breakfront* (3), *bulletproof* (3), *chrissakes* (3), *dollied* (3), *haircutting* (3), *highball* (3), *oldsters* (3), *precipitator* (3), *precipitators* (3), *pretense* (3), *antimachine* (2), *appraisingly* (2), *automagic* (2), *bagpipers* (2), *barnyard* (2), *baseboard* (2), *billers* (2), *bloodlines* (2), *braiders* (2), *chagrined* (2), *checkerpiece* (2), *checkroom* (2), *costumed* (2), *enameled* (2), *engineeringwise* (2), *gadgeteering* (2), *gages* (2), *grownup* (2), *inspirationally* (2), *interfraternity* (2), *invincibility* (2), *jeweled* (2), *maladjusted* (2), *managership* (2), *nationwise* (2), *oligomenorrhea* (2), *perplexedly* (2), *poniard* (2), *socialwise* (2), *somber* (2), *sonofabitching* (2), *suppertime* (2), *treeslaughter* (2), *tuxedos* (2), *voicebox* (2), *wainscoted* (2), *whatchamacallits* (2), *whimsey* (2), *abusively*, *addressers*, *affectionateness*, *aggressions*, *amplidynes*, *analyzers*, *anesthetized*, *antiaircraft*, *anticlimactic*, *antimagnetic*, *antiqued*, *archcriminal*, *archprophet*, *backstop*, *bandmaster*, *bargeman*, *bargemen*, *bearlike*, *becomingly*, *betch*, *bilinguality*, *bloodlettings*, *boltheads*, *breastworks*, *breechclouts*, *broomcorn*, *canners*, *captivatingly*, *celebrators*, *célèbre*, *cellmate*, *chinnings*, *chrissake*, *clickings*, *-coater*, *colorimeters*, *cosmopolite*, *countermove*, *coupla*, *cowshit*, *cruelest*, *cryostats*, *cump-*, *darnedest*, *demarkation*, *demoniacal*, *densitometers*, *devaluate*, *disjointedly*, *disposers*, *dokey*, *donnas*, *downdale*, *draggers*, *dreamworld*, *dynamiting*, *dynamotors*, *ejector*, *emasculating*, *-epauleted*, *-eradicator*, *excretor*, *fagots*, *filers*, *firings*, *fissionable*, *flashbulb*, *footlocker*, *fountainhead*, *freethinker*, *funereally*, *funnypapers*, *futilely*, *gadgeteer*, *gameroom*, *giveth*, *gracile*, *gravelthroat*, *greeters*, *-grenaded*, *groggily*, *gunrack*, *handsomest*, *harmlessness*, *healthful*, *heartwise*, *highballs*, *holdover*, *hoppity*, *hostilely*, *hummings*, *icebox*, *idee*, *inexpressibly*, *ironer*, *ironware*, *irrelevantly*, *keynoter*, *kilted*, *knobbed*, *lackadaisicalness*, *lazaretto*, *leathercraft*, *'lectric*, *leopardskins*, *leprosarium*, *likably*, *locomotor*, *lordy*, *lunchbox*, *manacled*, *managershipwise*, *managerwise*, *marionettelike*, *milkshake*, *mimeo*, *-muddying*, *mysteriousness*, *newy*, *noisemeters*, *oleomargarine*, *opaqued*, *-operationwise*, *oscillographs*, *outa*, *pacifistic*, *-paned*, *panelboards*, *pantomimed*, *pentathol*, *personalitywise*, *pesthouse*, *petshop*, *pigshit*, *pirouetting*, *piteously*, *pityingly*, *placatingly*, *playlet*, *poisonously*, *potentiometers*, *puncho*, *pushbuttons*, *pushups*, *quieted*, *radiophoto*, *rearmost*, *redskin*, *reducers*, *rehashed*, *righter*, *ringbarked*, *rinseless*, *roadforks*, *rockwool*, *roisterers*, *roughhousing*, *screamingly*, *'scuse*, *selsyn*, *selsyns*, *semper*, *sexologists*, *shadowbox*, *shorthanded*, *shoulda*, *shoveling*, *shovelman*, *sidelight*, *signaled*, *snuffbox*, *sobber*, *socked*, *songbook*, *sooted*, *sorters*, *spectrogoniometers*, *spiraea*, *squalled*, *squalling*, *squeegeed*, *squirto*, *stampeding*, *stickup*, *straingages*, *streetcorner*, *stupifying*, *superbrains*, *surlily*, *tailstocks*, *taketh*, *telemetering*, *telemeters*, *televiewer*, *tentmate*, *'thout*, *torquemeters*, *transitors*, *trapshooting*, *trenchfoot*, *tyrannis*, *unclenching*, *undedicated*, *undercoater*, *unloosed*, *unmachine-*, *unshocked*, *unwatched*, *vaultlike*, *vestigal*, *viscosimeters*, *waitful*, *washless*, *Weltschmerz*, *willfully*, *woodsmen*, *workshirt*, *zigzagging*, *zymometers*

Proper nouns, with inflected forms and derivatives: 101 types (1,027 tokens)

Finnerty (309), *Lasher* (135), *Khashdrahr* (97), *Berringer* (54), *Gelhorne* (43), *Haycox* (43), *Alfy* (42), *EPICAC* (35), *Takaru* (34), *Hacketts* (20), *Bratpuhr* (16), *Calhoun* (14), *Gottwald* (12), *Hagstrohm* (11), *Averageman* (9), *Sumklish* (9), *MacCleary* (7), *Eldgrin* (6), *Parmesans* (6), *Carlsbad* (5), *Scotchmen* (5), *Bigley* (4), *Cayuga* (4), *Constableville* (4), *Herpers* (4), *Herkimer* (3), *Kolhouri* (3), *Averagetown* (2), *Bratpuhrian* (2), *Brud* (2), *Castorland* (2), *Glocks* (2), *Hagstrohms* (2), *Holdermann* (2), *Homesteaders* (2), *Iliumites* (2), *Iliumwise* (2), *Injuns* (2), *Kandahar* (2), *Kisco* (2), *Merganthaler* (2), *Middleville* (2), *Ormand* (2), *Pythias* (2), *Surrasi* (2), *Tamanrasset* (2), *Adirondacks*, *Afyon*, *Algonquins*, *Americanwise*, *Amerikka*, *Breslaw*, *Calhouns*, *Chippewa*, *Cocoanut*, *Cohoes*, *Collester*, *Elbesan*, *Evansville*, *Farafangana*, *Glock*, *Goyas*, *Grecos*, *Halporn*, *Hanford*, *Harrisburg*, *Hickock*, *Hjoring*, *Iliums*, *Iroqois*, *Kaifen*, *Karahisar*, *Kiukiang*, *Kransystav*, *Krashdrahr*, *Kyakhta*, *Lionhearted*, *Lubbocks*, *McCarren*, *Michelangelos*, *Milankowitz*, *Mirro*, *Moskowitz*, *Pelrine*, *Pepkowitz*, *Philippino*, *Pittsburghs*, *Poznitsky*, *Rembrandts*, *Remsen*, *Renoirs*, *Rosenau*, *Sheperd*, *Takarus*, *Ust*, *Utica*, *Vincis*, *Vonnegut*, *Watervliet*, *Wheaties*, *Zandt*

Exclamations, expletives, and ideophones: 43 types (82 tokens)

urdle (13), *hmmmm* (5), *vaaaaaaa-* (4), *-zuzip* (4), *dinga-* (3), *-grumph* (3), *kaplowie* (3), *vuuuuzzzzzip* (3), *bazz-* (2), *furrazz-* (2), *kablooooom* (2), *ooooooooooooh* (2), *pfft* (2), *ssssh* (2), *ug* (2), *-wap* (2), *znick* (2), *aaaaaaah*, *aaaaaah*, *-aaaaah*, *agh*, *ahhhhh*, *-assed*, *-azz-*, *azzzzzzzzzz*, *azzzzzzzzzzzzzz*, *baloooooooowie*, *clunkle*, *fromp*, *froomp*, *furr-*, *kablooom*, *kablooooom*, *kaboom*, *mmmmmm*, *mmmmmmm*, *naaaaaah*, *naah*, *ohhhh*, *ohhhhhhh*, *phwew*, *upps*, *vuuuuzzzzzzip*

Abbreviations: 13 types (17 tokens)

DSM (2), *GCT* (2), *VFW* (2), *Wks.* (2), *BDH*, *Chev.*, *Cond.*, *grad*, *KTM*, *NNS*, *pfc.*, *phys.*, *Silv.*

Appendix B: 253 new types (2,155 tokens) found in *The Sirens of Titan*

OED2 base forms, with inflected forms and derivatives: 178 types (312 tokens)

-synclastic (26), *batball* (20), *squadmates* (12), *infundibula* (10), *concessionaires* (9), *infundibulum* (9), *thrup* (9), *infundibulated* (8), *instructress* (6), *billfold* (4), *wirehouse* (4), *batballs* (3), *blatted* (3), *builded* (3), *emptily* (3), *outwardness* (3), *suh* (3), *wavelet* (3), *chimneylike* (2), *coathanger* (2), *cruelest* (2), *cyclopedia* (2), *fortunetelling* (2), *gropingly* (2), *heartbrokenly* (2), *hisself* (2), *-mâché* (2), *palsied* (2), *peyotl* (2), *shriveled* (2), *twilit* (2), *undershorts* (2), *adoringly*, *aspera*, *bandmaster*, *-barreled*, *bottommost*, *breechclout*, *brummagem*, *brutishness*, *buggywhips*, *butcherknife*, *buttercrunch*, *calluses*, *-cheekboned*, *chinked*, *clawlike*, *-colonelcy*, *colorlessness*, *cookware*, *couldst*, *deadness*, *decillion*, *dematerializations*, *diagramed*, *dictu*, *dismayingly*, *dottles*, *dreamingly*, *droughte*, *dumfounding*, *earsplittingly*, *effeminacy*, *eightball*, *endoskeletons*, *ensnarled*, *exhaler*, *exhaustingly*, *fils*, *fitly*, *flews*, *footlocker*, *foreordained*, *frowzy*, *fuddled*, *gawkers*, *gayly*, *grindingly*, *honeymooning*, *hornblende*, *horseblanket*, *inabilities*, *incognita*, *instructresses*, *intergalactic*, *jazzing*, *lewdnesses*, *licour*, *limpingly*, *minareted*, *mirabile*, *mnemonically*, *mucilaginous*, *musingly*, *newsboy*, *noncombat*, *noncommissioned*, *nonsupport*, *oleomargarine*, *overexposed*, *overshoes*, *pantherlike*, *paroled*, *patria*, *perced*, *père*, *pinfeather*, *pityingly*, *-pleaser*, *predaceous*, *prismatic*, *prismatically*, *pulsingly*, *pushbrooms*, *puttered*, *puttering*, *quackery*, *quintuplicate*, *rakehell*, *raspingly*, *rechristened*, *recontamination*, *refurnished*, *residua*, *roguishly*, *rooftree*, *sashweights*, *scalplock*, *schoolmarmishly*, *semiprecious*, *sexer*, *shanghaied*, *shipfitters*, *shorted*, *-shotted*, *shoures*, *sideless*, *simperingly*, *skylarking*, *somber*, *somethingness*, *sote*, *spitted*, *spookily*, *suffocatingly*, *supportable*, *swaggerstick*, *swich*, *swishingly*, *temporariness*, *thumbtacked*, *tinnily*, *toodle-*, *toodleoo*, *treed*, *trifler*, *-trillionth*, *truthing*, *turquoises*, *uncreative*, *unslung*, *unsoldierly*, *unstitched*, *veyne*, *viscid*, *whan*, *whang-*, *whanged*, *whangee*, *whanging*, *whisperingly*, *windowpanes*, *wirehouses*, *wispily*, *wolfed*, *woodsman*, *writhingly*, *yessir*

Proper nouns, with inflected forms and derivatives: 65 types (1,817 tokens)

Rumfoord (542), *Unk* (534), *Salo* (202), *Chrono* (170), *Kazak* (54), *Brackman* (51), *Redwine* (51), *Tralfamadore* (32), *MoonMist* (14), *Tralfamadorian* (14), *Wilburhampton* (14), *Moncrief* (13), *Schliemann* (10), *Fenstermaker* (9), *Malachis* (8), *Koradubian* (7), *Tralfamadorians* (7), *Betelgeuse* (6), *Gomburg* (6), *Raton* (6), *Mauser* (5), *Canby* (4), *ELCO* (4), *Magellanic* (3), *Burch* (2), *Deimos* (2), *Garu* (2), *-Magnon* (2), *Minot* (2), *Phobus* (2), *Roamin* (2), *Rosenau* (2), *Vonnegut* (2), *Wataru* (2), *Aprille*, *Brownsville*, *Cagliostro*, *Chaucerian*, *Chrissakes*, *Cohoes*, *Dupree*, *Enfields*, *Groton*, *Imbrium*, *Khufu*, *Lavina*, *M13*, *Maffia*, *Mausers*, *Maxfield*, *McSwann*, *Nattaweena*, *Necker*, *Pulsifer*, *Sacre*, *Sani-*, *-Scoutish, Seward*, *Slivovitz*, *Sonnyboy*, *Stael-*, *Stivers*, *Tartufe*, *Trowbridges*, *Twelvetrees*

Exclamations, expletives, and ideophones: 7 types (15 tokens)

chuffa (6), *awwwww* (3), *graw* (2), *ohhhhh*, *pft*, *-scraws*, *ummmmmmmmmmm*

Abbreviations: 3 types (11 tokens)

UWTB (9), *CRR*, *FYI*

Appendix C: 172 new types (560 tokens) found in *Mother Night*

OED2 base forms, with inflected forms and derivatives: 90 types (110 tokens)

hangwomen (4), *aggressions* (3), *briquet* (3), *briquets* (3), *countersign* (3), *alles* (2), *chessmen* (2), *footstool* (2), *ragbags* (2), *ringless* (2), *submicroscopic* (2), *über* (2), *unfrocked* (2), *willful* (2), *amateurishly*, *amateurishness*, *anxiousness*, *arthritically*, *-baiter*, *ballgames*, *baseboards*, *behinds*, *beseechingly*, *billfold*, *birdless*, *birdshot*, *bootless*, *boxholder*, *catalepsis*, *catfood*, *chromo*, *clangingly*, *clutchless*, *coalbin*, *cobwebby*, *crematoria*, *datelines*, *disturbers*, *dragger*, *dropsical*, *dropsically*, *editorially*, *emptily*, *fountainhead*, *Geist*, *glutted*, *grommeted*, *gropingly*, *grownups*, *homicidally*, *howdy*, *inchworm*, *introducer*, *jailbird*, *jailbirds*, *latently*, *lispingly*, *memorialized*, *mildewed*, *miscast*, *muffed*, *namecard*, *nationless*, *nontechnical*, *overdrew*, *paintbox*, *patrolman*, *pityingly*, *propagandizing*, *rechartered*, *rededicated*, *refurnish*, *scabrous*, *'scuse*, *snaggle-*, *spavined*, *stairheads*, *strengthless*, *submachine*, *tid-*, *tinkerings*, *trunkful*, *unapologetic*, *unbowdlerized*, *unprintable*, *unpublishable*, *unsmilingly*, *willfully*, *wincingly*, *zippered*

Proper nouns, with inflected forms and derivatives: 78 types (446 tokens)

Resi (95), *Wirtanen* (74), *Noth* (44), *Krapptauer* (28), *Bodovskov* (16), *Hoess* (15), *Potapov* (14), *Mengel* (13), *Arpad* (11), *Westlake* (10), *Dobrowitz* (7), *Kahm-* (7), *Szombathy* (7), *Ohrdruf* (6), *Sonderkommando* (6), *-pileser* (5), *Tiglath-* (5), *Klopfer* (4), *Andor* (3), *Buchanon* (3), *Hazor* (3), *Schildknecht* (3), *Tiergarten* (3), *-Tru* (3), *Arndt* (2), *-Bodovskovian* (2), *Scharff* (2), *Schutzstaffel* (2), *Schwefelbad* (2), *Viverine* (2), *Vonnegut* (2), *Alexanders*, *Andaman*, *Ansel*, *Arafura*, *Auschwitzer*, *Bartlesville*, *Bernardsville*, *Bethune*, *Borisoglebsk*, *Coggin*, *Cotuit*, *Cyklon-*, *Delano*, *Dornberger*, *Dulcinea*, *Edens*, *Ehrens*, *Eichmanns*, *Gretels*, *Hansels*, *Hederich*, *Hersfeld*, *Heydrich*, *Hinkleyville*, *Horthy*, *Hottentots*, *Iwo*, *Killinger*, *Lomar*, *Maxfield*, *Mephistopheles*, *Miklos*, *Noths*, *Panza*, *Paulist*, *Peekskill*, *Riesengebirge*, *Schreiberhaus*, *Semites*, *Tartakover*, *Toboso*, *Tuvia*, *Vith*, *Wandervögel*, *Werther*, *Yadin*, *Yigael*

Exclamations, expletives, and ideophones: 4 types (4 tokens)

kapow-, *-roooooom*, *-vaaaaaaa-*, *vroooom*

Appendix D: 174 new types (751 tokens) found in *Cat's Cradle*

OED2 base forms, with inflected forms and derivatives: 106 types (144 tokens)

muddily (7), *blowtorch* (5), *mosaicist* (4), *billfold* (3), *doghouse* (3), *doodley* (3), *hatbox* (3), *painty* (3), *pissants* (3), *yessir* (3), *zo* (3), *amiability* (2), *blamming* (2), *catsy* (2), *emptily* (2), *grownups* (2), *incurious* (2), *patria* (2), *sulfathiazole* (2), *-woogie* (2), *alit*, *altruist*, *anesthetized*, *apathies*, *awarenesses*, *balistrariae*, *bartizan*, *behinds*, *beseechingly*, *bounceless*, *butterball*, *canted*, *caricaturist*, *-caroling*, *chickenlike*, *chiefest*, *chinked*, *choirgirl*, *chunking*, *conscienceless*, *constrictor*, *counterclockwise*, *crenels*, *-dahs*, *deathtrap*, *demoniacal*, *dropcloth*, *dropcloths*, *dulcitude*, *exercisers*, *flaccidly*, *goosed*, *greeter*, *gropingly*, *hearselike*, *highballs*, *hoptoad*, *howitzers*, *imploringly*, *irrelevantly*, *ithyphallic*, *jeweled*, *jigaboo*, *leeringly*, *loincloths*, *longness*, *machicolations*, *marcelled*, *meltingly*, *menthe*, *mistakemaker*, *moveth*, *oubliation*, *paddlewheels*, *-peen*, *peevishness*, *penciled*, *penes*, *perplexedly*, *petrescence*, *phonies*, *photographable*, *puddly*, *purifiers*, *raincapes*, *rakehell*, *rockabye*, *scraggly*, *snaggle-*, *snaky*, *spiraea*, *surrealistic*, *tartrate*, *telegrapher*, *tepees*, *tholepin*, *touchhole*, *tumid*, *twangingly*, *unabridged*, *unbundling*, *unsnapped*, *waterless*, *whistlingly*, *wouldst*, *-wristed*

Proper nouns, with inflected forms and derivatives: 61 types (589 tokens)

Bokonon (135), *Hoenikker* (105), *Monzano* (45), *Bokononist* (35), *Aamons* (22), *Koenigswald* (21), *Pefko* (16), *Crosbys* (14), *Hoosier* (13), *Zinka* (13), *Bokononists* (12), *Hoosiers* (11), *Mintons* (11), *Bokononism* (10), *Lorenzan* (10), *Horlick* (8), *Krebbs* (8), *Fata* (7), *Hoenikkers* (7), *Morgana* (7), *Pabu* (6), *Schlichter* (6), *Borasisi* (5), *Horvath* (5), *Lorenzans* (5), *-bumwa* (4), *Fabri-* (4), *Marmon* (4), *Rumfoord* (3), *Avram* (2), *Caz-* (2), *Chetniks* (2), *Tasmanians* (2), *Yancey* (2), *Alamogordo*, *Beautyrest*, *Brobdingnagians*, *Buchenwald*, *Cornellians*, *Delano*, *-Delt*, *Duco*, *Gamaliel*, *Hernando*, *Kret-*, *-licken*, *Littauer*, *Lowie*, *Moakely*, *Mohandas*, *Myrna*, *Navajos*, *Nilsak*, *Pigalle*, *Rumfoords*, *Sangre*, *Stater*, *Trenary*, *Vonnegut*, *Whitcomb*, *-yeen*

Exclamations, expletives, and ideophones: 7 types (18 tokens)

fugging (10), *uck* (2), *-whoom* (2), *ech*, *mmmmmmm*, *-phweet*, *pootee-*

Appendix E: 204 new types (694 tokens) found in *God Bless You, Mr. Rosewater*

OED2 base forms, with inflected forms and derivatives: 125 types (147 tokens)

bullfrog (9), *descendents* (4), *broomhandle* (3), *bayoneted* (2), *blueballs* (2), *chalkstripe* (2), *goosefish* (2), *lusted* (2), *paroled* (2), *predestinarian* (2), *razorblades* (2), *sparrowfarts* (2), *abridgment*, *agronomists*, *appraisingly*, *areaway*, *ascendency*, *aslop*, *bandsaws*, *-barreled*, *benefactions*, *blatted*, *boodles*, *bookjackets*, *boozily*, *boulevardier*, *boxturtles*, *briber*, *calculatingly*, *carpetbagging*, *chancres*, *chokingly*, *combustibles*, *communistic*, *conquerers*, *creakingly*, *cutenesses*, *deathgrip*, *disserting*, *djin*, *doghouse*, *dolled-*, *doubtingly*, *emptily*, *-enterpriser*, *firebugs*, *flatcar*, *floorwax*, *frowzy*, *fruitcake*, *fumingly*, *gaffhooks*, *garterbelt*, *-gartered*, *gigglingly*, *gropingly*, *gulched*, *gutturally*, *hackingly*, *hankerings*, *harborfront*, *hippity*, *humorist*, *idlers*, *incurious*, *instructress*, *lapstreak*, *lifejacket*, *listlessnesses*, *maggotty*, *marvelingly*, *midsection*, *mildewed*, *miraculousness*, *mooningly*, *parabolas*, *passings*, *penciled*, *phonies*, *predatorily*, *razorblade*, *realizer*, *rechristened*, *redeyes*, *revoltingly*, *rhapsodized*, *scalplock*, *seeketh*, *showgirl*, *skylarking*, *slaveringly*, *slurpers*, *sniveling*, *soapdish*, *squeegeed*, *stickmen*, *submachinegun*, *sulfuric*, *sunburned*, *suppertime*, *-swatters*, *syphilitic*, *tailcoated*, *telepath*, *thumbtacked*, *tinhorn*, *townies*, *twittery*, *tyrannous*, *unappeased*, *uncowled*, *unmindful*, *unpatched*, *unplayful*, *unum*, *vilely*, *washday*, *wastebasket*, *whinnyingly*, *whiskbroom*, *whiskbrooms*, *wincingly*, *wingchair*, *yahooism*, *yoni*

Proper nouns, with inflected forms and derivatives: 69 types (529 tokens)

Rosewater (236), *Mushari* (61), *Buntline* (35), *Pisquontuit* (30), *Rosewaters* (22), *Pena* (9), *Robjent* (9), *Ewald* (8), *Kilgore* (8), *Glampers* (7), *Merrihue* (7), *Selena* (6), *Elsinore* (5), *Cotuit* (4), *Hoosier* (4), *Rumfoord* (4), *Warmergran* (4), *Arrid* (3), *Buntlines* (3), *Foxcroft* (3), *Monon* (3), *Sawmakers* (3), *Antietam* (2), *Cleota* (2), *Finnerty* (2), *Kublai* (2), *Pittsfield* (2), *Swarthmore* (2), *Tralfamadore* (2), *Wakeby* (2), *Absorbine*, *Actium*, *Ambrosians*, *Andaman*, *Argylls*, *Barca-*, *Bloomington*, *Borgia*, *Chaplinesque*, *Dillingen*, *DuVrais*, *Elihu*, *Flemming*, *Glamperses*, *Glinko-*, *Gompers*, *Gonsalves*, *Hausmännin*, *Lavoris*, *Letzinger*, *Lucretia*, *Marott*, *Octavian*, *Ramba*, *Rumfoords*, *Rumpf*, *Sawmaker*, *Scomber*, *Shaltoon*, *Skat*, *Sunoco*, *Thermopane*, *Thorsten*, *Topsiders*, *Vitsayana*, *Vonnegut*, *Wakebys*, *Walpurga*, *Zetterling*

Exclamations, expletives, and ideophones: 6 types (14 tokens)

weet (8), *-scraw* (2), *aaaaah*, *mf*, *ohhhhh*, *wupps*

Abbreviations: 4 types (4 tokens)

FH, *-K2CP-*, *-TDM-*, *-W3K3-*

Appendix F: 154 new types (421 tokens) found in *Slaughterhouse-Five*

OED2 base forms, with inflected forms and derivatives: 93 types (114 tokens)

flatcar (4), *honeymooning* (3), *shriveled* (3), *blowtorch* (2), *boomingly* (2), *bulletproof* (2), *costumed* (2), *dartboard* (2), *dogtag* (2), *doorchimes* (2), *dragger* (2), *-focals* (2), *lumbermill* (2), *optometer* (2), *-rouser* (2), *socked* (2), *sweatsocks* (2), *alackday*, *amoretti*, *anesthetic*, *bandsaw*, *bearskin*, *beetfield*, *blackbread*, *blowouts*, *blutwurst*, *boohooing*, *brainlessly*, *calmingly*, *chinning*, *clatteringly*, *copilot*, *covetously*, *dashingly-*, *dogtags*, *doornail*, *drainings*, *droolers*, *-dums*, *exploratorily*, *flappingly*, *fragmentarily*, *frowsy*, *frumpishly*, *groggily*, *gropingly*, *haloed*, *hooved*, *howitzers*, *humorist*, *installment*, *jazzing*, *lethargical*, *loveplay*, *milkshakes*, *moonlike*, *moonscape*, *negligibility*, *overstuffed*, *pinhead*, *pitiers*, *potching*, *quieted*, *ratproofed*, *razorblades*, *rocketry*, *rodomontades*, *roweled*, *screamingly*, *semierect*, *shellbursts*, *shimmeringly*, *sniffingly*, *sniveling*, *snootful*, *soaringly*, *sobbingly*, *stingingly*, *stoppered*, *straightaways*, *suffocatingly*, *telephoners*, *-terrestrials*, *thrillingly*, *turnoff*, *unprotesting*, *uproariously*, *whorehouse*, *willfully*, *windburned*, *windmilled*, *windowpane*, *winkings*

Proper nouns, with inflected forms and derivatives: 56 types (297 tokens)

Lazzaro (45), *Rumfoord* (41), *Tralfamadore* (37), *Rosewater* (31), *Kilgore* (24), *Tralfamadorians* (22), *Tralfamadorian* (14), *Wildhack* (9), *Merble* (6), *Sugarbush* (4), *Yonson* (4), *Eaker* (3) *Fèvre* (3), *Slovik* (3), *Daguerre* (2), *Dresdeners* (2), *Frauenkirche* (2), *Kreuzkirche* (2), *Lucretia* (2), *Ostrovsky* (2), *Saundby* (2), *Stamboul* (2), *Tastee-* (2), *Alplaus*, *Ausable*, *Barca-*, *Billys*, *Breslau*, *Buchenwald*, *Carlsbad*, *Chemnitz*, *Corwin*, *Croesus*, *Dakto*, *Endell*, *Feodor*, *Friederich*, *GF-*, *Glatz*, *Huie*, *Karamazov*, *Königstein*, *Menjou*, *Plauen*, *Polack*, *Resi*, *Rinehart*, *Roadmaster*, *Scheherezade*, *Susann*, *Thiriart*, *Virginian*, *Vonnegut*, *WACs*, *WAFS*, *Zo-*

Exclamations, expletives, and ideophones: 4 types (8 tokens)

-weet (4), *hmmmm* (2), *hmmmmmmmm*, *ohhhh*

Abbreviations: 1 type (2 tokens)

Febs (2)

Appendix G: 181 new types (724 tokens) found in *Breakfast of Champions*

OED2 base forms, with inflected forms and derivatives: 98 types (135 tokens)

doodley- (7), *erns* (5), *parolee* (3), *rearview* (3), *soigné* (3), *tinhorn* (3), *windshields* (3), *banquette* (2), *grownup* (2), *hippity-* (2), *jailbird* (2), *jeweler* (2), *jism* (2), *livingroom* (2), *locomotor* (2), *motherless* (2), *plasticized* (2), *sandpapered* (2), *scepter* (2), *socked* (2), *transcontinental* (2), *trustingly* (2), *unitized* (2), *washday* (2), *allons*, *anticonservation*, *bandsaw*, *birdshot*, *blooped*, *blooping*, *bridgework*, *carload*, *catfood*, *copilot*, *copperheads*, *cumberbund*, *defenselessness*, *déjà*, *dinette*, *dismayingly*, *dreamhouse*, *enameled*, *everyplace*, *exterminators*, *extrasensitive*, *fairsize*, *formlessly*, *golfclub*, *gynecologist*, *heartrending*, *horsing*, *inkless*, *itchingly*, *jazzing*, *junked*, *leeringly*, *legbone*, *lonesomeness*, *loonily*, *lovingness*, *nonwhite*, *outspread*, *pantlegs*, *paroled*, *-père*, *poreless*, *quintillion*, *radioing*, *rubberbands*, *saddlebags*, *sexlessness*, *shamblingly*, *shriveled*, *signaled*, *slurpy*, *smilers*, *snaggle-*, *soapflakes*, *squirmy*, *sub-*, *syphilitic*, *syphilitics*, *tinhorns*, *tinsnips*, *trifocals*, *twittery*, *undippable*, *unfeelingly*, *unlighted*, *unornamental*, *untroubling*, *unum*, *waggishly*, *whatchamacallit*, *wheedlingly*, *whimsey*, *whorehouse*, *wingspreads*, *wolfed*

Proper nouns, with inflected forms and derivatives: 77 types (562 tokens)

Kilgore (113), *Karabekian* (43), *Rosewater* (33), *Shepherdstown* (30), *Keedsler* (28), *Hoobler* (27), *Maritimo* (18), *Ukwende* (18), *Kazak* (17), *Rabo* (17), *Robo-* (16), *Breedlove* (14), *Barrytron* (13), *Pefko* (13), *Drāno* (10), *Cohoes* (8), *Galaxie* (8), *Kago* (8), *Durling* (6), *Hurty* (6), *Indaro* (5), *Newcomb* (5), *Alfy* (4), *Garr* (4), *Gooz* (4), *Heliogabalus* (4), *Khashdrahr* (4), *Libertyville* (4), *Skag* (4), *ACTH* (3), *Attucks* (3), *BLINC* (3), *Crispus* (3), *Delmore* (3), *Pontiacs* (3), *Zog* (3), *Athanasius* (2), *Bagnialto* (2), *Crestview* (2), *CRF* (2), *Garterbelt* (2), *Indaros* (2), *Keedslers* (2), *Mahesh* (2), *Monon* (2), *Perma-* (2), *Philboyd* (2), *Queensboro* (2), *Stockmayer* (2), *Studge* (2), *Sunoco* (2), *WMCY* (2), *Ashtabula*, *Bearse*, *Dostoevski*, *Hellertown*, *Hoosier*, *Lembrig*, *Lieber*, *Lingamon*, *MDCCCLXXXI*, *MDCCLXXVI*, *München*, *Ojumwa*, *Olympiade*, *Omologato*, *Rathskeller*, *Scribner*, *Steinway*, *Tolstoi*, *Toronado*, *Turismo*, *Vonnegut*, *Wolfschmidt*, *Zeltoldimar*, *Zeltoldimarian*, *Zeltoldimarians*

Exclamations, expletives, and ideophones: 3 types (3 tokens)

torrrrrrrrrrrrrrrrrrrrrrrrrrrrrnnnnnnnnnnnn, *wooooo-*, *wooooooo-*

Abbreviations: 3 type (24 tokens)

CH_2 (14), *CN* (6), *WO1* (4)

Appendix H: 124 new types (300 tokens) found in *Slapstick*

OED2 base forms, with inflected forms and derivatives: 73 types (99 tokens)

diningroom (11), *lunchpail* (7), *frontiersman* (4), *grownups* (3), *descendents* (2), *doorkeeper* (2), *livingroom* (2), *palsied* (2), *sappy* (2), *batburgers*, *armorplate*, *awarenesses*, *axeman*, *calmingly*, *candlewax*, *cavalryman*, *centenarian*, *costuming*, *-crankum*, *creativeness*, *crinkum-*, *cripplingly*, *deadness*, *dinnerbells*, *disorderliness*, *dogpaddle*, *emptily*, *enlaced*, *fillingstation*, *flatcars*, *forgettery*, *gabblings*, *glans*, *gobblings*, *howitzers*, *humorlessness*, *jittering*, *laundress*, *lovelessness*, *maintainers*, *naptime*, *nays*, *Neanderthaler*, *Neanderthaloid*, *Neanderthaloids*, *normals*, *notetakers*, *overexcited*, *pantomimed*, *parquetry*, *ragingly*, *railroading*, *rechecking*, *saddlebags*, *screamingly*, *shirtmakers*, *shriveled*, *skeedaddled*, *snaggle-*, *soigné*, *squiggly*, *steamshovel*, *subhuman*, *suppertime*, *suppressants*, *talcumed*, *teemingly*, *tiddledy*, *treetrunk*, *twitted*, *weathertight*, *-weeniest*, *wiz*

Proper nouns, with inflected forms and derivatives: 51 types (201 tokens)

Isadore (22), *Mushari* (18), *Cordiner* (15), *Peterswald* (14), *-benzo-* (13), *-Deportamil* (13), *Urbana* (12), *Elihu* (11), *Oveta* (10), *Maxinkuckee* (6), *Villavicencio* (5), *Galen* (4), *Goatsucker* (4), *Tourette* (4), *Vonnegut* (4), *Sooners* (3), *Carmalt* (2), *Dostoevski* (2), *Fauntleroy* (2), *McBundy* (2), *Muskellunge-* (2), *Stankowitz* (2), *Tish* (2), *Woodpile* (2), *Berylliums*, *Fëdor*, *Gamaliel*, *Gumps*, *Hawkeyes*, *Hieronymus*, *Hoosiers*, *Jayhawkers*, *Kleindienst*, *Langhorne*, *Mellons*, *Mikhailovich*, *-Nellyism*, *Norvell*, *Pachysandras*, *Panza*, *Piatigorsky*, *Ponts*, *Rappahannock*, *Razorclam-*, *Rockmell*, *Sulfurs*, *Tcherkassky*, *Theodorides*, *Thirteens*, *Vanderbilts*, *Yamashiro*

Appendix I: 263 new types (859 tokens) found in *Jailbird*

OED2 base forms, with inflected forms and derivatives: 150 types (200 tokens)

jailbird (6), *mixology* (6), *pousse-* (6), *prothonotary* (6), *catfood* (4), *costumed* (4), *subbasement* (4), *abrasives* (3), *bartending* (3), *billfold* (3), *birdshit* (3), *cannery* (3), *inkpad* (3), *cochairman* (2), *coconspirators* (2), *coeditor* (2), *flannelcakes* (2), *flirtatiousness* (2), *imbecilic* (2), *incurious* (2), *jailor* (2), *keened* (2), *abutting*, *altitudinous*, *antiwar*, *appall*, *apso*, *awarenesses*, *batterers*, *bested*, *blackmarket*, *blonds*, *boulevardier*, *braining*, *brainlessly*, *breakfronts*, *bridgework*, *broadsword*, *brontosaurs*, *bunchy*, *cabinetwork*, *cavalryman*, *chamberpot*, *chemistries*, *clothiers*, *clowningly*, *coastwise*, *coconspirator*, *cofounder*, *coinbox*, *contendere*, *coquettish*, *crashingly*, *crematoria*, *crooningly*, *damagingly*, *defenseless*, *descendent*, *descendents*, *dinnerbells*, *distributorship*, *disturbers*, *doornail*, *drawerful*, *dustmop*, *earloops*, *earsplittingly*, *emptily*, *enterprisers*, *eyelessly*, *fiddlecase*, *forgivingness*, *foundryman*, *girllike*, *gloatingly*, *gropingly*, *grownup*, *handlessly*, *harmlessness*, *-hawing*, *highboys*, *horsewhipped*, *hounder*, *humorlessness*, *inspirer*, *jailbirds*, *jumpseat*, *kneepants*, *lampblack*, *laude*, *lazaret*, *locksmithing*, *-lousing*, *lowboys*, *-martialed*, *midtown*, *milkshakes*, *misbuttoned*, *motherless*, *newsmagazines*, *nonaggression*, *nonstrikers*, *overacting*, *overbooked*, *ownerships*, *placelessness*, *pretenses*, *raggedier*, *Reichsmarschall*, *respectfulness*, *restolen*, *retie*, *retoucher*, *ridiculousness*, *saddlebags*, *satirizing*, *sawblade*, *scuttlings*, *semiliterate*, *sentimentalize*, *serviceably*, *shirtstuds*, *shoepolish*, *shoon*, *shriveled*, *sidelight*, *signaled*, *silverfoil*, *socialistically*, *squiring*, *stamper*, *standers-*, *subhuman*, *suitcoat*, *teasings*, *telegraphers*, *thumbtacked*, *titillatingly*, *truckload*, *trudgers*, *trustingness*, *tuxedos*, *twilit*, *undreamlike*, *unhospitable*, *unjudgmental*, *unmilitary*, *unreflective*, *vengefulness*, *whorehouse*

Proper nouns, with inflected forms and derivatives: 107 types (650 tokens)

RAMJAC (100), *Clewes* (96), *Vanzetti* (44), *Arapahoe* (35), *Starbuck* (33), *Greathouse* (27), *Arpad* (26), *Edel* (25), *Hapgood* (25), *Cuyahoga* (20), *Ubriaco* (18), *Delmar* (8), *Izumi* (8), *Stankiewicz* (8), *Finletter* (7), *Kramm* (7), *Madeiros* (6), *Redfield* (6), *Strelitz* (6), *Piaf* (5), *Ponzi* (5), *Transico* (5), *Brockton* (4), *Kilgore* (4), *Bowery* (3), *Claycomb* (3), *Dürer* (3), *Figler* (3), *Gibney* (3), *McCones* (3), *Morristown* (3), *Peale* (3), *Salsedo* (3), *Sanza* (3), *Stegemeier* (3), *Ukey* (3), *Wyatts* (3), *Bangwhistle* (2), *BIBEC* (2), *Caryl* (2), *Celestino* (2), *Chessman* (2), *Failey* (2), *Honeybunch* (2), *Königstrasse* (2), *Männleinlaufen* (2), *Marcaccio* (2), *Petoskey* (2), *Tillie* (2), *Vicunians* (2), *Accutron*, *Amatis*, *Anheuser-*, *Anschluss*, *Bartolomeo*, *Bashevis*, *Bormann*, *Broun*, *Bulova*, *Carlsbad*, *Charlottesville*, *Christmastide*, *Cotuit*, *Crawdaddy*, *Delano*, *Edels*, *Ehrlichman*, *Fafner*, *Farben*, *Farfans*, *Frankfort*, *Frauenkirche*, *Garfinckel*, *Gebel*, *Gorki*, *Haldeman*, *Hänfstaengl*, *Hapgoods*, *Hitz*, *Hoosier*, *Kairys*, *Kaiserburg*, *Kappas*, *Kincaid*, *Lambchop*, *-lator*, *Leora*, *Limburger*, *Liv*, *Milland*, *Millay*, *Morais*, *Ophelias*, *Padwee*, *Passos*, *Pilates*, *Putzi*, *Remagen*, *Rolland*, *Stradivari*, *Tolstoi*, *Ullmann*, *Vicunian*, *Vonnegut*, *WAC*, *Welk*, *Whitcomb*

Exclamations, expletives, and ideophones: 6 types (9 tokens)

rowrr (3), *vrooom-* (2), *ohhhhhhhhhhhh*, *-roooooooooooooooooooooooooooooooom*, *-schnap*, *schnip-*

Appendix J: 164 new types (491 tokens) found in *Deadeye Dick*

OED2 base forms, with inflected forms and derivatives: 88 types (123 tokens)

shitbox (12), *laughingstock* (8), *sauerbraten* (4), *buttonless* (2), *drippings* (2), *filberts* (2), *honeybunch* (2), *legful* (2), *patrolman* (2), *playlet* (2), *playwriting* (2), *pled* (2), *semisweet* (2), *shitboxes* (2), *snaggletoothed* (2), *ungifted* (2), *whorehouses* (2), *-barreled*, *baseboard*, *boatload*, *brainlessly*, *bughouse*, *cabinetwork*, *calibers*, *carport*, *catshit*, *checkered*, *copperheads*, *countinghouse*, *cutest*, *daymare*, *dermatoses*, *downfield*, *draftiness*, *dumbfoundingly*, *dyslectic*, *emptily*, *everyplace*, *faceprinted*, *fakery*, *fatted*, *foreclosures*, *graffito*, *groused*, *handsomest*, *harrowingly*, *heartwarming*, *heebie-*, *holies*, *incurious*, *insultingly*, *inviters*, *jailbird*, *-jeebies*, *leprous*, *lunk*, *methaqualones*, *milkshake*, *miseducations*, *monologist*, *morepork*, *nonrepresentational*, *profundo*, *promisingly*, *prybar*, *repaved*, *resourcefully*, *rockpile*, *-scarum*, *seafoams*, *showplace*, *signaled*, *-skulled*, *skylarking*, *stoics*, *threescore*, *thrillingly*, *unblanched*, *ungraceful*, *unmusical*, *unrationed*, *unvain*, *voodooist*, *wastebasket*, *waxer*, *whatchamacallits*, *whorehouse*

Proper nouns, with inflected forms and derivatives: 76 types (368 tokens)

Metzger (64), *Ketchum* (37), *Maritimo* (24), *Morissey* (24), *Hippolyte* (20), *Hildreth* (16), *Hoobler* (13), *Shepherdstown* (10), *Keedsler* (9), *Metzgers* (9), *Oloffson* (9), *Drāno* (7), *Duveneck* (7), *Gatch* (7), *Rettig* (7), *Barrytron* (5), *Minorite* (5), *Pefko* (5), *Brokenshire* (4), *Furstenberg* (4), *Liederkranz* (4), *Linzer* (4), *Biphetamine* (3), *Oberlin* (3), *Pennwalt* (3), *Wetzel* (3), *Anyface* (2), *Arimathea* (2), *Barcalounger* (2), *Barrys* (2), *Durstine* (2), *Escadrille* (2), *Kokomo* (2), *Schramms* (2), *Sheperdstown* (2), *Volendam* (2), *Woollcott* (2), *WOR* (2), *Arjumand*, *Banu*, *Blaupunkt*, *Bucyrus*, *Cedarville*, *Cervantes*, *Courtland*, *Darvon*, *Dubuque*, *Fairchilds*, *Finkelstein*, *Fluoristan*, *Furstenbergs*, *Garand*, *Gleem*, *Jahan*, *Karabekian*, *Kenosha*, *Ketchums*, *Krementz*, *Learjet*, *Lusitania*, *Maytag*, *Meekers*, *Mephistopheles*, *Millay*, *Monon*, *Ostermans*, *Penfield*, *Piatigorsky*, *Rabo*, *RAMJAC*, *Seitz*, *Shitface*, *Tormé*, *Virginny*, *Vonnegut*, *Wetzels*

Exclamations, expletives, and ideophones: 1 type (1 token)

clackety-

Appendix K: 174 new types (999 tokens) found in *Galápagos*

OED2 base forms, with inflected forms and derivatives: 97 types (111 tokens)

pigmentosa (4), *mignons* (3), *susurruses* (3), *boxy* (2), *defenseless* (2), *honeymooned* (2), *mudbank* (2), *oceangoing* (2), *palsied* (2), *provisioned* (2), *airmailed*, *aliases*, *allez*, *amassers*, *antisocially*, *begetting*, *bellyband*, *bellyful*, *benignant*, *bighearted*, *birdcage*, *-blahing*, *bowline*, *brainpower*, *breechclout*, *bulletlike*, *bulletproof*, *deepwater*, *edentate*, *evanesced*, *eviscerator*, *exhibitionistic*, *fireproof*, *fisherpeople*, *fisherperson*, *fleeced*, *Fräulein*, *frequenters*, *furloughed*, *giveth*, *glottises*, *hallucinators*, *imperialistic*, *incognita*, *inheritable*, *insensate*, *landlubberly*, *larcenous*, *longboat*, *longboats*, *lunks*, *monopolar*, *motorship*, *napalmed*, *nonentities*, *ordainest*, *outsurvive*, *overelaborate*, *pinhead*, *potching*, *precut*, *printshop*, *prognathous*, *promisingly*, *ravening*, *resourcefully*, *retrogression*, *rocketry*, *roofer*, *sanitaire*, *screamingly*, *shipload*, *snaky*, *snarlingly*, *splitter*, *squishy*, *starlit*, *sundowns*, *susurruing*, *teletyped*, *trancelike*, *trenchknife*, *trillionaire*, *undammed*, *undershorts*, *underspoken*, *unfavorable*, *unheeding*, *unpotable*, *unreachable*, *urgencies*, *uteri*, *utile*, *warplanes*, *windowshades*, *wingspreads*, *worshiped*

Proper nouns, with inflected forms and derivatives: 77 types (888 tokens)

Mandarax (128), *Guayaquil* (75), *Kanka-* (74), *Akiko* (62), *Selena* (61), *Hisako* (60), *Zenji* (57), *Hiroguchi* (52), *Kleist* (52), *-bonos* (25), *Gokubi* (24), *Flemming* (22), *Hiroguchis* (22), *Mateo* (17), *Baltra* (15), *Hernando* (9), *GEFFCo* (7), *Cohoes* (6), *Ecuadorians* (6), *Colombianos* (5), *Donoso* (5), *Omoo* (5), *Quezeda* (5), *Wojciehowitz* (5), *Boström* (4), *Agosto* (3), *Arvid* (3), *Calle* (3), *Diez* (3), *Ecuatoriana* (3), *Folklórico* (3), *Genovesa* (3), *Hjalmar* (3), *Kilgore* (3), *Tibbets* (3), *Ziggie* (3), *Claggett* (2), *Dirno* (2), *Eleonore* (2), *Kirghiz* (2), *Kleists* (2), *Learjet* (2), *Mérida* (2), *-Neuburg* (2), *Quechuan* (2), *Rábida* (2), *Rosenquist* (2), *Bertolt*, *Bierce*, *Bonesana*, *Carryl*, *Cristóbal*, *Eyquem*, *Fernandina*, *Gokubis*, *Gömbös*, *Greenleaf*, *Hammerstein*, *Hedy*, *Hillis*, *Hitz*, *Ignacio*, *Kentuckian*, *Kenzaburo*, *Kenzaburos*, *Khufu*, *Lobsterville*, *Lor*, *Miklós*, *Nanno*, *Shinola*, *Sinka*, *Teodoro*, *Tiputini*, *Treveranus*, *Vonnegut*, *Watervliet*

Appendix L: 220 new types (606 tokens) found in *Bluebeard*

OED2 base forms, with inflected forms and derivatives: 118 types (149 tokens)

babyshit (9), *cannery* (4), *whatchamacallit* (4), *brownstones* (3), *chromos* (3), *womblike* (3), *asps* (2), *baseboards* (2), *bushwa* (2), *cowhide* (2), *defenseless* (2), *dismayingly* (2), *flyspeck* (2), *fubar* (2), *gutshot* (2), *postcoital* (2), *wowed* (2), *abashing*, *abutted*, *amiability*, *antimilitaristic*, *bandsaw*, *begrudgingly*, *biggie*, *bottommost*, *checkered*, *cherchez*, *chokingly*, *-cochère*, *cocksmen*, *coffinlike*, *connoisseurship*, *cosmopolite*, *costumed*, *crueler*, *dandified*, *doorpath*, *dribblings*, *dropsies*, *dumbfoundingly*, *earthshaking*, *effeminacy*, *emptiest*, *encourgement*, *evidentally*, *flashbulb*, *fleabag*, *flyspecks*, *foreclosures*, *fuckups*, *goosey*, *groggily*, *handshakers*, *holies*, *houseflies*, *houseroom*, *humorlessness*, *installment*, *interfertile*, *laughingstock*, *leeringly*, *lowborn*, *lunk*, *moodiest*, *moosehead*, *mothproofed*, *mudpies*, *negotiability*, *neutralities*, *oilcloth-*, *orchestrators*, *orgastic*, *outranked*, *overstuffed*, *pacifistic*, *painty*, *paperhanger*, *paperhangers*, *pigheaded*, *pleasers*, *portraitist*, *prefectly*, *prenursing*, *protological*, *quintupled*, *raillery*, *reenter*, *reprimed*, *restretched*, *rinky-*, *sailfish*, *sawteeth*, *semiliterate*, *shoveling*, *showgirl*, *simperingly*, *skintight*, *skullcap*, *spearpoints*, *squalling*, *streetlamp*, *sub-*, *subraces*, *suckered*, *supranaturally*, *transcontinental*, *treed*, *trembly*, *twangingly*, *unembarrassed*, *unserious*, *unspiked*, *unswaddled*, *whisperingly*, *willfully*, *witchlike*, *witticism*, *yclept*

Proper nouns, with inflected forms and derivatives: 98 types (451 tokens)

Slazinger (76), *Circe* (52), *Rabo* (40), *Karabekian* (33), *Ignacio* (28), *Mamigonian* (22), *Beskudnikov* (19), *Finkelstein* (14), *Portomaggiore* (11), *Pomerantz* (6), *Busto* (5), *Barbira* (4), *GEFFCo* (4), *Innocenzo* (4), *Jolson* (4), *Karabekians* (4), *Normandie* (4), *Bauerbeck* (3), *Bengals* (3), *Isadore* (3), *Karpinski* (3), *Leidveld* (3), *Lucrezia* (3), *Mashtots* (3), *Mesrob* (3), *Salibaar* (3), *Santayana* (3), *Tarkington* (3), *Arapahoe* (2), *Arshile* (2), *Barrani* (2), *Battista* (2), *Dorene* (2), *Girolamo* (2), *Guston* (2), *Hiawatha* (2), *Hildreth* (2), *Karpinskis* (2), *Kevork* (2), *Medicis* (2), *Mencken* (2), *Mintouchian* (2), *Mohammedan* (2), *Riverhead* (2), *Savonarola* (2), *Vonnegut* (2), *Algren*, *Baziotes*, *Brens*, *Bridgehampton*, *Brobdingnagian*, *Claessen*, *Clyfford*, *Dadaists*, *Dagh*, *Dürer*, *-Foinet*, *Frenchified*, *Frisian*, *Ghiberti*, *Hamptonite*, *Henrik*, *Honeybunch*, *Horadam*, *Hovanessian*, *Hovanissian*, *Ibos*, *Kasabian*, *Klimt*, *Kouyoumdjian*, *Lascaux*, *-Luxes*, *Marktich*, *Marmon*, *Masaccio*, *Mohammedans*, *Mussini*, *Oporto*, *Polacks*, *Portomaggiori*, *Praecox*, *Quogue*, *Roadmaster*, *Sagaponack*, *Saroyan*, *SHAEF*, *Skidmore*, *Spahis*, *Stens*, *Tallin*, *Tanglewood*, *Terpsichore*, *Tolstoi*, *Trippingham*, *Uccello*, *Verman*, *Vitelli*, *Ziegfield*

Exclamations, expletives, and ideophones: 4 types (6 tokens)

clickety- (2), *zingo* (2), *-pank*, *ploop*

Appendix M: 203 new types (626 tokens) found in *Hocus Pocus*

OED2 base forms, with inflected forms and derivatives: 121 types (165 tokens)

garterbelt (12), *townies* (9), *footlocker* (4), *umiak* (4), *carillonneur* (3), *handgrenade* (3), *unicyclist* (3), *unteachable* (3), *whorehouse* (3), *antipersonnel* (2), *antiwar* (2), *bottommost* (2), *chattier* (2), *crematoria* (2), *cuestick* (2), *garterbelts* (2), *tenured* (2), *unteacher* (2), *absquatulated*, *backslid*, *ballpark*, *baseboards*, *cannoneer*, *carbonizer*, *chokingly*, *chowderhead*, *crabbing*, *crackup*, *creampuffs*, *daid*, *debater*, *-defenestrators*, *devastators*, *dolled*, *draftee*, *draftees*, *-drowners*, *firebox*, *firstie*, *fishline*, *footsoldier*, *footsoldiers*, *fragged*, *fragging*, *futurology*, *garrote*, *gloatingly*, *goshdarned*, *grownup*, *grownups*, *handgrenades*, *headliner*, *hippity-*, *humanized*, *imbecilic*, *incurious*, *ineducable*, *inheritable*, *intergalactic*, *jackbooted*, *jailbird*, *jailbirds*, *kilovolts*, *livable*, *lordy*, *mapmaking*, *multidimensioned*, *nonacademics*, *noncombatant*, *nonmusicians*, *nonparticipation*, *nonscientists*, *nonunion*, *orals*, *paraplegics*, *parings*, *paroling*, *-peen*, *pillowed*, *poorhouse*, *pricelessly*, *probables*, *profundo*, *-pullers*, *pyrotechnician*, *pyrotechnicians*, *reburied*, *reinstalled*, *remount*, *resegregation*, *roisterers*, *roundtrip*, *sculptress*, *sentimentalized*, *sidearms*, *sodded*, *sozzled*, *spookily*, *-starvers*, *statesmanlike*, *stinkbombs*, *stinky*, *stupidities*, *suppertime*, *townie*, *tracings*, *ultrarealistic*, *ultrasophisticated*, *uncatchable*, *underclasspersons*, *unfavorable*, *unicycling*, *unsalable*, *unsmilingly*, *untightened*, *unvictory*, *upperclasspersons*, *weenie*, *westernmost*, *-whippers*, *windowpane*

Proper nouns, with inflected forms and derivatives: 79 types (452 tokens)

Tarkington (136), *GRIOT™* (27), *Mohiga* (23), *Moellenkamp* (21), *Hartke* (19), *Slazinger* (17), *Dubuque* (16), *Barrytron* (13), *Bergeron* (12), *Meadowdale* (10), *Tralfamadore* (10), *VanArsdale* (7), *Moellenkamps* (7), *Pahlavi* (7), *Samoza* (7), *Tarkingtonians* (7), *Madelaine* (6), *Fenstermaker* (5), *Hiscock* (5), *Fedders* (4), *Isuzu* (4), *Shultzes* (4), *Topf* (4), *Akbahr* (3), *Howdy* (3), *Swarthmore* (3), *Tarkingtons* (3), *Ainus* (2), *Blankenship* (2), *Capades* (2), *Clewes* (2), *Donners* (2), *Farben* (2), *Freethinkers* (2), *Ironsides* (2), *Nemours* (2), *Robo-* (2), *Roys* (2), *Tegucigalpa* (2), *Tralfamadorians* (2), *Turismo* (2), *Vonnegut* (2), *Waxahachie* (2), *Wheelock* (2), *Anheuser-*, *Arapahos*, *Arsdale*, *Barnegat*, *Betelgeuse*, *Bluebellies*, *Bratpuhr*, *Bucknell*, *Carib*, *Carpathia*, *Clarabell*, *Deerfield*, *Dreiser*, *Hiscocks*, *Injuns*, *Iwo*, *Krugersdorp*, *Laramie*, *LeGrand*, *Marthinus*, *Montagues*, *Oberlin*, *Onondaga*, *Pahlavis*, *Paso*, *Peale*, *Polk*, *Reb*, *Saabs*, *Shiloh*, *Sinatras*, *Tarkingtonian*, *Thorazine*, *Waynes*, *Westmoreland*

Exclamations, expletives, and ideophones: 3 types (9 tokens)

bloomp (6), *blankety-* (2), *ooof*

Appendix N: 263 new types (579 tokens) found in *Timequake*

OED2 base forms, with inflected forms and derivatives: 141 types (267 tokens)

timequake (70), *grownups* (8), *bashers* (6), *grownup* (5), *lidless* (5), *swoopers* (5), *artsy-* (4), *déjà* (4), *-fartsy* (4), *whoozit* (4), *clumpity* (3), *equivalency* (3), *erns* (3), *firepersons* (3), *scrooched* (3), *timequakes* (3), *ballpark* (2), *bankable* (2), *fatling* (2), *leadeth* (2), *lotsa* (2), *monopolar* (2), *sappy* (2), *airlanes*, *anesthetists*, *anointest*, *antihero*, *appoggiatura*, *armorplate*, *auricles*, *bakemaster*, *beanbag*, *beddy-*, *birdshit*, *blowtorches*, *blueballs*, *bughouse*, *cathouse*, *certifiable*, *classifiable*, *confessedly*, *consistencies*, *cookouts*, *-crankum*, *crinkum-*, *deader*, *decorticated*, *dogtags*, *dolled*, *doodley*, *doodoo*, *doornail*, *driverless*, *dumdums*, *escalier*, *excerpted*, *exculpatory*, *faut*, *faw*, *feds*, *-fleuve*, *functionless*, *groggily*, *haymow*, *heartwarming*, *hiccuping*, *hooty-*, *horsecrap*, *huffmobile*, *hulking*, *humorist*, *installment*, *-jailbird*, *jeeringly*, *junkyard*, *learnable*, *literates*, *loanshark*, *lordy*, *maketh*, *meathooks*, *microtome*, *midtown*, *mopery*, *motivationally*, *mousetrapped*, *multibillionaires*, *neglectful*, *-nellyism*, *numero*, *onrushing*, *overplanted*, *overshoes*, *overstuffed*, *paraphraser*, *penitentiaries*, *pictureness*, *pocketwatches*, *policepersons*, *preparest*, *providentially*, *rakehell*, *reexamines*, *rehashing*, *restoreth*, *retyping*, *saloonkeepers*, *scapegoating*, *schooler*, *scrooch*, *semiautomatic*, *shingled*, *-sixed*, *skeptic*, *smokable*, *somethings*, *sopper-*, *squiggly*, *stairstep*, *stinky*, *stupidities*, *submachine*, *sunburned*, *syph*, *testees*, *theatricals*, *unbolted*, *uncivil*, *unemployability*, *unfavorably*, *unmalleable*, *unsweet*, *villainously*, *vivant*, *-weensy*, *whistlestop*, *whizbang*, *whorehouses*, *windowpanes*, *wowed*, *yahooistic*

Proper nouns, with inflected forms and derivatives: 114 types (299 tokens)

Kilgore (67), *Vonnegut* (22), *Lieber* (14), *Sunoco* (10), *Booboolings* (8), *Booboo* (7), *Shortridge* (6), *Dalhousies* (5), *Fleon* (5), *Styron* (5), *Wynkoop* (5), *Banalulu* (4), *Barus* (4), *Hickenlooper* (4), *Ibo* (4), *Ibos* (4), *Swarthmore* (4), *Biafran* (3), *Boobooling* (3), *Dalhousie* (3), *Hoosier* (3), *Raye* (3), *Chinua* (2), *Dortmunder-* (2), *Enola* (2), *Hotchner* (2), *Junius* (2), *Kerfuit* (2), *Kosinski* (2), *Krementz* (2), *Littauer* (2), *-Masoch* (2), *Mbuti* (2), *Mencken* (2), *Whitcomb* (2), *Yarmolinsky* (2), *Achebe*, *Algren*, *Augie*, *Barkenhicker*, *Bysshe*, *Cayuga*, *Cheever*, *Cleopatras*, *Cohoes*, *Communistic*, *Delicto*, *Dictu*, *Disneyesque*, *Donoso*, *Dripper*, *Failey*, *Freethinker*, *Freethinkers*, *Gothics*, *Guaranty*, *Hickenbar*, *Hitz*, *Honeybunch*, *Hurty*, *Jeffersonian*, *Juans*, *Judeo-*, *Kappas*, *Klinkowitz*, *Kokomo*, *Krassner*, *Langmuir*, *Lascaux*, *Lockenlooper*, *Lockenbarker*, *Lockenbar*, *Loopenhick*, *Loopenlock*, *Loree*, *Ludd*, *MacDowell*, *Markson*, *Maxincuckee*, *Mayas*, *McKim*, *Mensas*, *Mihalich*, *Mirabile*, *Noam*, *Nolte*, *Offit*, *Passos*, *Piaf*, *Pieratt*, *Pinsky*, *Rackstraw*, *Rauch*, *Riah*, *Sacher-*, *Saroyan*, *Schutzstaffel*, *Seren*, *Sixpacks*, *Squibb*, *Stagg*, *Steinmetz*, *Stromboli*, *Themak*, *Tralfamadore*, *Urbana*, *Victrolas*, *Watusis*, *Weide*, *Wiesel*, *Willa*, *Xanthippe*, *Yarmolinskys*, *Zanesville*

Exclamations, expletives, and ideophones: 6 types (9 tokens)

deedly (3), *bloomp* (2), *-assed*, *bloompity*, *schnip-*, *-schnop*

Abbreviations: 2 types (4 tokens)

MTYOAP (2), *PFC* (2)

RESÜMEE

TARTU ÜLIKOOL
INGLISE FILOLOOGIA OSAKOND

Edmund Alexander Dalton

AntWordProfiler analysis of the novels of Kurt Vonnegut, Jr., as set in opposition to the GSL, the AWL, and the BNC/COCA word-family lists
Kurt Vonneguti romaanide analüüs arvutiprogrammiga AntWordProfiler neid GSLi, AWLi ja BNC/COCA sõnapereloenditele vastandades

Magistritöö
2014
Lehekülgede arv: 87

Annotatsioon:
Käesolev magistritöö kujutab endast Kurt Vonneguti idiolekti kvantitatiivset uurimust, mis rajaneb olemasolevatel korpuslingvistilistel meetoditel. Eesmärgiks on esile tuua järgmine faktide kolmikjaotus: esiteks, sõnavara maht tuhandetes sõnaperedes, millede piisav oskus võimaldab lugejal mõista vähemalt 98% sõnedest romaanide ingliskeelsetes originaalides; teiseks, ootusvastaste suhteliste püsivuste või kasvude arv esinemissagedusega määratud ja astmeliselt järjestatud tuhandesõnapereliste loendite liikmete arvus kõnealuste romaanide puhul; kolmandaks, kontroll- ja katserühma vahelistest erinevustest tulenev vea ülemmäär, mis on kindlaks tehtav kirjakeelesõnaraamatu „Oxford English Dictionary“ II trükiga.
Ülesehituselt on töö kaheosaline, hõlmates kokku nelja peatükki: sissejuhatus ja põhjalik ülevaade kõrvutatavatest tekstikogudest, mis moodustavad teoreetilise osa, ning vastavate tekstikogude üksikasjalik eritelu ja lõppkokkuvõte, mis moodustavad empiirilise osa. Teine peatükk jaguneb omakorda kaheks paragrahviks, selleks et käsitada vastandlike rühmadena nii sõnapereloendeid kui ka romaanide tekste, kolmas peatükk seevastu neljateistkümneks paragrahviks, selleks et keskenduda teostele kordamööda nende ilmumisaastate järjestuses, juhindudes sealjuures autori 45aastast karjääri vastaval alal kajastavast bibliograafiast.
Valdav enamik analüüsi metoodikast on omistatav Paul Nationile, nagu näiteks seisukoht, et kui lugeja sõnavara hulka kuuluvad vähemalt 98% ühe konkreetse kirjutatu sõnedest, on põhjust eeldada, et kirjutatu sisu on tervikuna lugejale piisavalt arusaadav. Analüüs seisneb romaanidest koosneva katserühma võrdluses tekstikogude General Service List, Academic Word List ja nii British National Corpus kui ka Corpus of Contemporary American English alusel loodud sõnapereloenditest koosneva kontrollrühmaga ning väljundi tabeldustes.
Teooriaosa raames on võetud kokku nii sõnapereloendite ja iseseisva tekstikogu puudused kui ka vaatlustulemused rakendatavast metoodikast Euroopa Nõukogu keeleoskustasemete süsteemi ja Vonneguti enda hinnete kontekstis. Empiirilised andmed aga käsitlevad sõnede, sõnatüüpide ja -perede arvu ja protsenti juba fikseeritud sagedusastmetel ning täpsustavad suhteliste muutuste põhisuundi, samal ajal kui vea ülemmäär selgub lisades toodud sõnede arvust. Statistikast nähtub muuhulgas, et teoste keerulisus on tihti alla 8 000 sõnapere.

Märksõnad:
Ameerika, ilukirjandus, korpuslingvistika, romaanid, sõnavara

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina Edmund Alexander Dalton
(autori nimi)
(isikukood: 38307120333)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
„AntWordProfiler analysis of the novels of Kurt Vonnegut, Jr., as set in opposition to the GSL, the AWL, and the BNC/COCA word-family lists“
,
(lõputöö pealkiri)

mille juhendaja on dotsent Reet Sool,
(juhendaja nimi)

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus 1. mail 2014. a *(kuupäev)*

Edmund Alexander Dalton
(allkiri)