PHONOLOGICAL CONSTRAINTS AND FREE VARIATION IN COMPOUNDING: A Corpus Study of English and Estonian Noun Compounds

by

MARY SEPP

A dissertation submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Doctor of Philosophy,
The City University of New York

2006

UMI Number: 3213181

Copyright 2006 by
Sepp, Mary

All rights reserved.

UMI Microform 3213181
Copyright 2006 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

© 2006
Mary Sepp
All Rights Reserved

This manuscript has been read and accepted for the
Graduate Faculty in Linguistics in satisfaction of the
dissertation requirements for the degree of Doctor of Philosophy

_____	_____
Date	Dr. Martin Chodorow, Chair of Examining Committee
_____	_____
Date	Dr. Gita Martohardjono, Executive Officer

Dr. Martin Chodorow

Dr. Dianne Bradley

Dr. Robert Vago

_____

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

PHONOLOGICAL CONSTRAINTS AND FREE VARIATION IN COMPOUNDING: A Corpus Study of English and Estonian Noun Compounds

by

Mary Sepp

Advisor: Dr. Martin Chodorow

This research was designed to examine the patterns of variation in the phonological and/or orthographic form of Estonian and English noun compounds. Estonian noun compounds generally occur in one of two forms: $N_{1(nominative)}$+ N_2 , as in kool + meister ("schoolmaster"), or $N_{1(genitive)}$ + N_2, as in kooli + õpetaja ("schoolteacher"). Some Estonian compounds vary freely in form – e.g., veebsepp/veebisepp ("webmaster"). English noun compounds exhibit orthographic variation, as they may be written in three ways: closed ("bookstore"), hyphenated ("dot-com"), or open ("space station"). Many English compounds also vary freely – e.g., cellphone/cell-phone/cell phone. The principal goal of this study was to use statistical data derived from corpora to determine which variables best account for the choice of variant compound forms.

The 1,094 Estonian compounds used in this research came from a one million word corpus of Estonian literary and news texts. Data on variation of form were obtained from Google searches of the World Wide Web. Results showed a strong preference for

genitive forms, and it was posited that this preference is due to general principles of ease of pronunciation and ease of perception.

Phonology is also a factor in the distribution of English compounds. A number of phonological variables were examined in the current study: number of syllables, presence of compound stress, vowel sequences across internal lexical boundaries, and double consonants across internal lexical boundaries. Frequency data for these variables were extracted from a fourteen million word English corpus. Results of multiple regression analyses showed that the number of syllables in the compound is a stronger predictor of orthographic form than the other phonological features that were tested. Phonology was not assumed to be the only influence, however; lexical features were also examined. Results indicated a substantial contribution of the second constituent in predicting whether the compound would be open or closed, and a lesser, though important, contribution of the first constituent. A regression analysis combining phonological and lexical variables accounted for about 68% of the variance in the orthography of 707 high frequency English noun compounds.

Acknowledgements

In the course of my research for this dissertation, I have been influenced by many individuals, but there are certain among them to whom I owe a debt of gratitude. The first of these is my advisor, Martin Chodorow. His expertise in corpus linguistics, statistical methods, and programming helped guide me through the process and, to a great extent, made the successful completion of this project possible. Moreover, his patience and generosity have been unparalleled in my academic experience.

I would also like to acknowledge and thank my committee members, Dianne Bradley and Robert Vago, for their patience and invaluable feedback. In addition, I am grateful to Robert Vago for giving me the opportunity to teach linguistics at Queens College, which allowed me to deepen my knowledge and develop ideas relevant to my research.

During my stay in Estonia in the summer of 2003, I received valuable input from Heiki-Jaan Kaalep, Kadri Muischnek, and Kati Lepajõe of the University of Tartu, as well as from my friends and relatives in Tallinn and Saaremaa. In addition, I am very grateful to Heiki-Jaan Kaalep and Kadri Muischnek for creating the tools to facilitate a corpus study of Estonian.

I extend my thanks also to Juta Kurman for helping me learn the Estonian language and to my father, Elmar Sepp, for patiently sharing his native intuitions.

I am grateful to my fellow students Hope Cotton and Ylana Beller for their help with the stress judgments for the English data. I thank my classmate Eiji Nishimoto for so thoughtfully sharing some of his research materials. Thanks also to Hope for her

feedback, and to my brother EJ for his proofreading help, on this work and previous work. Finally, I would like to express my gratitude to my friends and my entire family, especially my mother and father, for their encouragement and prayers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter One

INTRODUCTION: Background to the study

Morphology is concerned with word formation. Although some types of word formation, such as blends ("blog" - weblog, "mochaccino" - mocha cappuccino) and acronyms ("WMD" - weapon of mass destruction, "TESOL" - teaching of English to speakers of other languages), have become popular in recent years, derivation and compounding are still the most productive by far. Of course, the productivity of word formation processes varies from language to language (Bauer 2001: 117-119), so that compounding, for example, is highly productive in some languages and not so common in others. The present study examines the form of compounds in two languages where that process is pervasive and dynamic. A corpus-based, statistical approach is taken to investigate the relationship between variant forms of noun compounds and certain phonological and other variables.

1.1 Why compounds?

Studies have shown that, as a process, compounding is quite simple. For example, an experiment reported by Clark (1993: 143-150) showed that English-speaking children under the age of four relied heavily on compounding for the creation of novel

words – 80% of the words created were compounds, while only 20% were derivations. Though compounds are apparently easy to construct, a greater difficulty may lie in the comprehension, perception, or interpretation of compounds.

Compounds have been and remain a problem for computational systems. Challenges to the efficient machine recognition of compounds have highlighted some of the complexities that stem from the process. For example, text-to-speech systems need to know whether a sequence of nouns is a compound or a syntactic phrase in order to determine its correct prosodic structure. Machine translation needs to make this distinction for correct interpretation of the string. However, the greatest challenge is that of novel words. Since compounding is so productive, there are constantly new compounds being created. Some of these are transparent, but many are not. For instance, on a website called "The Word Spy" (www.wordspy.com), at least one new word or word combination is posted each day from Monday through Friday. During the weeks of November 14th and November 21st, 2005, the following twelve novel "words" were posted:

1. Cyber Monday
2. Black Friday
3. Googlejuice
4. acoustic ecology
5. rescue call
6. Swiss army phone
7. social bookmarking
8. playlistism
9. playlist anxiety
10. targeted Trojan horse
11. folksonomy
12. placeshift

Without even pondering the meanings of these intriguing coinages, we may surmise that ten of the twelve are compounds. The other two are derivations: "folksonomy" "playlistism". The latter even has a compound stem – i.e., "playlistism" → "play+list" + ism. Moreover, one might guess that most of these compounds are nouns. In fact, only "placeshift" is not. As for what they mean (see footnote)[1], a little creativity may be required. Most of them are "opaque", which means that the sense of the compound is not straightforwardly equal to the sum of its parts (Bauer 1983: 19-20). Humans could probably obtain a close approximation of the meaning but would require context. Machines, on the other hand, would not be able to reason out the meaning of a new word in this way.

Another problem for natural language processing, and for machine translation in particular, is the fact that compounds pattern very differently cross-linguistically. This is true both in terms of how they are formed and how productive they are. For example, some languages have right-headed compounds, while others have left-headed compounds. Some languages (e.g., German) have special linking morphemes which

[1] The definitions for items 1-12 as stated on www.wordspy.com are the following (8 and 11 are parenthesized to distinguish them from the compounds): 1. The Monday after the U.S. Thanksgiving holiday, when online retailers reportedly experience a surge in purchases; 2. The Friday after the U.S. Thanksgiving holiday, considered to be the busiest retail shopping day of the year; 3. The presumed quality inherent in a Web site that enables it to appear at or near the top of search engine results, particularly those of the Google search engine; 4. The study of the relationship between living organisms and their sonic environment; 5. A call to a cell phone placed at a prearranged time to give the person being called an excuse to end a date or other social engagement; 6. A cell phone that includes multiple non-voice features such as a digital camera, digital audio player, and electronic organizer; 7. Saving and applying keywords to one's personal collection of web site bookmarks on a site that enables other people to share those bookmarks; (8. Judging a person based on what songs are on the playlist of his or her digital music player); 9. Anxiety felt by a person who fears what other people might think of the music on his or her digital music player; 10. A Trojan horse program sent as an attachment in an e-mail message that has a subject line, body text, and return address that have been crafted to fool the recipient into opening the attachment; (11. An ad hoc classification scheme in which Web users apply their own keywords to site content as a way of categorizing the data they find online); 12. To redirect a TV signal so the viewer can watch a show on a device other than his or her television.

connect compound components, while others have no linking elements. Many languages spell compounds as a single orthographic word, while English, for example, often leaves spaces or puts hyphens between constituents. Languages may also differ in which concepts are expressed with compounds vs. some other type of word. And finally, some languages simply do not use compounding as much as others. All of these factors create challenges for humans and still more for machines.

1.2 Why a corpus study?

Every method of linguistic investigation has its merits, and old methods are not necessarily made obsolete by new ones. The empiricists may have been overtaken by the rationalists during the Chomskyan "revolution", but they did not perish. With advances in language technology, a trend towards more empirical approaches in the study of language has re-emerged. In the technological domain, where statistical natural language processing (NLP) has become the modus operandi, the focus has shifted from the binary grammaticality judgments of generative linguistics to calculations of frequency and probability. In the latter case, the choice is no longer whether a sentence is "well-formed" or not. Rather, the question is whether it is "usual" or "unusual" (Manning and Schütze, 2000: 7). Depending on the goals of a particular study, a generative approach might serve one's purposes better. But certainly determining what is "usual" seems to be more straightforward than determining what is "grammatical". Judging grammaticality requires intuition, and not everyone's intuitions are the same. On the other hand, judging

the "usual" vs. "unusual" would seem to be a mere matter of counting. If a word or a phrase or a sentence occurs frequently, we can say that it is usual; by the same token, if it occurs infrequently, then we might say that it is unusual. One might take this a step further and say that if a word or a phrase or a sentence occurs frequently it must be grammatical. The form that occurs infrequently might then logically be considered a candidate for ungrammaticality.

There are several problems attached to such a notion. First, the low frequency of a linguistic unit in no sense entails that it is ungrammatical (Lapata and Lascarides, 2003). "Hapax legomena", or words which only occur once in a corpus (Pinker, 2000: 172), may be neologisms (e.g., "overworking class")[2], outdated words or entities (e.g., "soda fountain"), technical words (e.g., "nucleotide phosphoanhydride"), words occurring with an alternate spelling (e.g., "aluminium can")[3], colloquialisms (e.g., "motormouth"), or words that may be familiar but simply do not come up in many of the contexts sampled (e.g., "carousel horse"). In some cases, seemingly usual lexical items may have a low frequency or perhaps not even occur in the particular corpus one is using. For example, "windbreaker", "sheet metal", and "paint roller" were among the hapaxes identified in the English corpus used in the present study. There is nothing ill-formed about any of these examples, so it would seem that low frequency does not preclude grammaticality.

---

[2] "Overworking class" is defined by Paul McFedries, author of "The Word Spy" (www.wordspy.com/diversions/neologisms.asp) as "A segment of society in which the chief characteristic is the desire or need to work long hours."

[3] "Aluminium" is the conventional British spelling; "aluminum" is the standard spelling in American English, so it is not surprising that it would have a low frequency in an American English corpus.

Secondly, while the high frequency of a linguistic unit may actually suggest grammaticality, it does not necessarily mean that the expression is usual. This is because its "usualness" greatly depends on how balanced the corpus is. A balanced corpus is one consisting of texts from a variety of genres and on a variety of topics. For example, if the corpus were composed only of texts from clothing catalogues, then "windbreaker" would probably not be a hapax legomenon. On the other hand, if "nucleotide phosphoanhydride" occurred 200 times in a corpus, one could not say that this expression is usual, except perhaps in a particular domain. So it could be that "nucleotide phosphoanhydride" occurs very frequently in one particular context, but not at all in any other context. The problem of misleading frequencies is discussed briefly in Kaalep and Muischnek (2002) and addressed more directly in Nishimoto (2004). In his work on measures of morphological productivity, Nishimoto proposes that the problem of distinguishing "high frequency" words from "wide usage" words can be resolved by a cross-comparison of corpus segments.

In general, the frequency statistics obtained from large corpora provide reliable data for linguistic investigation. These data can also be used to train computer systems, so they may predict linguistic behavior. High frequency words do usually suggest grammaticality (i.e., well-formedness) and low frequency may often coincide with ungrammaticality or "ill-formedness". Nevertheless, the usefulness of such statistics depends greatly on the nature of the corpus. A corpus must be balanced in terms of content and style.

There are several large corpora available for English that cover a wide variety of texts. The most famous corpus for written American English is probably the Brown

Corpus (Kučera and Francis, 1967). This is a corpus of approximately one million words compiled during the 1960s and 1970s by H. Kučera and W. Francis of Brown University (hence the name). It had been widely used in the decades following its creation since it was the only English corpus of its size available. However, studies on current linguistic phenomena may no longer find this corpus useful. Furthermore, the size of so-called "large" corpora has increased substantially. The British National Corpus (BNC), released worldwide in 2001 by Oxford University Press, consists of more than 100 million words. A comparable corpus is being compiled for American English. Upon its completion, the American National Corpus (ANC)[4] will also contain 100 million words. A segment of this corpus (11 million words) has been released in the interim.

There are a number of considerations in choosing a suitable corpus. One, of course, is its size. It is difficult to know before a study begins exactly how large a corpus may be needed. Generally, one million words would be considered too small. However, if the study is on determiners, a million words may be sufficient since determiners tend to occur very frequently in most texts; for example, in the Brown Corpus, the determiner "the" by itself constitutes nearly 7% of the word tokens sampled (http://en.wikipedia.org/wiki/Brown_Corpus). In a 514,000 word segment[5] of the ANC, the same proportion was calculated.

Another factor is the content. If one is studying the morphology of medical terminology, the corpus should be domain-specific. As a result, it need not necessarily be

[4] The "ANC First Release" (Ide, Reppen, and Suderman, 2003) is distributed by the Linguistic Data Consortium of the University of Pennsylvania. Information on this and various other corpora is available through the LDC website: www.ldc.upenn.edu .

[5] This segment consisted of Berlitz travel texts.

so large. Where more general domains are required, the corpus should be chosen accordingly.

Yet another issue in choosing a corpus is its age. Since corpora, and especially large corpora, take years to compile, we may find they are not as up-to-date as we would like. One solution to this problem suggested by Volk (2002) was to use the web as a corpus. This sounds like a logical and timely solution. However, as Volk noted, this scheme is not without its problems. First of all, web content changes every nanosecond. Analyzing data that is in constant flux may present challenges. Even if we assume this issue could be resolved, web content may not be balanced, despite its breadth. There are many errors in the content and there are also certain topics that are very popular online but may not be of much interest linguistically – e.g., pornography sites. Volk's proposal finally was to create a search engine that is designed specifically for linguistic inquiry. He suggests that in addition to the Boolean operators (e.g., "and", "or", "not") available for ordinary search engines, a linguistic search engine should also include tools such as *syntactic class operators* to restrict searches to a specific part of speech, *subject domain operators* to restrict the search to certain domains such as computer science or linguistics, and *environment operators* to restrict searches to a particular section of a document, e.g., headers or lists.

Notwithstanding the sometimes arduous task of finding an appropriate corpus, corpora provide a tremendous resource for linguistic research. Using small sample sets to make linguistic generalizations seems naïve and unnecessary when there are large corpora available to provide more substantial evidence. There are certain phenomena whose analyses would be difficult without the benefit of a large amount of data. As

Anttila (1997) noted, corpus data are critical for the analysis of phenomena such as variation and speaker preferences. Since the present study addresses the issue of variation in compounding, a corpus analysis seemed the most logical approach.

1.3 Why Estonian and English?

Estonian is among the very few European languages which are not Indo-European. Specifically, it is a Finno-Ugric language, related to Finnish and Hungarian. English, by contrast, is Germanic and thus Indo-European. The genetic distance between Estonian and English suggests considerable differences in the grammars of these two languages. Nevertheless, they share some interesting characteristics. First, compounding is very productive in both languages. Secondly, both of these languages exhibit variation in the form of compounds. In some cases, this variation is tightly constrained by the grammar, and in other cases it is free. For Estonian, the first component of compounds varies generally between the genitive and nominative singular forms, e.g., *kool+meister* ("schoolmaster") and *kooli+õpetaja* ("schoolteacher"). This difference often entails an additional syllable in the genitive, as in [kɔɔ:l][6] vs. [kɔɔ.li]. English compounds, on the other hand, vary in the orthography – i.e., they are written closed (as one orthographic word), open, or hyphenated: *bookstore, self-consciousness, space station*. Some English compounds also may vary freely – e.g., *cellphone/cell-phone/cell phone*. For Estonian

[6] The transcription of the vowel here, [ɔɔ:], represents the extra long vowel; minus the colon it is simply long. This notation is a slight variation on that used in the literature on Estonian quantity. E.g., Hint (1998) uses an accent mark instead of the colon to indicate overlength: [ˋɔɔ].

and for English, where there is free variation, there are clear preferences. Determining what motivates these preferences is the primary goal of this dissertation.

1.4 Purpose and organization of the work

This research was designed to examine the patterns of variation in the phonological and/or orthographic form of English and Estonian noun compounds. The principal goal, as stated above, is to determine the motivations for the choice of variants. In addition, the following goals were pursued:

a. To show the extent to which free variation occurs in both English and Estonian noun compounds and to discover the conditions under which a compound's form may vary freely.
b. To clarify the role of phonology in compounding.
c. To account for the distribution of open/hyphenated/closed orthographic forms in English noun compounds and the distribution of genitive/nominative forms in Estonian noun compounds.
d. To work towards a greater understanding of what defines a compound.

This work is essentially set up as two corpus studies with a common goal set. It is organized into six chapters. The next chapter explains key terms used in this work, such as "word" and "compound". Chapter 3 discusses some of the literature relevant to the issues under consideration in this study.

The heart of the work is contained in Chapters 4 and 5. Chapter 4 begins with a sketch of the Estonian phonological system and a discussion of compounding in Estonian. This is followed by a discussion of the corpus, the procedure for extraction of the data, and its subsequent analysis.

The corpus study of English compounds is presented in Chapter 5. The selection of the corpus and the procedure for compound extraction are explained first, followed by a discussion of the analysis of the data. Specifically, the discussion focuses on the results of a statistical analysis of some phonological and distributional variables that account for much of the variability in the orthographic form of English compounds.

Finally, Chapter 6 recapitulates the principal findings of the study, and discusses their implications and possible applications.

Chapter Two

DEFINING COMPOUNDS AND THEIR COMPONENTS

2.1 Defining "word"

Since morphology is, in part, about word formation, it is essential in any morphological study to first establish what a word is, or at least, what one intends it to be. To some, the notion of word may seem straightforward. For example, among literate populations, one might think of a word as a string of letters without spaces, which is pronounceable and carries meaning. But certainly words exist independent of writing systems. Children learn words without knowing how to read or write. For that matter, so do many adults. Furthermore, new words and even entire lexicons have been created absent a writing system.

In search of a definition for "word", the painstakingly descriptive volume of Marchand (1969: 1) seemed a logical starting point. Marchand defines a word as "the smallest independent, indivisible, and meaningful unit of speech that is susceptible of transposition in sentences." This sounds reasonable. However, the meaning of the term "word" largely depends on one's perspective. The *Lexicon of Linguistics* (Kerstens, Ruys, and Zwarts, 1995-2005, http://www2.let.uu.nl/UiL) accordingly defines "word" by domain:

> *i. in MORPHOLOGY: words are morphological objects which may but need not be the output of processes of affixation and compounding. (cf. Aronoff, M. (1976), Selkirk, E.O. (1982), Spencer, A. (1991))*
>
> *ii. in SYNTAX: words are generally considered atomic elements: they are the indivisible building blocks of syntax, which may be the input but not the output of syntactic processes, their parts presumably being inaccessible for syntactic rules.*
>
> *iii. in PHONOLOGY: words are phonological objects which constitute the domain for lexical phonological rules.*

Marchand's (1969) definition insightfully encompasses two of the three domains, but ignores the notion of the phonological word. Defining a word in terms of its prosodic boundaries typically does not figure into discussions of word formation. This is not to say that it should not. But in general, it does not seem to matter how one uses the term "word" as long as the user specifies precisely how it is intended.

Bauer (1983) uses the term "word" to blur the distinction between word-forms and lexemes. Word-forms, according to Bauer, are the oral or written realizations of lexemes. A lexeme by contrast is the abstract reference of a word-form. So by Bauer's account, the word-forms "words" and "word" refer to the same lexeme "word"; likewise, "saw", "see", and "seen" all refer to the lexeme "see". Following Bauer, we refer to *words* in the context of this study as "word-forms" or "words". In addition, we distinguish between a "word" or "word-form" and a word in the phonological sense, i.e., a "phonological word" or "prosodic word". A prosodic word may, but need not, be a word in the morphological or syntactic sense. Contractions such as *John'll* or *should've*

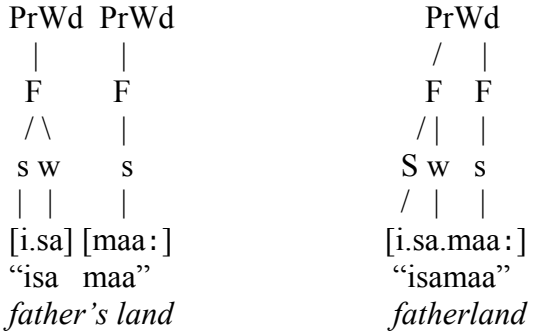
do not fit into a single syntactic category, nor are they the result of morphological rules (DiSciullo and Williams, 1987: 106-107). However, they are prosodic words. In phonological terms, a prosodic word must consist of at least one foot and should contain at least two moras or two syllables (McCarthy and Prince, 1995: 321).

2.2 Compounds

2.2.1 Defining "compound"

Compounding is a process that is seemingly simple enough for even the non-expert to describe. If you ask an Estonian what a compound ("liitsõna") is, he or she would most likely say something like "two or more words combined to form one word" or alternatively, "two or more words *written* as one word". And he or she would be correct, at least in the context of Estonian. This is because, by convention, all Estonian compounds are written without spaces between the lexical constituents. "Dvandva" compounds, a designation borrowed from the Sanskrit word meaning "two and two" and describing a conjunction of two independent elements (Spencer 1991: 311), are an exception and are hyphenated: *sekretär-masinakirjutaja*, "secretary-typist" (Erelt, Kasik, Metslang, Rajandi, Ross, Saari, Tael, and Vare, 1995: 476). And occasionally a writer might separate longer compounds (i.e., of three or more components) for greater clarity. But as a rule, compounds with a modifier + head structure, the vast majority, are always orthographically contiguous. Furthermore, there are prosodic differences between

compounds and phrases. As a result, there is no confusion, in the mind of an Estonian speaker as to whether a given string of nouns is a compound or a syntactic phrase. The rules seem clear whether they are reading or listening. Take, for example, a minimal pair such as “isa maa” (father’s land) and “isamaa” (fatherland). The written form identifies the former as a syntactic phrase and the latter as a compound. In spoken language, the intonational contrast does the same: in both cases, stresses fall on the first and third syllable [i.sa.maa:], but in the compound the first syllable has more prominent stress than the third [ˈi.sa.ˌmaa:], and in the syntactic phrase the stresses have equal prominence [ˈi.sa. ˈmaa:]. Duration is a factor as well since the compound is a single prosodic word in this case and would normally be pronounced more quickly than the phrase. The two structures are shown below (“PrWd” = Prosodic word, “F” = foot, “s” + strong syllable, “S”= strong syllable with more prominence than “s”, “w” = weak syllable).



The problem of defining compounds in English is not as straightforward. For example, a typical English speaker might also say that a compound (in the linguistic sense) is a combination of two words, or perhaps even two words that are written as one word. But in the case of English, the latter would not be accurate. While it is true that

two words may be written together to form a compound, there are many compounds that are not spelled as one word and there are some which are optionally combined. Whether or not a native speaker of English recognizes open compounds (e.g., "tax return" or "home plate") as compounds is not the issue, however. The greater question is how we really define compounds and whether the form, be it phonological or orthographic, has anything to do with it.

A review of the literature on compounding reveals a multitude of characterizations. Depending on the definer's perspective, a compound may be deemed a product of syntax, semantics, phonology, and/or morphology.

Marchand (1969: 11) defines compounds in terms of "determinant" and "determinatum". Semantically, the determinatum is limited by the determinant. For English and languages with similar morphological structure, in a given wordform XY, where X and Y are morphemes, X=the determinant and Y= the determinatum. According to Marchand, if both X and Y are free (i.e., independent) morphemes, then XY is a compound. In addition, Marchand maintains that stress position is a criterion for compounds.

Bauer (1983: 12) characterizes a compound as "a lexeme containing two or more potential stems that has not subsequently been subjected to some derivational process." Bauer views compounding, therefore, as a morphological process, and he takes issue with Marchand's (1969) position that stress is criterial for compounds. The distinction between compounds and syntactic phrases on the basis of stress location has been posited by a number of scholars, including Marchand, who takes the rather extreme view that a lexical group cannot be a compound in English if primary stress is not on the first

constituent. He uses the example (1969: 21) of "black bird" (a bird which happens to be black) and "blackbird" (a type of bird) to point out that the latter has forestress (*blá̀ckbìrd*) and is thus a compound, while the former does not have forestress (*blà̀ckbírd)* and so cannot be a compound. It should be pointed out that he in no way attributes this dichotomy to the orthographic form. Though the contrast of *blackbird* and *black bird* is indisputably clear in terms of stress, Marchand goes on to say that *black market* is not a compound either since the primary stress is not on "black". Few linguists would support such a narrow view (Bauer 1983, Sadock 1998, Fabb 1998, Aronoff 1994). Bauer points out that it is precisely because of the "black market" variety of compounds that stress cannot be a criterion for identifying these complex forms.

Syntactic criteria for compounds are generally not controversial. Compounds are expected to behave as single words (Bauer 1983, Ryder 1994). Among the tests for wordhood discussed in the literature are uninterruptibility (**black old market, *potato new peeler*), and modification of the whole as opposed to a part (**a very black market* , **a starchy potato peeler*) (Ryder 1994, Bauer 1983). Using these tests, *blà̀ck márket* and *potáto pè̀eler* are compounds, though they have opposite stress patterns.

Anderson (1992: 293) places compounding within the syntactic component of the grammar, claiming that compounds exhibit word-internal structure in that "elements of compounds typically fill argument positions in the semantics of other elements". Anderson also introduces the notion of analogy, suggesting that some compounds, specifically pseudo-compounds, which are combinations consisting of at least one bound morpheme (e.g., "cranberry", "Afro-American", and neoclassical compounds), may be formed analogically on the basis of other compounds (1992: 297).

Semantics has also been touted as the key player in compounding. There have been arguments for a semantic basis for compounds, with criteria ranging from noncompositionality or "idiomatization" (e.g., *guinea pig*) to permanence (e.g., *water bug* or *earthworm*) (Ryder, 1994: 14-15). Noncompositionality is unarguably an indicator of compound status, though clearly not criterial. Noncompositional compounds often bear the so-called "compound stress" (Chomsky and Halle, 1968) prescribed by Marchand, but so do forms such as *bookcase* and *wristwatch*, which are obviously compounds and completely transparent. In the broadest semantic terms, compounds may be "compositional" (transparent), as the latter two examples, or "noncompositional" (opaque). Studies involving a semantically-based approach generally look at the specific relations between constituents. This is a challenging task, and may be hindered by the opacity of some compounds.

The importance of a semantic characterization of compounds has been acknowledged in the domain of language technology. The problems of machine interpretation and/or translation of compounds have convinced many researchers that syntactic models alone do not suffice. The trend seems to be towards a combined syntactic/semantic model (ten Hacken 1999). Ten Hacken's model will be discussed in Chapter 3.

2.2.2 Compound Classes

There are various classes of compounds described in the literature. Following

Spencer (1991) and Bauer (1983), the distinguishing characteristics of these classes as they relate to the present study are outlined below.

First, there are three general classes related to syntactic headedness: endocentric, exocentric, and dvandva. Endocentric compounds are the typical modifier-head type, such as "candy bar" or "toothache". This is the most common type in English. Exocentric compounds are headless because, generally, they are non-compositional or "opaque": *guinea pig, moonshine*. The dvandva type of compound, mentioned earlier, may also be called a "copulative" (Bauer, 1983: 31) or a "coordinate" compound (Spencer, 1991: 311), as two "equal" parts are joined in one composite expression: "singer-songwriter", "secretary-typist". Since the components are equal, this type could be interpreted as having either two heads or no head. The operative word here, however, is "equal", whatever else that entails.

Another important class of compounds, which may be subsumed by the endocentric group, is synthetic compounds. These are verbal compounds and are most commonly manifested as one of the following structures:

$[[\text{NOUN}] + [\text{VERB+er}]]_n$ $[[\text{tax}]_n[\text{payer}]_n]_n$

or

$[[\text{NOUN}] + [\text{VERB+ing}]]_n$ $[[\text{decision}]_n[\text{making}]_n]_n$

Finally, "neoclassical" compounds are designated as such based on the nature and origin of the elements to be combined. That is, Greek and/or Latin roots combine to form words. If one defines compounding as simply combining two or more roots, then combining classical roots, which are usually bound morphemes, would result in a

compound, as in “sociopath” or “psychosomatic”. But the status of neoclassical combinations as compounds is contentious (Bauer, 1983: 216), and this type of “compounding” is excluded from the present study.

2.2.3 Overlapping Structures: Defining the boundaries of compounding

A number of scholars, including ten Hacken (1994) and Booij (2004), have noted that the line between derivation and compounding is somewhat blurred. For example, Booij argues against Anderson (1992), who draws a distinction between derivation and compounding, claiming that once an affix has been attached, the internal structure of a derived word is not accessible to the grammar. Since linking particles, such as those found in German (e.g., *-s-* and *-en-*), have access to the internal structure of compounds, Anderson contends that the two processes operate on different principles. Booij’s counterclaim is that the internal structure of derivations is in fact accessible to the grammar. Specifically, he cites the example “nominee” (2004: 24), which is derived via the process of truncation from “nominate”. He further points to the potential confusion between free and bound morphemes of the same form (e.g., “radio”, “counter”), as well as to the problem of suffix-like words that occur in compounds, such as “like”, “worthy”, and “monger”. Booij calls the latter “affixoids” (2004: 5). The essence of Booij’s argument is that derivation and compounding are overlapping processes and cannot be accounted for separately.

While ten Hacken (1994, 1999) notes the problem of a fuzzy boundary, his approach to the problem is quite different from that of Booij (2004). Motivated by his work in natural language processing, ten Hacken has worked extensively on this topic. In acknowledging the overlap in the processes of compounding and derivation, he decided that defining these boundaries was crucial both for analysis within theoretical linguistics and for the successful development of NLP systems. Ten Hacken (1994: 299) ultimately proposed three defining distinctions:

1. Compound elements can be either the head or non-head of the structure, while prefixes and suffixes are of a different nature.
2. Headedness does not play a role in derivation, whereas for compounds the whole is a subset of the head (i.e., assuming a right-headed structure, $[[N]_x[N]_y]_z$, Z is a subset of Y).
3. Virtually any pair of nouns can be interpreted as a compound, whereas in a derivative, the relation between a functor and its argument is fixed.

Ten Hacken's conclusions clearly hinge on the idea that compounds have a modifier + head structure. This follows from his claim that only "headed" compounds, which he dubs "H-Compounds", are actually compounds. For example, he calls the exocentric variety "derivations" (1994: 141). This approach might be viewed as a convenient case of excluding classes that do not fit the theory. This point will be discussed in greater detail in the next chapter.

It is clear, in any case, that defining the boundaries of compounding is a challenging and necessary endeavor. But the boundary between compounds and derivatives is not the only gray area in need of clarification. In the case of open

compounds, determining what can be called a compound vs. what should be called a syntactic phrase is an equally daunting task. This distinction has been attributed to a stress criterion (Burstein 1992), a position that seems to oversimplify the issue.

In order to avoid the gray area of the adjective+noun combinations in English (e.g., “modern ballet”, “old man”, etc.) and obtain a less controversial set of compounds, this work focused on the noun+noun variety of nominal compounds. The latter criterion was maintained for the Estonian data as well, for the sake of consistency.

In the present study, all “nounnoun” combinations in the English and Estonian corpora were accepted as compounds, whether they were endocentric or not. Likewise, all English “noun-noun” combinations were accepted once the part-of-speech of each constituent was verified. In addition, all English “noun noun” sequences which were not part of separate constituents were accepted as compounds, following Bauer (1998), who takes the position of a “lumper” (1998: 65), treating all noun+noun collocations as compounds in the absence of sufficient evidence to the contrary. Constituency was judged by looking at the “noun noun” plus the following word. In this way, the noun sequence “bedtime shop” followed by “talk” would be rejected, for example.

2.3 Collocations vs. n-grams

In NLP, and perhaps especially in corpus linguistics, there is a particular vocabulary used in addition to the vocabulary which comes out of theoretical linguistics. For instance, any corpus study will likely discuss “collocations”, “bigrams”, “trigrams”

or “n-grams”. In this study, the terms *collocation* and *trigram* will be used in the context of the English data. As such, some explication of these terms is warranted here.

While it is true that people create new utterances all the time, it is probably equally true that people often repeat what they hear (Bock, 1986; Pickering and Branigan, 1999). Rather than reusing entire utterances, we reuse phrases or strings of words and combine them to form new sentences. Words that frequently occur together are generally referred to as “collocations”. Since this designation is based on frequency of use, corpus linguistics and NLP are its natural domains. But statistical analyses of collocations extracted from large corpora can provide insights valuable to both applied and theoretical linguistics.

What one calls a “collocation” may vary (Nesselhauf 2004), but it certainly extends beyond a pair of adjacent words. It happens that the most commonly co-occurring words are function words like “of the” or “in a” (Manning & Schütze, 2000), which, linguistically or otherwise, may not be particularly interesting. Thus, a distinction should be made between simple *n-grams* and *collocations*. Let us take, for instance, an n-gram where n=2, i.e., a *bigram*. A bigram is a sequence of two words. For example, in the preceding sentence, we can find the following eight bigrams:

1. a bigram
2. bigram is
3. is just
4. just a
5. a sequence
6. sequence of
7. of two
8. two words

Bigrams may contain elements of two different syntactic phrases, as in number 2 above. Likewise, bigrams may have no particular relevance to prosodic structures. The term "collocation", on the other hand, is used in the NLP literature (Manning and Schütze, 2000) to characterize co-occurring content words, open compounds, idioms, phrasal verbs, etc. Collocations may well be bigrams (e.g., "real estate", "small business", etc.) since as a rule they consist of contiguous words[7]. But they may just as easily be trigrams (i.e., n-grams where n=3), such as "real estate developer" or "small business loan". Even longer n-grams are possible, depending on the researcher's goals. In obtaining the data set for this study, collocations consisting of two nouns were targeted.

The constituents of closed and hyphenated compounds such as "football" and "self-determination" are already unambiguously connected by virtue of their respective concatenation and hyphenation. Open compounds, conversely, are not orthographically linked, but may be discovered in a corpus by targeting noun collocations.

2.4 Concatenation

In this study, the term "concatenation" is used to refer to the orthographic alignment of words. For example, in English, the words "book" and "case" are concatenated to form "bookcase", as opposed to "book cover", in which "book" and "cover" are not concatenated.

---

[7] Manning and Schütze (2000: 158) also count non-contiguous expressions as collocations, so that "knock…door" would be a collocation, despite the necessary intervening material.

The two languages to be considered here – Estonian and English – differ markedly at the level of orthographic concatenation. That is, Estonian concatenates its compounds freely, while English does not. In Estonian, one can almost equate concatenating words with compounding, while in English many compounds are not concatenated. If we can say that, for English at least, compounds do not need to be orthographically linked, can we also say that concatenated forms are necessarily compounds? What does concatenation entail?

Languages like Swedish and German use compounding extensively and spontaneously (Clark, 1993: 154-159). Compounds with multiple components are frequent and nearly always concatenated (Mellenius, 1997: 23; Sproat, 1992: 41). There is no limit on the number of components in these languages (Bauer, 1983: 66). Are these really morphological units? Or does concatenation have a different function in these languages than it does in English? According to a website devoted to the subject of long words (Miller 2005, http:// members.aol.com/ gulfhigh2/words11.html), citing Guinness 1996 as its source, the longest Swedish word is the following:

> ***NORDÖSTERSJÖKUSTARTILLERIFLYGSPANINGSSIMULATORANLÄGG-NINGSMATERIELUNDERHÅLLSUPPFÖLJNINGSSYSTEMDISKUSSIONS-INLÄGGSFÖRBEREDELSEARBETEN,*** ("preparatory work on the contribution to the discussion on the maintaining system of support of the material of the aviation survey simulator device within the northern Baltic coast artillery")
>
> (130 letters, 44 syllables, 19 words)

Similar constructions can be found in German. The following example from Bauer (1983:67) is slightly less cumbersome, but still a bit shocking for an English speaker:

ÜBERSEEREICHWEITENFERNSEHRICHTFUNKVERBINDUNG
Über+see+reich+weiten+fern+seh+richt+funk+verbindung
("over sea reach distance distant see direction radio connection" or "overseas range television microscope link")

(44 letters, 13 syllables, 9 words)

While these are rather extreme cases even for Swedes and Germans, they do underscore the lack of consistency in compound forms cross-linguistically.

English does allow lengthiness in derived forms, particularly in the medical domain. For example, *pneumonoultramicroscopicsilicovolcanoconiosis* (45 letters, 19 syllables) – "a lung disease caused by breathing in certain particles" - is purported to be the longest word in any English-language dictionary (Fromkin, Rodman, and Hyams, 2003: 71). And for those of us outside the medical profession, there are also opportunities to challenge our short-term memories. *Floccinaucinihilipilification* (29 letters, 12 syllables) - "an estimation of something as worthless" - has apparently been used publicly by prominent political figures (http://en.wikiquote.org/wiki). However, unlike its Germanic cousins and Estonian, English limits concatenation of nouns to just two. Perhaps this trend will change due to the influence of the Internet. While English-speakers would probably be baffled by long strings such as the Swedish example, they generally do not have a problem with URLs. While some domain names use underscores or hyphens to separate words, they are usually concatenated. The owner of the following domain name (Dr. Subrahmanyam Karuturi) claims it to be the longest there is (63 letters, 20 syllables, 15 words):

Iamtheproudownerofthelongestlongestlongestdomainnameinthisworld.com

(from Miller 2005, http:// members.aol.com/ gulfhigh2/words11.html)

Chapter Three

A REVIEW OF KEY ISSUES

As in any comprehensive study, a review of the existing literature related to the topic is necessary at the very least to provide context to any claims which may be put forth. This section is organized into four parts, each devoted to an issue central to this work. First of all, several works on compounding are discussed, specifically those that attempt to define the process. Second, recent work on morphophonological variation is considered. This is followed by a discussion of isochrony, a phenomenon that has been the subject of several studies on Estonian prosody, and which may play a role in the variation of Estonian compounds. And finally, we examine some of the current views on orthography within linguistics.

3.1 Other work on compounding

3.1.1 Bauer (1983, 1998)

Bauer has written extensively on English morphology, giving considerable attention to compounding. Bauer (1983) provides a comprehensive account of the compounding process and the various types of compounds which occur in English. His

approach is simple and straightforward, focusing largely on the morphological qualities of compounds. Of particular interest in this work is Bauer's argument against the position of Marchand (1969) and others, who claim that compound stress is a criterion of compounds. Bauer disagrees with Marchand, who essentially argues that any noun collocation without compound stress is not a compound but a syntactic phrase. Marchand's (1969: 28) one exception to this is his acknowledgement that pronoun-noun combinations such as "self-determination" and "self-esteem" are compounds even though they do not take compound stress.

Distinguishing between compound nouns and syntactic phrases can be a challenge, so explaining this dichotomy based solely on stress contours would of course be an attractive solution. Unfortunately, it is an inadequate one. It is certainly true that the typical noun compound bears compound stress – e.g., BOOKcase, SUITcase, etc. However, one need not look too hard to find the "atypical" sort, such as "world WAR" or "home RUN", which could not reasonably be denied compoundhood. Bauer (1998) points to counterexamples such as these to make the argument, as mentioned earlier, that proposed criteria for distinguishing compounds from syntactic phrases do not reliably delimit the two sets.

Bauer (1983) refers to compound stress as "single stress" and phrasal stress as "double stress". The implication here is that noun collocations either get main stress on the left component or equal stress on both components. Other accounts describe this opposition instead as left vs. right stress or "initial" vs. "final" stress, including Bauer (1998). But the important point to make is that some of these collocations have a more prominent stress on the left component (i.e., "compound stress", henceforth "CS"), while

others do not. Bauer offers numerous examples of compounds which do not have CS, and even identifies a semantic class which requires what he calls "double stress". Specifically, the class of compounds which bears the relationship material/object, as in "mahogany table" or "cherry brandy" (Bauer 1983: 108), does not get CS.

Bauer (1983) further argues against the notion expressed by Marchand (1969: 24) that CS occurs when there is an *implicit* contrast being made, as in "day time" vs. "night time" or "milkman" vs. "mailman". Bauer contends that neither "cherry brandy" nor "apricot brandy" gets CS, and that they are thus counterexamples to this. Clearly, in a situation where the contrast was conscious and intentional, the stress contour would shift, producing *CHERRY brandy* and *APRICOT brandy*. But, in such cases, the placement of main stress on the left constituent should not constitute compound stress. Bauer's contention concerns implicit contrast rather than intentional contrast, and his counterexamples are therefore justifiable. In addition, he maintains that compound stress is not lexically motivated (i.e., it is not a function of constituent type), citing examples such as "TROOP leader" and "world LEADER" (1983: 107-108). In Bauer (1998: 70), it is further claimed that there is rarely a consensus on where stress should be placed in a given noun+noun collocation since there may be inconsistencies in stress judgments elicited from groups of speakers and even from individual speakers. Moreover, Bauer states that even dictionaries do not always agree on stress locus. As a result, he asserts that stress cannot be a criterion for compounds.

While Bauer (1983: 102-112) calls the view "defective", the idea that compound stress is required in compounding has been maintained by other scholars. This notion is central to Burstein's (1992) dissertation.

3.1.2 Burstein (1992):

Burstein (1992) argues that nominal compounds are syntactic objects. This claim is based primarily on her observation that, cross-linguistically, the headedness of compounds corresponds to the structure of syntactic phrases expressing modification. For example, in English, typical noun phrases such as "beautiful children" or "warm weather" have a modifier-head construction. Noun compounds in English, typically, have the same right-headed structure: "goldfish", "flight attendant", "credit card". At the same time, French noun phrases are generally left-headed: *les yeux verts*>the eyes green>"green eyes", *un homme faible* >a man weak>"weak man"). And similarly, French noun compounds are left-headed: *homme-grenouille*>man frog>"frogman", *programme machine*>program machine>"computer program". Burstein clearly establishes the structural correspondence between syntactic phrases and noun compounds. She further notes that compounding is more productive in languages that favor right-headed nominal constructions. This would seem to be true since French and its sister languages count few compounds and are left-headed, while in the right-headed Germanic languages compounding is highly productive. Finnish and Estonian also fit this pattern, as they are both strong compounding languages and favor right-headed noun phrase structure.

In addition, Burstein (1992) makes a peripheral observation that is perhaps more interesting and relevant to the current study; specifically, she raises the issue of

compound length. In her analysis of compound forms Burstein includes the structure [NP N], acknowledging that NPs may be quite long. Since this would open the door for potentially cumbersome strings, she contends that there must be some length limitation, stating simply that the likelihood of "acceptable" compounding decreases as the length and complexity of the nonhead increase. Despite the fact that a language like German seems to allow compounding ad infinitum, the idea of a lexical "maximum" certainly is plausible. Furthermore, it begs the question of what we can reasonably call a compound, or a word for that matter. Does concatenation of words equal compounding? As Burstein (1992) only touches upon this topic, it will be set aside here and addressed later in this chapter (section 3.4.2).

Another key component of Burstein's study is her analysis of the role of stress in distinguishing compounds from phrases. Her principal argument is that when the modifier in a modifier-head construction is functioning as a noun, whether or not it looks like a noun, it gets compound stress. If the modifier functions as an adjective, whether or not it looks like an adjective, then it takes phrasal stress. For instance, she posits that ambiguous constructions such as "legal work" receive stress assignment according to the function of the adjective/modifier. In this case, the word "legal" can clearly be interpreted as an adjective so that "lègal wórk" would mean work that is legal, as opposed to illegal. By this interpretation, primary stress should be on "work". However, "legal" can also be used as a noun, in the sense of the "legal department" of a company. In the latter case, "légal wòrk" could refer to the kind of work a lawyer does. Thus "legal" functions as a noun here and stress is assigned accordingly, on "legal". This seems an

important distinction and a sound one. There are other issues raised in relation to stress which are somewhat more controversial, however.

Burstein (1992) claims that proper nouns do not have compound stress and are thus not compounds, "unless otherwise stated". She later goes on to say that "Main Street' is a compound since it has compound stress, but "Madison Avenue" does not have CS and is not a compound. Burstein attempts to explain this with her "Broad Reference Head Rule", which states that "no narrow reference heads may appear in nominal compounds, where 'narrow' implies that a broader head exists" (1992: 106). A "Broad Reference Noun" is defined as "a noun which is underspecified" (1992: 105). It may be a head noun or a modifier, but in relation to stress, Burstein refers only to those cases where the Broad Reference Noun is the head noun. She explains that in "Main Street" the head noun "street" can substitute for the more specific or narrower "avenue", "lane", etc. This broad vs. narrow dichotomy may be a factor in the prosodic shape of the word, but to claim that this prosodic distinction entails that one is a compound and the other is a noun phrase seems rather overstated. Semantically, these are of the same type. Could we really claim, for example, that "Greenwich Street" and "Greenwich Avenue" have different internal structures?

In her discussion of stress assignment in three and four-part English compounds, Burstein (1992: 52) posits that primary stress is on the first and third components, respectively. While many compounds of the type [[N1 N2] N3] may conform to the stated formula – e.g., "SEAT-belt fastener", "AIRplane pilot", "BASEball player" – those with the structure [N1 [N2 N3]] (e.g., "arena FOOTball", "company CHAIRwoman") do not. Citing "data acquisition board" as an example of a three-part compound, Burstein

claims that “data” receives main stress. Prominence in this case may be speaker-dependent, but it is certainly plausible that the third syllable of *acquiSItion* receives main stress. Assuming the latter pattern, *data acquisition board* would seem to follow the stress pattern Burstein attributes to a four-part compound. This appears to be a function of its foot structure:

```
  FOOT    FOOT       FOOT    FOOT
   /  \    /  \       /  \      |
   σ  σ    σ  σ       σ  σ      σ
   |  |    |  |       |  |      |
[de . tə  . æ . kwɪ  .  zɪ . šən  . bɔɹd ]
( da . ta  . a . cqui  .  SI . tion  . board )
```

It actually consists of four feet, the prosodic structure often found in four-part compounds. A parallel example to this would be “tourist information office” ([ˌtʊ . ɹɪst . ˌɪn . fəɹ . ˈ**me** . shən . ˌɔ . fɪs]). Here again there is prominence on the third of four feet. “Cathode **ray** tube” [ˌkæ . θod . ˈ**ɹeɪ** . ˌtub] also gets main stress on the second constituent, which in this case coincides with the second and penultimate foot. What these cases show is that the placement of stress in three-part compounds cannot be attributed solely to the constituent position. A more likely scenario might be that it is the penultimate stressed syllable which receives prominence. This could also work with the *baseball player* type; though it is indisputably the first constituent that is stressed, this compound consists of two feet, the first being also the penultimate:

```
   FOOT      FOOT
    /  \      /  \
    σ  σ      σ  σ
    |  |      |  |
[bes . bɔl  .  pleɪ . əɹ]
(BASE . ball  .  play . er)
```

Fabb (1998: 79) presents a similar account to Burstein's for four-part compounds, attributing stress placement to constituent structure; his example also fits the pattern described above: *student essay RECord book.*

Burstein (1992) attempts to identify the distinctive characteristics of nominal compounds, and overall seems to accomplish the task of separating compounds from noun phrases fairly well. Nevertheless, the role of stress in determining what is or is not a compound has been overstated.

3.1.3 ten Hacken (1999):

As a result of his work in NLP, ten Hacken (1999) recognized the importance of delimiting the domains of morphological processes (i.e., inflection, derivation, and compounding) in a way that would hold cross-linguistically. With regard to compounding, he acknowledged its potential overlap with derivation. Thus, in order to distinguish compounding as a separate and unique process, he formulated the following definition (1999: 41):

> *A compound is a structure $[XY]_Z$ or $[YX]_Z$ such that:*
>
> > *1. the denotation of Z is a subset of the denotation of Y*
> > *2. if S is a possible way of specifying Y, the denotation of Z is determined by the range of S's that are compatible with the semantics of X*
> > *3. X does not have independent access to the discourse*

Condition 1 of the definition requires that the construction have a head, thereby excluding exocentric and coordinate compounds (ten Hacken 1994, 1999, 2004). Thus, he means to exclude "nut job" and "worrywart" from compoundhood since *nut job* is not a type of "job" and *worrywart* is not a type of "wart". It is not clear however, whether ten Hacken also intends to exclude "semi-opaque" expressions such as "bullfrog", where the head is transparent but the non-head is not. Applying ten Hacken's compound criteria, *bullfrog* is a subset of "frog" and would thus satisfy condition 1. By condition 2, for *bullfrog* to be a compound, the relation of X to Y must not be structure-dependent. Rather, the specification of "frog" should be determined by the semantic relation between the components "bull" and "frog". In other words, minus a context, *bullfrog* should be interpretable in several ways, based on its lexical components "bull" and "frog". If we could say, for instance, that a bullfrog is "a frog that is large", "a frog that is strong", etc., then it satisfies condition 2. Finally, condition 3 would be more difficult to test. Ten Hacken's test (1999: 47) for condition 3 goes as follows:

> *"Construct a discourse with the alleged compound Z in one sentence, and a pronoun unambiguously referring to the non-head of Z in the next sentence. If the non-head is not a proper noun and the discourse is correct, Z is not a compound."*

In a sentence such as, "The children saw the bullfrog in the pond, and they started throwing stones at it", one could perhaps say that "bull" cannot be referred to independently since its meaning is not transparent in this case. However, this conclusion

could not be reached on the basis of the Pronominal Reference Test (PRT). It simply doesn't apply here, since "bull" in this case refers not to "a bull" but to some bull-like characteristic(s). Ten Hacken's own examples (1999: 46-49) suggest a gap. To illustrate the test for conditions 1 and 2, he uses the example "baby duck", but for the PRT (for condition 3), he uses a different example, with no mention of whether or not *baby duck* was finally determined to be a compound. There are probably at least two reasons for this omission. One is that the non-head of *baby duck* (i.e., "baby") is polysemous; therefore, finding a context which would eliminate any potential ambiguity might be difficult. Secondly, assuming the most likely interpretation of *baby duck*, the non-head "baby" does not refer to a baby, but to a characteristic of a baby, namely that it is young.

The tests work nicely in many cases, but not in all cases. Nevertheless, ten Hacken's (1999) idea is interesting in that he excludes some of the more prevalent criteria posited for noun compounds. First, he excludes the stress criterion from his definition, noting that stress is subject to variation, both within and across languages. He also excludes the cohesiveness criterion (Bloomfield 1933: 232, Ryder 1994, Bauer 1983), which states that compound constituents cannot be split by other morphological material. For example, in English we can say "He is a sort of couch potato", but not "He is a couch sort of potato". Again here, ten Hacken points out that cohesiveness is not a cross-linguistic phenomenon, since compound components may be coordinated (i.e., joined with a conjunction) in Dutch and other languages. Ten Hacken (1999: 45) cites a Dutch example from Booij (1985): "zons- an maansverduisteringen" (sun and moon eclipses); Mellenius (1997: 24) provides a Swedish example, "pojk- och flickcyklar" (boys' and girls' bicycles).

While ten Hacken (1999) seems to be on the right track in attempting to create a language-independent definition for compounding, his downsizing of the domain does not account for longer compounds, as those found in German and Swedish, for example. Furthermore, there seems to be a contradiction in calling exocentric compounds "derivations", while classifying neoclassical composites as "compounds".

Coordinate compounds such as "writer-editor" or "murder-suicide" are also not compounds by ten Hacken's definition. They both can be excluded on the basis of condition 1, since they lack the required modifier + head structure. Specifically, a "writer-editor" is both a writer **and** an editor, not a type of editor. Similarly, a "murder-suicide" describes an event involving both a murder **and** a suicide, rather than a type of suicide. Ten Hacken's tests clearly place coordinate and exocentric compounds outside the realm of compounding. The question is, is that where they belong?

Historically, many exocentric compounds have evolved semantically from endocentric compounds, a phenomenon sometimes referred to as "semantic drift" (Aronoff, 1976: 43). Consider "swan song". A plausible folk etymology attributes this expression to Plato (Morris 1995-2005, www.word-detective.com) who claimed that the swan sings a beautiful song before it dies, in joyful anticipation of meeting Apollo, the god of music. In current usage, *swan song* more frequently refers to "the last words or last great works of a person" (Oxford American Dictionary, 1999). A similar example is "headhunter", which is defined as "a recruiter of personnel (especially for corporations)" (WordNet 2.1). Obviously, this has nothing to do with either heads or hunters. Another sense of *headhunter* is "a savage who cuts off and preserves the heads of enemies as trophies" (WordNet 2.1). The latter sense, which is completely transparent, was the

original meaning of the expression and is still in use. In short, *swan song* and *headhunter* are both endocentric and exocentric, and by ten Hacken's account, one sense would be a result of compounding and the other a result of derivation. He spells out the latter process as one applying an empty affix with an "agentive possessive" meaning to an NP or a noun compound. Ten Hacken's claim seems to be based on the premise that all exocentric compounds have such a structure. The examples he provides (ten Hacken 1994: 141), *loudmouth* and *birdbrain*, might well be interpretable as agentive possessive: a *loudmouth* is a person who speaks in a loud voice and a *birdbrain* is a person who has a brain like a bird. However, the same cannot be said of *swan song* and *headhunter.* They do not bear the same constituent relationship. While it would be fairly easy to identify an agent in *headhunter*, there is no obvious possessive relationship here. The semantics of *swan song* are even further removed since it refers to an event, rather than a person. Thus, if one accepts the idea of zero affixation in exocentric compounds, it would be necessary to account for a wider set of semantic relationships. The more opaque the compound, the more difficult this would be, as in a compound such as "guinea pig", which has two senses: one denoting a hamster-like rodent and the other a person used as the subject of an experiment (encarta.msn.com).

In terms of improving the efficiency of natural language processing systems, ten Hacken's (1994, 1999) focus on endocentric compounds may ultimately improve the handling of compounds. However, there are still some gray areas with respect to exocentric expressions. For example, it is not clear how the definition proposed can help distinguish computationally between a compound such as *headhunter,* and the "derivative" of the same form.

3.2 On "Free" Variation

Variation occurs at all linguistic levels and in all languages. The degree to which language varies at a particular level is language-dependent. For example, syntactic variation probably occurs more in so-called "nonconfigurational" languages (Hale 1978, Golumbia 2004) where the word order is relatively free. Lexical variation (i.e., synonyms) may be more pervasive in languages that are more widely spoken, such as Spanish and English. For example, Von Schneidemesser (1996) reported on a survey done in the United States on alternate terms for "soda" (i.e., carbonated drinks like Pepsi or Sprite). The results showed that *soda* could alternatively be called "pop", "soda-pop", "coke", "soda water", "cold drink", "bottle drink", "soft drink", "tonic", or "dope". Similarly, a study by Chambers (2000) on this phenomenon showed considerable variability among English-speaking Canadians. In any case, all languages contain some variation. Perhaps the most variable aspect of all in language is phonological. This is not surprising since phonology is concerned with the smallest of linguistic units – phonological features. Phonological variation may involve the use of two different phonemes, as in the two common pronunciations of the initial vowel in *economic* [ɛ.kə.ˈna.mɪk] vs. [i.kə.ˈna.mɪk]. Or variants may differ in a single segmental feature, as in *hot* with an unaspirated final /t/ ([hat]) vs. *hot* with an aspirated /t/ ([hat^{h}]), or a suprasegmental feature such as the stress in *harássment* ([hə.ˈɹæs.mənt]) vs. *hárassment* ([ˈhæ.ɹəs.mənt]).

The present study is concerned with the variation found in the compounding of English and Estonian nouns, whether in the written or spoken form. More than identifying the amount and type of variation, what is of concern here is how the choice of one variant over another is made and what generalizations, if any, can be drawn from this information. Anttila (1997) has studied this question in the context of Finnish, a close relative of Estonian.

3.2.1 Anttila (1997)

The phenomenon of free variation has received relatively little attention in the literature. But Anttila, in his 1997 dissertation, suggests there may be some reason to change this trend. Presenting an optimality theoretic account of free variation and statistical preferences in Finnish morphophonology, Anttila ultimately attributes free variation to a partial ranking of constraints, and makes a case that morphological variation and statistical preferences are driven by phonology. For instance, he establishes that the variation found in the Finnish genitive plural (e.g., the gen. plural of *naapuri,* "neighbor", may be realized as *naapurien* or *naapureiden*) is due to an interaction between vowel sonority on the one hand and syllable weight and stress on the other. This interaction, he claims, is only found in the context of variation. Thus, based on his findings, Anttila concludes that variant phonology reveals phonological patterns that invariant phonology does not.

Anttila (1997) also acknowledges that "free" variation is often not truly free in the sense that there are nearly always strong statistical preferences and that these preferences are the result of principled choices. In short, statistical preferences imply the existence of constraints. And as Anttila maintains, these constraints tend to be phonological. The Estonian compound data, which will be examined in Chapter 4 of the present study, support this claim. For English, the situation is slightly different, as the variation lies in the orthographic concatenation or nonconcatenation of constituents. Nevertheless, we intend to show that the variation found in English compounds is also at least partially constrained by phonology.

3.3 Isochrony:

The term "isochrony", from Greek *iso,* "equal" and *chrono,* "time" (OAD 1999), refers to the tendency in some languages for speakers to adopt a regular rhythm or equal timing of prosodic units (Catford, 1988: 182). According to Abercrombie (1967: 97-98), this rhythm helps to facilitate communication between speaker and listener, as the listener perceives the isochronous beats by means of "phonetic empathy".

Languages have been typologically defined in terms of their rhythm as *syllable-timed* or *stress-timed.* Japanese, for example, may be referred to as a *syllable-timed* language, which means that syllables occur at roughly equal intervals. English vowels vary in duration, so rather than being syllable-timed, English is a *stress-timed* language. That is, "prominent" or stressed syllables, as opposed to weaker syllables, occur at more

or less equal intervals (Abercrombie, 1967: 96-98; Clark and Yallop, 1995: 323). Thus, a speaker might utter two sentences with different syllable counts in the same amount of time, assuming each sentence has the same number of stress beats. As noted in Clark and Yallop (1995: 340) and in Abercrombie (1967: 98), this sort of isochrony can easily be achieved in poetry, for example. By contrast, it may not be as evident in conversational speech. As a result, the acceptance of isochrony as a phenomenon in English phonology is not pervasive (Bertran 1999). Support for isochrony in Estonian, however, is somewhat stronger.

Estonian syllables also vary in length. This is underscored by its phonemic quantity distinctions. While quantity is sometimes characterized in terms of weight, in Estonian it is generally perceived as a difference in length, so segments, and syllables containing such segments, are referred to as either short, long, or extra-long. Thus, as for English, syllable isochrony is not possible. Instead, it has been suggested that Estonian has either word or foot isochrony. The more accepted view is that Estonian tends toward foot isochrony. Studies on the Estonian quantity system seem to support this idea (Eek and Meister 1997, Gordon 1997, Lehiste 2004). For example, Eek and Meister claim that total foot durations in Estonian are essentially the same. So if we compare an extra-long monosyllabic word, e.g., “maa” (land), with a disyllabic word, e.g., “maja” (house), each of which constitutes a foot, the durations should be more or less the same.

One common thread in the literature on Estonian quantity is the emphasis on the importance of relative duration (Lehiste 1997a, 1997b, Krull 1999). There is no fixed duration for a syllable, but there are clear contrasts in the relative duration of syllables. This is true across speakers and also for individual speakers. It may be difficult to

distinguish between long (Q2) and extra-long (Q3) syllables pronounced in isolation since the duration of each may vary. However, put in the proper context, syllable quantity is much easier to perceive. Linguistic performance varies in all modules of the grammar, but perhaps most profoundly at the phonological / phonetic level. The relative values of acoustic phonetic features, however, remain remarkably constant and thus are far more revealing than absolute measurements.

Assuming a binary foot, the precise duration of a syllable in Estonian adjusts according to the other syllable in the foot. So ultimately, neither the syllable nor the foot would have a consistent duration. Krull (1997) cites this fact as evidence against isochrony. However, the idea of isochrony is that it is a tendency, rather than a hard and fast rule (Lehiste 1977). It is doubtful that anyone would suggest that every foot in a language could possibly occupy the same durational space. Moreover, it has been pointed out that what a speaker perceives is not necessarily equal to the phonetic reality of what was uttered (Lehiste 1977, Roach 1982, Tatham and Morton 2001). So one might actually perceive isochrony, though it is not there physically.

What is important here is that where isochrony exists in Estonian, it appears to be at the level of the foot. In Chapter 4, a connection will be drawn between this notion and variation in Estonian compound forms.

3.4 On Orthography:

If free variation has received little attention, orthography has been even more neglected. There has been considerable psycholinguistic research on orthographic effects

in language processing (e.g., de Jong, Feldman, Schneider, Pastizzo, and Baayen, 2002), but the general perception is that orthography is extralinguistic. With respect to compounding in particular, ten Hacken (1994) categorically rejects the idea that orthography warrants any consideration. Marchand (1969) claims that "spelling is of no help" in the analysis of compounding and that the spelling variants seen in English compounding show a "complete lack of uniformity" (1969: 21). Bauer (1983: 105-106) is slightly more vague about it, referring only to the stress dichotomy in English and saying simply that "spelling does not provide an accurate guide" to whether a compound takes stress on the left component or not. He points to the fact that some open and hyphenated compounds take CS, while others take phrasal stress (e.g., "**GAR**DEN party" vs. "garden **CI**TY", "**CAR**BON-paper" vs. "carbon DI**OX**IDE"). In his later work, Bauer (1998: 69) devotes a bit more attention to the issue of orthography, even suggesting a possible connection between word length and concatenation.

Despite the resistance of many, some recent work has emerged touting the linguistic relevance of orthography. Among the proponents of an "orthographically relevant level" are Sproat (2000), Neef (2002), Neijt (2002), and Noack (2002).

3.4.1 An Orthographic Level

Researchers in computational phonology, specifically those working on text-to-speech systems, have explored the relation between spelling and phonology, positing an orthographically relevant level (ORL) of the grammar. According to Sproat (2000: 177)

the ORL is where lexical representations are orthographically encoded. Sproat's ORL occurs "at a fixed point in the grammatical derivation where the derivation of the writing system branches off" (Neef, Neijt, and Sproat, 2002: 4). The "depth" or "shallowness" of the level depends on the language, but the level is consistent in that language. For English, the ORL is deep because of the lack of direct "phoneme to grapheme" correspondence. Several other researchers, including Neijt (2002), Neef (2002), and Noack (2002) also support the idea of an orthographic level, albeit with slightly different opinions on the precise mechanics of the model. But since our interest here is in the broader implications of an ORL, we will skip the details of the competing versions.

For the present study, Sproat's (2000) comments on the role of orthography in phonology are most relevant. In particular, he points out that spelling may directly influence the phonology of a language. To support this claim, Sproat (2002: 20) cites an example from Italian, indicating that the Northern Italian dialects, which historically had no geminates, have adopted the standard pronunciation for word-internal geminates, but not for gemination across words. Word internal gemination is reflected in the spelling, whereas gemination across words is not. This seems compelling evidence that orthography is not language-external.

A second piece of evidence offered by Sproat (2000: 20) with regard to the influence of orthography on phonology is that spelling may affect a speaker's judgments about how certain words are actually pronounced, so that "toe" and "tow" may not be judged to be homophonous. Assuming this is true, one might posit analogously that a speaker would judge a phonological difference in closed vs. open or hyphenated compounds in English. The difference here would lie in the position of main stress. For

instance, a closed nominal compound would likely be recognized as a single word and receive stress on the first syllable, as is typical for nouns. Open forms by contrast would not be perceived as single words and would receive phrasal stress. Compare "death ROW" and "DEATHbed", for example. Though it is not our intention to make the generalization that stress is wholly determined by such a perception (since there are numerous instances of nonconcatenated compounds with CS), it seems reasonable to suggest that it is a contributing factor.

3.4.2 The anomaly of s p a c e s

Perhaps one of the more intriguing questions related to the orthographic form of compounds is the striking variation cross-linguistically. While differences in spelling conventions are to be expected across languages, it is nevertheless rather puzzling that languages of the same type (e.g., two Germanic languages) should be so dissimilar in this process. As mentioned in Chapter 2, English differs markedly from its Germanic relatives in the form of compounds. An expression that would be called a compound in German or Swedish would by many accounts be called a syntactic phrase in English (Sproat 1992). A German example is provided by Sproat (1992: 230): *Lebensversicherungsgesellschaftsangestellter*, "life insurance company employee". The Swedish example given earlier (in 2.4) would clearly be an unlikely compound in English. Still, more mundane instances exist in Swedish: *småflicka* (små+flicka), "little girl"; *sjuttioårsdag* (sjuttio+år+s+dag), "seventy years' day" → "seventieth birthday";

skolbokshylla (skol+bok+s+hylla), "schoolbook's shelf" (Hedlund 2002). This issue is not dealt with in the literature in any depth. So the question remains – What does the concatenation of words entail?

De Jong et al. (2002) looked at processing latencies for Dutch and English compounds by means of visual perception experiments. Their experiments showed that Dutch compounds, which are all concatenated, are processed in the same way as English closed compounds, and differently than English open compounds. That is, for both English closed compounds and Dutch compounds, processing was affected by the position-related token frequency of the "morphological family" (e.g., the total count of all compounds with "man" as the second constituent). The finding that constituent position contributes to processing times suggests that the subjects recognized certain words as being likely to either begin or end a compound. De Jong et al. explain these effects in terms of a conditional bigram probability, computed peripherally to the lexicon, and they conclude that lexical processing is therefore sensitive to morphological structure.

In the case of English open compounds, position-related type frequency was a better predictor of processing latencies than position-related token frequency, and therefore, de Jong et al (2002) concluded that open compounds are processed differently than closed compounds. However, there was a common effect for all three: for English open and closed compounds, as well as for Dutch compounds, compound frequency was found to be a reliable predictor. Furthermore, it was found that English closed and open compounds had something else in common: only the left constituent showed significant effects, whereas in Dutch compounds both the left and right constituents showed effects.

The possible reason for this was not explored by de Jong et al., but it seems to suggest that Dutch compounds may be processed differently from English compounds in general.

To test the effect of the orthography, another experiment was conducted, inserting an artificial space between components of the Dutch compounds. Results showed that processing of the Dutch compounds did not change due to the space. De Jong et al. (2002) thus suggest that the processing of compounds may not be affected by orthography, so that perhaps the English "open compounds" were "phraselike". This, however, would not explain compounds such as "data+base" or "health+care", which vary quite freely in form.

The issue of whether and why English compounds are concatenated is one of the goals of this study. The following chapter on Estonian compounding is meant to establish a basis for positing that compound variation is largely driven by phonology. By comparison with English, the Estonian pattern is more traditionally within the domain of linguistics. It does not involve a dichotomy between open and closed forms because Estonian compounds, as a general rule, are closed. As has been acknowledged from the beginning, these two languages appear to pattern differently with respect to compounding. But as we will see, there are nevertheless some commonalities.

Chapter Four

ESTONIAN COMPOUNDS

4.1 Background

4.1.1 A few words about Estonian phonology

As a Finno-Ugric language, Estonian is characterized by a complex case system and trochaic stress pattern. There are fourteen overt grammatical cases in Estonian (Collinder 1969, Tauli 1973, Viks 1994a, Mürk 1997), while its sister language, Finnish, has at least fifteen, and Hungarian, another relative, has substantially more than that.[8] Through its history of foreign occupation, Estonian morphology has been influenced by various languages, including German, Swedish, and Russian. The influence of these languages can be seen most easily in the lexicon. In this respect, English has also been influential, but more recently.

In terms of morphological structure, Estonian is still generally grouped with the "agglutinative" languages like Finnish and Hungarian. However, it has actually moved away from this pattern, to some extent, in the direction of inflectional systems. Finnish, for example, makes wide use of clitics and possessive suffixes, and its morphemes are fairly transparent (Viks 1994b). The possessive suffix, as opposed to the genitive affix marking the possessor, is a morpheme which attaches to the object possessed. It is also found in Hungarian (Vago 1980), but absent from modern Estonian. Furthermore, while

[8] The precise number of cases found in Hungarian is somewhat controversial and is analysis-dependent (Vago 1980).

Estonian has clitics (e.g., the emphatic particle "-ki/-gi", as in *kesklinnaski* (*kesklinnas,* downtown-inessive,sg. + *ki,* even, "even downtown" (Oinas, 1975: 112-115)), there are fewer in comparison to Finnish (Viks 1994b, Kiparsky 2005: 30). In addition, through the shortening of many stems and affixes, Estonian morphology has become less transparent (Viks 1994b), a feature more typical of an inflectional language, as opposed to an agglutinative one (Sproat, 1992: 20).

What distinguishes Estonian even further are its three-quantity system of vowels and consonants, and the phonological phenomenon known as *gradation*.[9] Much of the research done on Estonian has focused on its quantity system. Gradation is relevant to this system in that one type of grade alternation is quantity alternation. Both of these phenomena play an integral role in the grammar of this language. Thus, one can hardly discuss anything that happens in Estonian without mention of quantity and gradation. In the present study, the morphological process of compounding in Estonian and its phonological/prosodic motivations are examined. But first some background on the phonology of Estonian is provided.

Setting quantity aside, Estonian has a nine vowel system (Tauli 1973), as shown in Table 4.1:

	Front		Back	
	unrounded	rounded	unrounded	rounded
High	/ i /	/ ü /		/ u /
Mid	/ e /	/ ö /	/ ɤ /	/ o /
Low	/ æ /		/ a /	

Table 4.1: Estonian vowel inventory

[9] Consonant gradation also exists in Finnish. But since Estonian has both consonant and vowel gradation, the process is more pervasive in Estonian.

The back, mid, unrounded vowel /ɤ/ is peculiar to Estonian and often represented in the literature as /õ/, the same symbol used in the orthography. The reluctance to use a standard transcription, such as the IPA symbol /ɤ/, may relate to the fact that there is some disagreement as to whether this vowel is mid or high. For example, Hint (1998: 95) classifies it as a high vowel; Tauli (1973: 13) and Erelt et al. (1995: 104) describe it as a mid vowel. The latter classification is adopted here. In addition to its nine monophthongs, Estonian also has a broad inventory of diphthongs.

The basic consonant inventory, again setting quantity aside, is /p, t, ţ, k, h, j, l, ļ, m, n, ņ, r, s, ş, v/ (Mürk, 1997: 3-8)[10]. The consonants are displayed in Table 4.2 according to place and manner of articulation. Note that while the phonemes /f/ and /š/ are counted as part of the consonant inventory, they are only used in foreign borrowings (Saagpakk 1955, Erelt et al. 1995).

MANNER	PLACE					
	labial	labiodental	alveolar	palatal	velar	glottal
fricative		/ v / / f /	/ s / / ş /	/ š /		/ h /
(oral) stop	/ p /		/ t / / ţ /		/ k /	
nasal	/ m /		/ n / / ņ /			
lateral			/ l / / ļ /			
trill			/ r /			
glide				/ j /		

Table 4.2: Estonian consonant inventory

[10] The consonant phonemes represented with a comma beneath them, as in /ţ/, are palatalized.

There are no voicing distinctions for stops in Estonian (Collinder 1969). "B", "d", and "g" appear in the orthography, but actually represent the short version (i.e., quantity 1) of the voiceless unaspirated stops /p, t, k/. In fact, all of the native phonemes enumerated above have corresponding long forms. Quantity differences in Estonian vowels and consonants are contrastive[11]. The following minimal triples illustrate this fact:

Quantity 1	GLOSS	Quantity 2	GLOSS	Quantity 3	GLOSS
sada [sata]	*hundred*	saada [saata]*	*send (imp.)*	saada [saa:ta]**	*get (infinitive)*
lina [lina]	*linen*	linna [linna]	*city (gen.)*	linna [linn:a]	*city(partitive,illative)*

Table 4.3: Examples of minimal triples showing phonemic quantity distinctions

*[aa] indicates a long segment (Q2); **the colon is used here to denote overlength (Q3)

Estonian nominal cases are realized through a complex system of stem alternations, particularly in the singular form of the three primary cases – nominative, genitive, and partitive. The form of a noun or adjective that would be used in the subject position of a sentence, and the citation form for dictionary listings, is the nominative. The so-called "genitive" case is actually much more than that, but this designation is used here to remain consistent with the literature. The Estonian genitive form may function as a possessive marker - *Jaani onu* (*Jaani*(John-gen.sg.) + *onu*(uncle)), as a direct object – *Mina sõin leiva ära*. (*Mina*(I) + *sõin*(ate) + *leiva*(bread-gen.sg.) + *ära*(all up)), or it can just serve as the modifier or nonhead of a compound – *piimapudel* (*piima*+*pudel*, milk-gen.sg.+ bottle, "milk bottle"), *lennujaam* (*lennu*+*jaam,* flight-gen.sg.+station, "airport"). As is evident from the latter examples, in compounds the genitive form is concatenated

[11] Overlength is only phonemically contrastive in the initial syllable of a word (Odden 1997).

with the noun(s) it is modifying - e.g., *rannapall* (*ranna*+*pall*, beach-gen.sg.+ball, "beach ball"), *lehepoiss* (*lehe*+*poiss*, paper-gen.sg.+boy, "paperboy"). The same is true of the nominative: *kingsepp* (*king*+*sepp*, shoe-nom.sg.+smith, "shoemaker"), *raudtee* (*raud*+*tee*, iron-nom.sg.+road, "railroad"). It can also be seen from these examples that neither the nominative nor the genitive singular form has a specific corresponding suffix. And although the genitive is more consistent than the nominative in the sense that it always ends in a vowel, the particular vowel is often unpredictable: where the nominative ends in a consonant, the genitive ends in either /a/, /e/, /i/, or /u/. Precisely how the vowel is chosen remains somewhat of a mystery (Viks 1994b).

There are numerous words which have identical nominative and genitive forms (e.g., *isa* (nom.), *isa* (gen.) "father"; *ema* (nom.), *ema* (gen.) "mother"). But more often than not, stem alternation does exist, resulting from a variety of phonological processes including gradation and deletion. The complexity of the system of stem alternations is evident in the following examples:

Nominative sg.	Genitive sg.	*Gloss*
laud	laua	*table*
vesi	vee	*water*
sepp	sepa	*smith*
koer	koera	*dog*
pidu	peo	*party*
hobune	hobuse	*horse*
aeg	aja	*time*
süda	südame	*heart*
aken	akna	*window*

Table 4.4: Estonian stem alternations

Gradation is a phonological process whereby a segment is either strengthened (fortition) or weakened (lenition). In Estonian, this change serves the grammatical function of case-marking. There are three general categories of gradation in Estonian: "quantitative", "qualitative", and "double" gradation. As the term suggests, "double gradation" is a combination of qualitative and quantitative gradation. This type typically occurs as a nominative / genitive alternation. A consonant is deleted from two-syllable words of the type (C)VCV, and the syllables are conflated creating a longer segment in the remaining monosyllable (e.g., *vesi* [vesi] "water" – nom.sg., *vee* [vee:] "water" – gen.sg.; *pidu* [pitu] "party" – nom.sg., *peo* [peo:] "party" – gen.sg.). In addition, this process is frequently accompanied by a lowering of the vowel, as in the latter example, *pidu/peo*, where the [i] of [pidu] is lowered to [e] in [peo:], forming the diphthong [eo:]. This deletion/conflation phenomenon in effect is a kind of compensatory lengthening (Lehiste 1997a), as the deleted consonant in this case is apparently being compensated for. It must be compensated for, since monosyllabic words must be in the third quantity. Deleting the consonant of the second syllable in such words would create a Q2 syllable (i.e., quantity 2). Since this would not meet the minimum word requirement for Estonian, the syllable has to be lengthened to quantity 3 or "Q3".

The question of how quantity relates to the mora has been entertained by a number of scholars, including Lehiste (1997a), Hayes (1995), and Prince (1980). The mora is by most accounts likened to weight. It is generally agreed also that a long segment is heavier than a shorter one, so that a long vowel would receive two moras while a short vowel would receive only one. In these respects, the notion of quantity

seems nearly synonymous with the mora. But within the prosodic hierarchy, the mora maps directly to the segment level below, linking it to the syllable level above. Although quantity was initially thought of as a function of the segment, current views tend to place it within the domain of the syllable (Hint 1997, Odden 1997). As Hint points out,

if we assign quantity at the segmental level, there would be no room for syllable quantity, as it would be reduced to "a sum of segmental quantities" (1997: 126).

Hayes (1995) draws a direct correlation between the mora and the degree of quantity, assigning one, two, and three moras to Q1, Q2, and Q3 respectively. The idea of a trimoraic syllable is of course controversial. Lehiste's (1997a) argument against this idea is convincing: Q3 and Q2 may vary greatly in duration so that in some cases a Q2 syllable may actually be phonetically longer than a Q3 syllable. The important distinction between degrees of quantity, according to Lehiste, is not the actual but the relative durations. That is, within the Estonian phonological system these degrees of quantity cue the listener by virtue of their relative differences. This holds especially for Q2 and Q3. The true durations vary among and within speakers, so they cannot be relied upon as acoustic cues, nor can they be directly associated with the mora as Hayes proposes. It is more likely that Q2 and Q3, the long and overlong syllable quantities, are both bimoraic. A further motivation for a bimoraic Q3 lies in the fact that the minimal word in Estonian is a binary foot – CVCV (Q1 + Q1) or alternatively, a Q3 syllable.

Finally, the Estonian stress system is fairly regular, and it is trochaic. The first syllable gets primary stress in virtually all native words and assimilated borrowings. Stress is persistent from left to right on every odd syllable. Estonian thus typically avoids "stress-clash" (Hayes, 1995: 322-329), i.e., the occurrence of adjacent stressed syllables.

However, the extra long syllable (i.e., quantity 3), which is stress-bearing regardless of its position, appears to be an exception to this pattern. According to Hayes, though, this can be explained in terms of Estonian foot structure. The minimal foot in Estonian and the most prevalent is binary, but ternary feet exist (*hobune* -"horse", *arvuti* -"computer"), as do unary feet (*linn* -"city", *sepp* -"smith", etc.). The latter are manifested as Q3 syllables. But Hayes equates the Q3 syllables theoretically with binary feet.

There is a compelling analysis of Estonian prosody in Hayes (1995). In his Chapter 8 "Ternary Alternation and Weak Local Parsing", Hayes offers a somewhat controversial account of Estonian quantity. Taking some of the insights of Prince (1980), he modified the framework from one based on foot structure to one based on the mora. His claim is that overlong syllables are trimoraic and are better explained in terms of moras than feet. Furthermore, a trimoraic syllable does not necessarily have to be considered a monosyllabic foot. In Hayes's analysis, the overlong syllable may behave as a disyllable, or it may not. Because of this, in his view, defining a Q3 syllable on the mora rather than the foot makes more sense. He claims, therefore, that if Q3 syllables are treated as bisyllabic, there is essentially no stress clash.

4.1.2 The structure of noun compounds in Estonian

Compounding in Estonian is highly productive. Multiple components are allowed - generally up to five, though the vast majority[12] contain only two or three. Estonian

[12] In the dataset used for the present study, 97% of the noun compounds identified were composed of two lexical components; the other 3% had 3 components.

compounds are formed by concatenating words. These may be from a variety of grammatical categories: *allmaaraudtee* - "subway" [[all]$_{prep}$"under" [maa]$_{n}$"ground" [raud]$_{n}$"iron" [tee]$_{n}$"road"]$_{n}$, *vanaema* – "grandmother" [[vana]$_{adj}$"old" [ema]$_{n}$"mother"]$_{n}$, *veripunane* - "blood red" [[veri]$_{n}$"blood" [punane]$_{adj}$"red"]$_{adj}$, *alahindama* - "to underestimate" [all]$_{prep}$"under" [hindama]$_{v}$"estimate"]$_{v}$. In this study, however, the investigation concerns specifically noun compounds of the form [[word 1]$_{n}$[word 2]$_{n}$]$_{n}$.

Noun-noun compounds in Estonian, with few exceptions, are formed with the first component in either the genitive singular or nominative singular; the second component is case-marked according to the rules of syntax. The choice of the nominative or genitive in the first component is not always predictable. In fact, sometimes both forms may be used.

A compound with a nominative first component will henceforth be referred to as a "nominative compound" or "NC"; similarly, a compound with the first component in the genitive will be referred to as a "genitive compound" or "GC". In addition, the first component will be called "W1" and the second component, "W2".

Tauli (1973) offers perhaps the most comprehensive linguistic account of Estonian compounding and concedes that "nominative and genitive compounds do not constitute clearly delimited groups" (Tauli, 1973: 176). Still, he suggests that there are some semantic parameters to help in choosing one form over the other. Tauli defines two broad categories; the first involves a substance/material relation and the second, a character/form relation. For example, when W1 describes the substance or material from which W2 is made, W1W2 is a nominative compound, as in *kulduur* (*kuld*+*uur*, gold-nom.sg.+watch, "gold watch") and *palkmaja* (*palk*+*maja*, log-nom.sg.+house, "log

house"). When W1 describes the external shape or character of W2, an NC is generally used as well, for example, *lillkapsas* (*lill+kapsas*, flower-nom.sg. + cabbage, "cauliflower"), and *nurksulg* (*nurk+sulg*, corner-nom.sg. +bracket, "square bracket"). Both of these categories show many exceptions, as Tauli (1973) attests.

Exceptions may result from a phonological constraint. For example, there appears to be a constraint against first component monosyllabic words ending in a consonant cluster (Tauli, 1973: 176) – e.g., *nartsunukk* (*narts*[u]+*nukk*, rag-gen.sg. + doll, "rag doll"), **nartsnukk*.

Further semantic motivations for the use of NC are restricted to small word groups. For instance, an NC will occur when W1 = {tali | kevad | suvi | sügis} (i.e., *winter*, *spring*, *summer*, or *fall*) or when W1 = {jalg | käsi} (*foot* or *hand*), the latter being used in the sense of operating W2 by means of W1: *sügisõhtu* (*sügis+õhtu,* fall-nom.sg.+evening, "fall evening"), *jalgratas* (*jalg+ratas*, foot-nom.sg.+wheel, "bicycle"). Yet even these microcategories are not without exception. Compare ***käsi****kohver* (*käsi+kohver,* hand-nom.sg.+luggage, "suitcase") to ***käe****kott* (*käe+kott*, hand-gen.sg.+bag, "handbag") where the nominative käsi and the genitive käe, respectively, are used in parallel contexts. Counterexamples can also be found for the "season" compounds; for instance, while "fall evening" (*sügisõhtu*) is an NC, "summer night" (*suveöö, suve+öö /*suviöö)* is a GC.

Beyond semantic explanation, Tauli (1973) posits that foreign borrowings effect a bias in compound form. That is, if W1 = an international word (or "IW" as Tauli calls it), then an NC is used: ***mootor****ratas* (*mootor+ratas*, motor+wheel, "motorbike"). In addition, when W1 = a noun ending in *–us*, the result is usually in the nominative:

majandus*ministeerium* (*majandus+ministeerium*, economics+ministry, "economics ministry"). Again, in both cases there are a multitude of exceptions: ***telefoni****raamat* (*telefoni+raamat*, telephone-gen.sg.+book, "telephone book"), ***saatuse****kaaslane* (*saatuse+kaaslane,* destiny-gen.sg.+mate, "soulmate").

The remaining compounds are generally GCs. According to Tauli (1973), there are many compounds which vary freely between the two forms. Tauli further claims that colloquial Estonian favors the genitive construction while more formal and technical domains favor the NC. This study does not bear specifically on this issue; however, evidence does emerge which suggests that Tauli's claim may be correct.

Notwithstanding Tauli's (1973) diligent effort to shed light on the nominative-genitive dichotomy in Estonian compounding, his account nevertheless leaves some aspects of the phenomenon unresolved. According to Tauli, there should be room for considerable variation. This portion of the study attempts first to find evidence that this variation exists, why and to what extent it exists, and then to determine the motivation for choosing one option over the other. The method used in this study is corpus analysis. The procedure for obtaining the data to be analyzed is described below.

4.2 The Corpus:

The University of Tartu (Estonia) Computational Linguistics Group has compiled and made available several corpora and a number of subfiles including frequency lists based on the corpora. The data on which the present study is based were extracted from a

frequency list of 10,000 lemmas (henceforth referred to as the "Tartu Frequency List"), drawn from one million words of the group's "Corpus of the 1990s". The latter consists of literary and newspaper texts of approximately 500,000 words each. The corpus was designed to reflect "standard widespread neutral Estonian literary language" (Kaalep & Muischnek, 2002: 1). To ensure that the word forms in the Tartu Frequency List fit this requirement, and also to minimize the chances of extracting proper nouns that might look like common nouns, Kaalep & Muischnek included only words occurring in both genres – i.e., in both the newspaper and literature texts. In addition, a lower limit of five tokens per type was set. A natural consequence of such criteria is that some high frequency words would not be included in the list because they appeared in one text genre but not the other (2002: 5).

4.3 Extracting the compounds:

Since the list contained lemmas of the original, there was no need for morphological analysis of the second components. Compounds were obtained by first running the frequency list through ESTMORF (Kaalep 1997), a morphological analyzer designed for Estonian. The freeware was created by Kaalep, a researcher from the Tartu University Computational Linguistics Group and co-founder of Filosoft, Inc. (www.filosoft.ee), a company which produces online linguistic tools for Estonian. ESTMORF identifies compounds by splitting the components with an underscore. The noun compounds were then extracted by means of a Perl program which we wrote for

that purpose. Though the resources available from Tartu University and from Filosoft, Inc. made this task much easier, it was not without its challenges.

Compounds were part-of-speech tagged as a unit, so the first component (W1) had to be analyzed separately. Since the goal was to extract all noun+noun compounds, W1's that were not analyzed as nouns were deleted. ESTMORF also identifies grammatical case. In instances where the nominative and genitive were identical, the form was labeled "nom/gen". Forms were checked manually for accuracy, but overall it was found that ESTMORF had a fairly low error rate. For example, among the 221 noun lemmas beginning with the letter "a", ESTMORF identified 98 as compounds. Of these, only one was incorrectly parsed – "atmo_sfäär" ("atmosphere"), showing an error rate of about 1%.

Another problem was character recognition. Since ESTMORF was built specifically for Estonian, it naturally reads Estonian orthography without a problem. However, when further manipulation of the data was necessary, some of the Estonian characters were not compatible with Perl. Specifically, those characters with diacritics had to be modified in order to be recognized. As Estonian has four vowels spelled with diacritics (õ, ä, ö, ü) and three consonants (š, ž, č), a substantial number of words were affected. The characters were encoded for Perl by replacing them with the diacritic (or a substitute for the diacritic) followed by the plain character. For example, *ä* was replaced with *"a,* so *äri_mees* was modified to *"ari_mees,* and so on.

4.4 Results and observations:

The above extraction process produced 1,094 different noun compounds with a combined token count of 18,714. A breakdown of type and token frequency by W1 category is provided in the table below. The results show that W1 is by and large either nominative or genitive. Occasionally, other cases such as the genitive plural, partitive plural, and illative singular occur in compounds, but such instances are relatively few, as can be seen in Table 4.5 below. In terms of type, the genitive is strongly favored, but the ratio of types to tokens is about the same for NCs and GCs, at approximately 1:15. Proportions of types and tokens in each category are shown in the pie charts.

	total	**# of GCs**	**# of NCs**	**# of GC=NC***	**Others**
# of types	1094	558	158	314	64
# of tokens	18714	8747	2388	6568	1011

Table 4.5: Compound frequencies by W1 component category/form
*"GC=NC" refers to compounds whose W1 has the same form in nominative and genitive

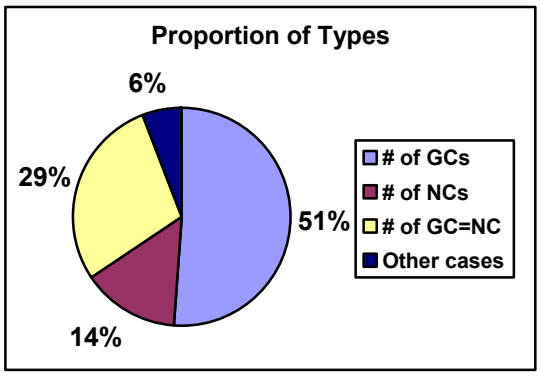


Figure 4.1a

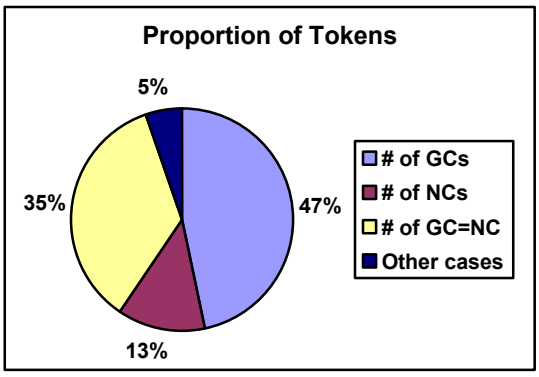


Figure 4.1b

Figure 4.1a and b: Proportion of 1,094 Estonian compound types and tokens by W1 category

Compounds were examined further in terms of source genre. The table below shows the frequencies of NC and GC types and tokens in each of the two general domains: news and literature. If we recall that the latter two subcorpora are of equal size, the figures suggest that compounds occur with greater frequency in news texts than in literary texts. This is not surprising: one of the advantages of compounds is that they are economical: they allow a writer to say more with fewer words. Depending on the semantic relationship of its constituents, the meaning expressed by a particular compound may be expressed in another way, though it rarely is. Since in some cases, Estonian does not have an equivalent word form of the English preposition "for", (as in "for the purpose of")[13], conversion from compound to syntactic phrase may be difficult or impossible in Estonian. For example, "supelrannamaja" ("beach house") might also be expressed as "maja supelranna juures" (a house at the beach), but converting "kontserdisaal" ("concert hall") is a little more complex -- "saal kus peatakse kontserte" ("a hall where concerts take place"). In each of these instances, the former conforms better to the more "direct" style typical of news writing. While the compound form is more common overall in Estonian, a phrasal variant may be used occasionally for stylistic reasons. And in general, one would anticipate this sort of variation more in literary texts. It is important to note, however, that a compound is often the only practical means of expression in Estonian, so that this is not a question of choice to the extent that it is in English.

[13] While the allative case (-le) may be used in certain cases where "for" would be used in English, some translations of English "for" are not possible in Estonian, and in these cases compounding is used. For instance, where we could say either "a machine for making bread" or "a bread-making machine" in English, only the latter is possible in Estonian.

In terms of nominative vs. genitive tokens, the ratio between genres is consistent. Both genres favor the genitive compound to virtually the same degree.

Frequency category	NEWS TEXTS	LITERATURE
NC TYPES	162	162
GC TYPES	554	554
GC=NC TYPES*	314	314
NC TOKENS	1480 (60.4%)	970 (39.6%)
GC TOKENS	5543 (63.7%)	3165 (36.3%)
GC=NC TOKENS	4127 (62.8%)	2441 (37.2%)

Table 4.6: Breakdown of most frequent compounds by text genre

Remarkably, the compounds obtained from the Tartu Frequency List did not show variation. As this was unexpected, the possibility that the W1 forms had been normalized during creation of the corpus was discussed with Kaalep (p.c., July 2003), who confirmed that they had not. That is, variant forms were not modified to a "standard" form. This does not preclude the possibility, however, that the source documents (i.e., pre-corpus) followed a particular standard which dictated the use of one form over the other. That is, it is not unusual for newspapers, for example, to follow a set of prescribed editing guidelines. Consequently, it was decided to take the data set and do an Internet search for all compounds containing a W1 with different nominative and genitive forms. Using the Google search engine (www.google.com), two searches were done for each compound: once as an NC and again as a GC, with W2 in the nominative singular. This produced more realistic results than those obtained from the original data set. Table 4.7 shows a sample (25 of 92) of the results. This subset was selected to include the most variable types, as well as to represent the range of alternation patterns among the varying forms. The variation figures are calculated as the percentage of GCs vs. NCs:

Noun Compounds	English Gloss	W1 Case*	Variation (% GCs)
aed+vili	garden vegetable	nom	4.1
alkoholi+joove	alcohol drink	gen	94.4
ameti+mees	businessman	gen	83.5
ameti+post	business position	gen	82.4
arvamus+avaldus	declaration of one's opinion	nom	17.9
diplomaadi+kohver	attaché case	gen	78.7
hommiku+söök	breakfast	gen	99.6
kauba+hall	supermarket	gen	66.5
kontserdi+saal	concert hall	gen	94.2
kuuli+pilduja	machine gun	gen	96.1
liidu+vabariik	federation	gen	5.5
linna+osa	city district	gen	99.5
looma+aed	zoo	gen	98.7
märtsi+kuu	March	gen	98.4
mere+mees	sailor	gen	95.1
mõtte+käik	train of thought	gen	97.2
paberi+leht	sheet of paper	gen	95.4
peo+laud	party table	gen	59.3
piir+joon	border line	nom	14.9
piiri+valve	border security	gen	99.4
seisu+kord	state, position	gen	98.7
telefoni+number	telephone number	gen	90.6
toidu+aine	foodstuff	gen	78.7
vabaduse+kaotus	loss of freedom	gen	95.5
vere+soon	blood vessel	gen	73.1

Table 4.7: Sample results from Google search of compounds in Tartu Corpus

*The form listed here corresponds to the form found in the corpus.

Of the 1,094 compound types, only 716 had potentially alternating nominative/genitive forms (see Appendix A). Among those that could, 92 types showed variation in the data generated by the Internet searches. A number of these, however, showed very skewed proportions, which may suggest performance errors (i.e., typos or misspellings). Still,

some compounds showed substantial variation. Table 4.8 below outlines the characteristics of these alternations.

NC	GC	Alternation pattern
pidu+laud	peo+laud	consonant deletion, vowel lowering
kaup+hall	kauba+hall	consonant gradation + [a]
diplomaat+kohver	diplomaadi+kohver	consonant gradation + [i]
toit+aine	toidu+aine	consonant gradation + [u]
veri+soon	vere+soon	[i] → [e]
amet+post	ameti+post	ø → [i]
arvamus+avaldus	arvamuse+avaldus	ø → [e]

Table 4.8: Most common patterns among variant types

These types can be divided into three general groups – one defined by syllable deletion, (e.g., *pidu/peo)*, the second by vowel lowering (e.g., *veri/vere*), and the third by syllable insertion (*arvamus/arvamuse*).

In general, Tauli's (1973) account suggests more variation than the present study reveals. However, some of the patterns Tauli (1973) mentions, in terms of the genitive-nominative dichotomy, were confirmed by the data. For example, foreign borrowings often vary between the two forms. In such cases, Tauli (1973) claims that the first component tends to be assigned nominative case, especially when its final syllable is heavy (i.e., CVC, CVV(C), or CVV:(C)). Nevertheless, the data reveal that this type may actually favor the genitive.

Of the 92 varying types, 23 (25%) were NCs and 69 (75%) were GCs. Furthermore, the majority of these showed a strong bias for one or the other form. For example, only 22 compounds fell within the 4% and 96% range for either NC or GC. The distribution in this midrange set was 7 with an NC bias and 15 with a GC bias. In

terms of stem alternations, the same patterns appear in the midrange group and the polarized group (i.e., where the bias is <4% or >96%). The one lexical pattern that emerged in this distribution was that for 5 of the midrange compounds, W1 was a foreign borrowing, with all 5 compounds favoring the GC form. There were an additional 10 compounds with foreign W1's among the remaining 70 compounds; all but one (*kommerts+direktor,* "commerce director") showed a GC bias.

NC	**GC**	**GLOSS**	**% GC**
alcohol+joove	alkoholi+joove	alcoholic drink	94%
diplomaat+kohver	diplomaadi+kohver	attaché case	79%
kontsert+saal	kontserdi+saal	concert hall	74%
telefon+number	telefoni+number	telephone number	91%

Table 4.9: Examples of compounds in the corpus with borrowed W1's

In foreign borrowings which did not vary, the GC also prevailed. For example, while the Google search showed *telefonnumber* alternating with *telefoninumber*, the GC *telefoniraamat* (*telefoni+raamat*, telephone-gen.sg.+book, "telephone directory") did not vary with its corresponding nominative form.

As explained earlier in this chapter, since the Tartu Frequency List included only words found in both the news and literary texts, there may have been some high frequency words that occurred in only one genre. Thus, we also searched a few apparently common compounds which contain borrowed constituents and found again that there was variation, and that the genitive variant occurred more often: *veeb+sepp / veebi+sepp* ("webmaster") – 73% GC, *aatom+pomm / aatomi+pomm* ("atom bomb") – 55% GC.

To some extent, the tendency for foreign borrowings to trigger variation might be attributed to the relative "newness" of the compound. The two external examples mentioned above, at least, seem to support such a hypothesis.

Apart from the foreign borrowings, there appears to be little uniformity in terms of the classes of compounds that vary. Thus one can conclude that the variation is lexically motivated only to a limited extent, and that further considerations are necessary. But before exploring these possibilities, we will attempt to further constrain the set of compounds that are allowed to vary. Tauli (1973) offered a few semantic constraints and one phonotactic constraint. We begin by reconsidering the issue of phonotactics. First, was Tauli's analysis correct, and are there any other phonotactic constraints causing certain words to use the NC rather than the GC?

4.5 Phonological analysis

4.5.1 Phonotactic Constraints

There is no apparent constraint on vowel-vowel sequences since there are numerous cases of V-V sequences at a compound's internal boundary: "linna-osa" ("city district"), "looma-aed" ("zoo"), "jää-ääre" ("edge of the ice"), "korteri-uks" ("apartment door"). Not only are vowel sequences permitted, but it seems that virtually every permutation of the vowel inventory, is allowable in Estonian. In fact, there were 53 different cross-boundary vowel combinations found among the 1094 compounds examined in this study. These combinations are presented in Table 4.10 below. It should be noted that these combinations are not exhaustive for Estonian, as they represent only those found among the compounds in the Tartu Frequency List.

Word-Final| Word-Initial →

↓	a	e	i	o	u	ɤ	æ	ö	ü	aa	ae	ai	au	ei	öö	uu	ɤu
a	+	+	+	+			+		+		+		+		+	+	
e	+	+	+	+		+				+	+	+			+		
i	+	+	+	+	+	+			+	+	+		+				
o	+										+						
u	+		+	+		+			+		+	+		+			+
ɤ																	
æ																	
ö																	
ü																	
aa			+	+													
ao																	
ea	+			+													
ei													+				
eo													+				
oo						+											
uu				+													
ɤe				+													
öö	+								+	+	+						

TABLE 4.10: Cross-boundary vowel sequences in 1094 compounds

The data showed no occurrences of W1's with /ɤ/, /æ/, /ö/, or /ü/ in word-final position. This is in fact the case in Estonian generally for the short form of these vowels, as well as for the mid back vowel /o/, which nevertheless frequently occurs as the final vowel in foreign borrowings such as *auto* ("car") and *kino* ("cinema") (Hint, 1998:102). In addition, the mid front rounded short vowel /ö/, did not occur in the word-initial position of any W2 in the data. Again this reflects a tendency in the language in general. A search in an Estonian dictionary (Silvet, 1964: 485-486) for words beginning with the short monophthong /ö/ revealed only ten wordforms, seven of which included *ökonoom* ("economy") and six of its derived forms. By contrast, the long versions of these vowels, except for the mid back unrounded /ɤ/, are allowed and do occur: *öö* ([öö:],"night"), *jää* ([jææ:],"ice"), *soo* ([soo:],"swamp"), *süü* ([süü:],"blame").

Thus, taking into account the allowable word boundaries in the language, the concatenation of vowel-final and vowel-initial words appears to be unconstrained. High front unrounded vowels may be followed by mid back rounded vowels, low back unrounded with high front rounded, and so on. So even at the level of the feature, there is no perceptible constraint on such sequences. Neither height, nor backness, nor rounding appears to affect the phonotactics here.

While vowels may combine freely, consonants may be subject to some constraints. As mentioned earlier, Tauli (1973) contends that monosyllabic W1's ending in a consonant cluster should take the genitive rather than the nominative form, as in the GC *nartsunukk* vs. the NC **nartsnukk*. There appears to be some truth to this, though it is slightly more involved than Tauli suggests. First of all, he uses this phonological constraint to explain an exception to the semantic constraint he had posited (material+object). It is not clear whether he meant to generalize this rule, but if a generalization cannot be made, then his proposal would seem somewhat ad hoc. Thus, we will assume that Tauli did intend this as a generalization across the language, so that this constraint holds for all compounds with monosyllabic W1's, regardless of the semantic class. Then stated formally, Tauli's rule should look something like the following:

"Consonant Cluster Constraint" (CCC) 1

> *Where the second component of a compound is consonant-initial, monosyllabic first components ending in a consonant cluster must occur in the genitive form.*

One can certainly find evidence of this in the data: *mänguruum* (*mäng+ruum*, game-gen.,sg.+room, "playroom"), *kunstikriitik* (*kunst+kriitik*, art-gen.,sg.+critic, "art critic"), *pangarööv* (*pank+rööv*, bank-gen.sg.+robber, "bank robber"), *palgapäev* (*palk+ päev*, pay-gen.sg.+day, "payday"). On the other hand, there are several counterexamples among the data: *turmtuli* (*turm+tuli*, storm-nom.sg.+wind, "storm wind"), *vintpűss* (*vint+pűss*, shot-nom.sg.+gun, "shotgun"), *võrkpall* (*võrk+pall*", volley-nom.sg.+ball, "volleyball") , *rongkäik* (*rong+käik"*, train-nom.sg.+walk, "procession"), *metssiga* (*mets+siga*, wild-nom.sg.+pig, "wild boar"), and *käsk+kiri* (*käsk+kiri*, command-nom.sg.+letter, "decree"). Perhaps a little refinement of the rule would help. For example, if we look more closely at the consonant sequences and their phonological features, we can see that for the counterexamples the onset of the second component in each case is voiceless. But recalling that Estonian has no voiced stops, this observation does not actually reveal anything. There is another common feature, however. The onsets /t/, /p/, /k/, and /s/ are all obstruents. So we can say that an NC is allowed when the onset of W2 is an obstruent. The constraint can thus be reformulated as follows:

"CCC" 2

> *Where the second component of a compound begins with a sonorant, monosyllabic first components ending in a consonant cluster must occur in the genitive form.*

Stated in this way, the possibility of a GC being used when the onset of W2 is an obstruent is not ruled out, as in *kunstikriitik* and *palgapäev*, though this certainly makes the argument weaker. Another, more plausible explanation might be that the GC is used when it facilitates a shift in the place of articulation, i.e., when the value of the feature

"anterior" changes. For instance, without the intervening vowel, *kunstkriitik* would be more difficult to articulate because the coda of "kunst" is [+anterior] and the onset of "kriitik" is [–anterior]. This argument becomes more compelling when we look at some of the counterexamples to version 1 of the rule: *rongkäik* ("procession"), *metssiga* ("wild boar"), and *käskkiri* ("decree"). In each of these cases the coda and following onset have the same place values: *ro*[ŋ+k]*äik*, *met*[s+s]*iga*), and *käs*[k+k]*iri*. However, among Tauli's (1973) examples, the compound "palk+maja" (log house) contradicts all three hypotheses.

Finally, it seems that perhaps what Tauli (1973) meant to say was the following:

"CCC" 3

Where the nominative of W1 is a monosyllable ending in three consonants and the onset of W2 consists of one or more consonants, W1 should occur in the genitive form.

This constraint is supported by the data:

a. narts[u]nukk (Tauli 1973)
b. kunst[i]muuseum
c. arst[i]teaduskond
d. kunst[i]kriitik
e. märts[i]kuu

In addition to breaking up the consonant sequence, the intervening vowel in these cases separates two heavy syllables. In all five examples, the CCC thus serves to avoid stress clash. Moreover, in *a* and *e*, the vowel preserves the transparency of morpheme boundaries since /sn/ and /sk/ are both possible onsets in Estonian.

Though the CCC addresses a relatively small group of compounds, it explains the behavior of this subset with relative consistency. The Google search actually showed some variation between märtskuu and märtsikuu (see Table 4.7), but the proportion that were NCs was minimal – just 16 of 1,026 or 1.6% of the tokens retrieved. It is not unreasonable to assume that the NCs in this case may have resulted from performance errors.

Despite the fact that some of the other observations proved less reliable, it may still be possible to make some generalizations. For example, it is clear that the place of articulation hypothesis does not hold across the board; however, it nevertheless seems plausible to posit that it represents a salient articulatory preference, particularly in the choice of the NC in cases where like segments are concatenated, as in “metssiga”. So phonotactics not only constrains the set of compounds that may be either NC or GC, but it also motivates a preference for one over the other. Given a choice, it the option requiring the least effort invariably wins.

4.5.2 The role of prosody

In addition to phonotactic effects, the notion of foot isochrony in Estonian may be at work here. Perhaps free alternation simply occurs when the two variants are isochronous as in “peolaud” vs. “pidulaud” (41%-NC / 59%-GC) or “diplomaatkohver” vs. “diplomaadikohver” (21%-NC / 79%-GC). According to Gordon (1997) two types of isochrony are present in Estonian - foot isochrony and word isochrony. Unsurprisingly, the foot and the word in Estonian often coincide. For the present analysis at least, both seem to hold true.

The process of gradation reduces the disyllable /pi.tu/ in “pidulaud” to a monosyllable /peo/ (/eo/ is a diphthong). To compensate for this loss, /peo/ becomes /peo:/, a Q3 monosyllable, thereby meeting the minimal word requirement. In the case of “diplomaatkohver” and “diplomaadikohver”, the final Q3 syllable of /ti.plo.**maa:t**/ is reduced to Q2 in the genitive /ti.plo.**maa**.ti/ and forms a disyllabic foot with /ti/, thus filling the same metrical space.

While these examples may explain why some variants are ok, they do not explain why the genitive is preferred. The following analysis, however, examines two possible motivations for this preference.

One possibility is that the extra syllable which the genitive marker often creates would help preserve the normal stress pattern, since Estonian tends to avoid adjacent stressed syllables. Thus, /pii:r.val.ve/, which aligns two heavy syllables, may alternate with /pii.ri.val.ve/, restoring the more natural HL(+HL) trochaic rhythm typical in Estonian. Likewise, the binary trochees of /al.ko.ho.li.joo.ve/ are preferred over the

ternary HLL(+HL) pattern of /al.ko.hol.joo.ve/. Performance preferences of this nature would seem to make Hayes's (1995) proposal to treat Q3's as disyllabic less compelling. That is to say, if Q3 syllables are actually covertly disyllabic, why would a native speaker be inclined to convert them to overt disyllables? There are, at a minimum, three possible answers to this:

1. Q3 syllables are underlyingly monosyllabic and may cause stress clash; therefore, native speakers avoid stress clash by using a GC.
2. Q3 syllables are underlyingly disyllabic and do not cause stress clash; therefore, the GC variant cannot be motivated by stress preferences.
3. Q3 syllables may be underlyingly disyllabic, but speakers prefer the more transparent disyllabic surface form, i.e., the GC.

Stress may or may not be the only factor involved in choosing one form over another, but it is reasonable to suggest it plays a role here. And assuming that Hayes (1995) is correct in his characterization of Q3 as a disyllable, answer three seems to be the most promising option. As a preference for transparency emerges as a motivating factor in the phonological systems of many languages (Hume and Johnson, 2001; Clark 1993), this presents a plausible explanation for the choice of the genitive over the nominative variant.

One further, but not remote, possibility is that the genitive marker, which is fairly consistent in form, creates a natural word boundary, thus making the compound easier to parse and preserving its semantic transparency. For example, in fast speech a listener might perceive *toitaine* as "toit" + "taine" (food + lean_{adj}), though they would have to be part of different collocations, instead of "toit"+ "aine" (food + $\text{stuff}_{\text{noun}}$). On the other

hand, the GC *toiduaine*, “toidu” + “aine”, presents no such ambiguity. Needless to say, lexical ambiguity would not arise in every case. However, even without this problem, it seems logical that parsing compounds should be facilitated by the presence of a recognizable morphological boundary.

4.6 Conclusions:

In many cases, Estonian compounds are assigned either the NC or GC form based on the semantic relation between components. Or there may be a phonotactic constraint such as the Consonant Cluster Constraint discussed above. While the extent to which free variation between NCs and GCs actually occurs is somewhat less than expected, there clearly are some cases where there is substantial variation between the two forms. Based on the data, and in the absence of semantic or pervasive phonotactic effects, it appears that Estonian compounds may vary as long as the prosodic requirements are met. It is true as Tauli (1973) claimed that foreign borrowings tend to vary, as do nouns ending in “-us”. However, the preference in these cases, contrary to Tauli, is the genitive form. Moreover, such lexical motivations are secondary.

Variation is allowed when the prosodic integrity of the compound is not jeopardized. For Estonian this means preserving its quantity. Thus, in an alternation like piirvalve <-> piirivalve, “piir”, which is Q3, could alternate with a Q2 + Q1 bisyllable “piiri” since the extra weight of [pii:r] compensates for the loss of the genitive vowel.

Among the compounds that vary, the evidence shows that the genitive form is usually preferred to the nominative. The preference for the genitive appears to be driven by Estonian's strong bias for trochaic stress and its resistance to stress clash. In more general terms, the preference can be attributed to common phonological principles: "ease of pronunciation" and "ease of perception". The genitive is the most consistent form in the sense that one can rely on its ending in a vowel; consequently, it may be more easily recognized than the nominative when concatenated to another string. In other words, the genitive marker may form a more salient boundary than the nominative. Furthermore, a vowel intervening between two consonants may be for many languages, and certainly is in most cases for this one, an articulatory facilitator.

Chapter Five

ENGLISH COMPOUNDS

The English compounds examined in the present study were of the noun+noun variety, as was the case in our treatment of Estonian compounds. However, English compounds, as discussed previously in this work, exhibit a different type of variation pattern than do Estonian compounds. That is, they may occur in three orthographic forms: closed, hyphenated, and open. Recall also that the definition of compounding followed here is more inclusive than exclusive, generally following Bauer (1983). Thus, noun+noun combinations were not excluded if they lacked compound stress or if they were semantically opaque. The intent was to capture all possible compounds that occurred in the corpus. We begin this portion of the study with a description of the English corpus.

5.1 Compiling the Corpus:

The English corpus for this study came from a variety of sources. The criteria for choosing a corpus was that it be as current, balanced, and unedited a sampling of standard American English as possible. To accomplish this, we initially endeavored to compile an original corpus. Creating one's own corpus has the advantage of customized content; it also allows for incorporation of more current material. However, creating a corpus of

substantial size is a long-term project in and of itself. Thus, the original corpus was ultimately combined with two corpora compiled by the Linguistic Data Consortium (LDC) of the University of Pennsylvania (www.ldc.upenn.edu).

First, we combined the original corpus of 1.42 million words (henceforth, the "Sepp Corpus") with texts from the Department of Energy consisting of an additional 1.06 million words. This amalgamation was called "Corpus A". After an initial analysis, it was determined that more data were needed, and an additional 11.5 million words from the LDC's American National Corpus (ANC) were added. The latter will be referred to alternately as "Corpus B" or the ANC.

5.1.1 Corpus A

Corpus A contains a total of 2,490,758 words. The first major component, the Sepp Corpus, contains 1,428,342 words and was compiled from the following sources:

SOURCE	**SIZE** (in number of words)
Letters to the Editor (from a wide range of American newspapers including the *Anchorage Daily News, the Village Voice, the Los Angeles Daily News, the Honolulu Advertiser, Newsday,* and *Computer World*	46,983
Speeches and Essays (Bill Clinton; George W. Bush; Bill Gates; Noam Chomsky; Kurt Vonnegut, novelist; William Buckley, journalist/political commentator; Mike Bloomberg, NYC mayor; Louis Gerstner Jr., former IBM chairman/CEO; Ralph Nader, activist/lawyer)	713,749
U.S. Federal Trade Commission (press releases)	577,427
Internet newsgroups	60,063
Newsweek (Jan.28, 2002)	30,120
Total	**1,428,342**

Table 5.1: Contents of the Sepp Corpus

All of the above were obtained online. Some of the text files (e.g., most of the speeches) were extracted by means of a webcrawler. This made extraction a bit faster, as a list of relevant files could be extracted all at once, rather than having to copy and paste individual web pages. The crawler did not work well on the newsgroups, however. That is, it retrieved too many extraneous files; therefore, extracting these texts one page at a time was found to be more efficient.

The choice of texts for this corpus was intended to reflect, to the extent possible, nonstandardized or "unprescribed" spellings of compounds. For example, "letters to the editor" were chosen with this goal in mind. The letters are written by readers, generally responding to a particular news piece. Unfortunately, it was found that many newspapers, though not all, do edit the letters they print for grammar, which usually includes spell-checking. The length of the letters is also restricted in most cases, so compilation may be time-consuming. Still, they provide a wider range of content than some other genres might. Internet newsgroups, on the other hand, are not subject to editing and seem to reflect spontaneous language use. The one problem with the newsgroups is that the language is sometimes a little too "real", as it is more likely to be substandard, and spelling errors are not uncommon. As for the speeches and essays, since they come from different sources, it is safe to assume that these authors were not edited by the same individual. Whether the forms used reflect the authors' personal spelling habits or not is of no consequence in this study. In short, an ideal genre may be hard to find, but the combination of texts used here should at least offer ample opportunity for variation.

The other major component of Corpus A is the U.S. Department of Energy Corpus (1991), part of the Association of Computational Linguistics "Data Initiative Project" and distributed by the LDC. This is a very large corpus, but due to its technical nature, only a subset (files 1-7) was used, which totaled 1,062,416 words.

5.1.2 Corpus B

After a preliminary analysis of the data, it was determined that a larger corpus was needed. The ANC First Release seemed an appropriate choice. The American National Corpus, modeled after the British National Corpus (BNC), will eventually total 100 million words. In the meantime, the LDC released a subset consisting of 11.5 million words. The contents of the first release were compiled from the following sources:

SOURCE	**SIZE** (in number of words)
New York Times	3,207,272
Switchboard corpus (spoken language transcript)	3,056,062
Berlitz Travel Guides (travel industry texts)	514,021
OUP non-fiction (Oxford University Press)	224,037
Charlotte Narrative (spoken language transcript)	117,832
Call home (spoken language transcript)	50,494
Slate Magazine (msn.com)	4,338,498
TOTAL	**11,508,216**

Table 5.2: Contents of Corpus B

Corpus B, like corpus A, reflects a variety of language styles and covers a broad range of topics. In addition to the apparent added advantage of length, corpus B contains spoken language transcripts. The speech represented by the transcripts is quite varied in

terms of topic and register. Nevertheless, since the speakers in these transcripts did not actually write them, there is no measurable benefit here in terms of orthographic variation. Instead, the orthography is likely subject to the bias of individual transcribers. Where spoken transcripts do enhance the corpus, however, is in validating the broad usage of a given compound.

Finally, corpus A and corpus B combined brought the total corpus size to 13,998,974 words.

5.2 Procedure

5.2.1 Extracting the compounds

The Sepp Corpus and the Department of Energy Corpus were part-of-speech tagged by means of the Brill Tagger (Brill 1992, 1994). Compounds were extracted using Perl programs, and then the compound lists obtained from the two subcorpora were combined. The ANC corpus was already part-of-speech tagged.

The output files from Corpus A and the ANC were edited to obtain only noun+ noun compounds. This task was done manually. As the part of speech tags were not always accurate, the data had to be checked for errors. Strings in which one or both components were not nouns were eliminated.

The method for extracting the closed (*healthcare*), open (*health care*), and hyphenated (*health-care*) compounds was of course different, each presenting its own

particular challenges. Obviously, the extraction programs had to be customized to search for the appropriate form of the compound. Closed compounds are the most difficult form to extract since part-of-speech taggers for English do not parse them into their component parts. For open and hyphenated compounds, there are problems of ambiguity.

The extraction program for closed forms searched a string from the left edge and if it found a match in the dictionary (i.e., the Wordnet lexical database, Miller,G. Beckwith, Fellbaum, Gross, and K.J. Miller, 1990), the remaining right string was searched. If two adjacent strings were found in the dictionary, they were extracted as compounds. Closed compounds, while not ambiguous in the same way as open and hyphenated forms, may show internal ambiguity. For example, looking purely at the orthography, the word "internally" can be parsed into two nouns, i.e., "intern" and "ally"; though interns may well have allies, this is clearly an erroneous parse. Misparsing of this sort occurred in many instances, as illustrated in Table 5.3.

word-final string	…ally e.g., direction ally emotion ally lyric ally magic ally	…ion e.g., direct ion contract ion elect ion exhibit ion	…ted e.g., car ted demo ted marina ted weigh ted	…ton e.g., can ton car ton hunting ton simple ton
Number of misparsed types	219	241	83	124
Number of misparsed tokens	1327	51624	351	8281
percentage of raw output (413,700 tokens)	.3%	12.4%	.08%	20%

Table 5.3: Percentage of closed compound output misparsed for four common word endings

As can be seen from Table 5.3 above, the misparsing of "ion" as the second component of a closed compound in itself constitutes a considerable portion of the output of the extraction procedure. While the proportions for "ally" and "ted" shown above may seem small, the accumulation of parsing errors triggered by these and other lexically ambiguous strings make filtering essential. Thus, after a visual scan of the raw output list, the extraction program was modified to skip the following strings as potential W2's: *ability, ally, als, ate, ded, ern, ert, ese, est, ion, iou, ism, led, ler, let, ley, ling, less, ness, out, ping, sumer, ted, ting, tic, ton,* and *ute*. The likelihood that these strings were actually occurring as closed compound constituents is minimal. Moreover, many of these are suffixes and highly productive, so this type of filtering was effective in minimizing the manual post-extraction task. At the same time, this program modification of course did not completely obviate the problem. For example, a suffix like "-ship" (e.g., friendship, statesmanship, scholarship), which is quite productive as well, could not be filtered out since "ship" is also a noun that could easily occur in a closed compound: "war+ship", "space+ship", "steam+ship", etc. To deal with this and similar cases, the files were sorted by the second potential constituent ("W2"), and suffixed forms were removed manually.

In extracting open and hyphenated compounds, programs were designed to retrieve the maximum number of tokens, so all forms which fit the general category "noun" were extracted, including proper nouns, common nouns, pronouns, gerunds, and acronyms. In addition, noun+noun sequences were targeted in the context of trigrams, i.e., noun+noun+anything (including punctuation). The purpose of the trigrams was to disambiguate sequences, verifying that the target bigrams were actually compounds, and

specifically noun+noun compounds. It is not uncommon for a lexical item to be mistagged in these large corpora, as in the following string from the ANC:

nn>**host** nn>**tour** nns>**groups**

The tags "nn" and "nns" refer to singular and plural nouns, respectively. In this case, the target bigram "host tour" is tagged as a noun+noun sequence. While the possible sense of "host tour" is not evident, it is nothing that a little creative thinking could not manage. The point is that whether or not it is a compound, it certainly could be. But within the context of the corpus it is not. The trigram unwraps this conundrum, revealing that "host" in this context is probably used as a verb.

Sometimes the trigram is not even necessary for disambiguation. For example, while "African" and "Italian" are both noun and adjective, in the case of NPs such as "African-American" or "Italian-American" they are adjectives. The constituent structure is $[\text{adj n}]_n$ and not $[\text{n n}]_n$. This is evidenced in parallel constructions like "Swedish-American" or "Polish-American" (compare *Swede-American and *Pole-American).

Proper nouns were extracted because this produced a substantial number of hits, many of which were legitimate compounds, e.g., "**Christmas** tree", or "**Police Commissioner**" (i.e., in a title such as *Police Commissioner Smith*). In some instances, common nouns were simply mistagged as proper nouns. Of course, there were also a significant number of useless hits, which had to be edited out: "Monica Lewinsky", "President Clinton", etc.

Ultimately, 91,868 compound types were extracted from corpus A and B taken together. Of these, 5,118 were closed, 9,804 were hyphenated, and 79,955 were open. Since some compounds varied in orthographic form (e.g., "database" and "data base" or "website", "web-site", and "web site"), the number of total types is less than the sum of the closed, hyphenated, and open forms. The token count for the corpus was 265,991: 73,001 closed, 19,953 hyphenated, and 173,037 open.

For the purpose of analysis, we have focused on the free variants (compounds that occurred in more than one orthographic form) and the most frequent types (frequency ≥ 35) in the corpus. This latter subset consists of 707 compound types, many of which varied in form (364 closed, 231 hyphenated, and 473 open), and will henceforth be referred to as the 35+ group. Prior to data analysis, the free variants were grouped into four lists: those that occurred with variation between closed/open, closed/hyphenated, open/hyphenated, and closed/open/hyphenated (see Appendix B, Table 3). These lists were examined for distinguishing features that might serve as good predictors of a compound's orthographic form.

5.2.2 The distributions of compound forms

As the goal of this work is to account for differences in the distributions of compound forms, the first step was to obtain these distributions. Perl programs were used to compute the frequencies of closed, hyphenated, and open compound forms for each W1+W2 compound type identified in the corpus (where W1 and W2 are the first

and second lexical components of the compound, respectively), and these values were then converted into relative frequencies. An illustrative sample of the results is presented in Table 5.4.

W1	W2	closed	hyphen	open	total	P(closed)	P(hyphen)	P(open)
guide	line	764	0	0	764	1.00	.00	.00
coca	cola	0	103	0	103	.00	1.00	.00
school	system	0	0	256	256	.00	.00	1.00
front	runner	7	58	0	65	.11	.89	.00
work	place	211	0	45	256	.82	.00	.18
credit	card	0	42	644	686	.00	.06	.94
health	care	70	152	517	739	.09	.21	.70

Table 5.4: Relative frequencies for three orthographic forms in compounds

The examples in Table 5.4 represent the seven logical distribution patterns for English compounds. Among the 707 most frequent compounds in the corpus, 318 (45%) varied in orthographic form: 43 compounds occurred in all three forms, 101 occurred as closed or open, 161 as hyphenated or open, and 13 as either closed or hyphenated. Of the 389 (55%) types that did not vary, 207 were closed and 168 were open, while just 14 compound types in this group occurred in the hyphenated form.

The frequencies of the three orthographic forms for each of the compounds in the 35+ group are listed in Table 1 of Appendix B, which displays types in descending order of their total frequency of occurrence. The correlations among the relative frequencies for this subset of the corpus are shown below in Table 5.5. The negative correlations of course reflect the fact that the three relative frequencies must sum to 1.00. What is striking about these results is that the occurrence of a compound as open is strongly predictive of its non-occurrence as closed.

Bivariate Correlations
N = 707

Criterion variables	P(closed)	P(hyphen)	P(open)
P(closed)	1	-.275**	-.913**
P(hyphen)	-.275**	1	-.142**
P(open)	-.913**	-.142**	1

Table 5.5: Correlations of relative frequencies
** significant at the .01 level

5.3 Phonological features: Results and observations

A primary goal of this study is to reveal the possible phonological motivations underlying compound formation. The relationships between the phonological features and the orthographic forms of the compound are presented below.

5.3.1 Syllable Counts

An analysis of the syllable counts strongly suggests that there is a limit on the number of syllables allowed in a closed compound. By extension, this could be looked at as a limit on the number of metrical feet normally allowed within an orthographic word. In this case it seems that closed forms may be limited to a maximum of two feet. So any string that exceeds this limit must be separated by a space or a hyphen. Figures 5.1 and 5.2 highlight the rather polarized distribution of compound types and tokens with respect to syllables counts.

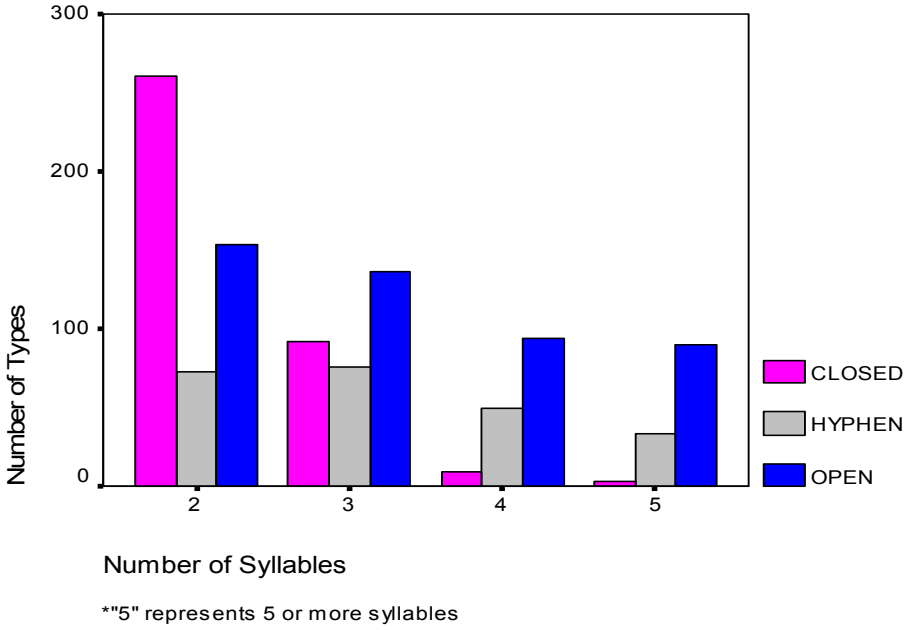


Figure 5.1: Number of Syllables for Three Forms (Compound Types)

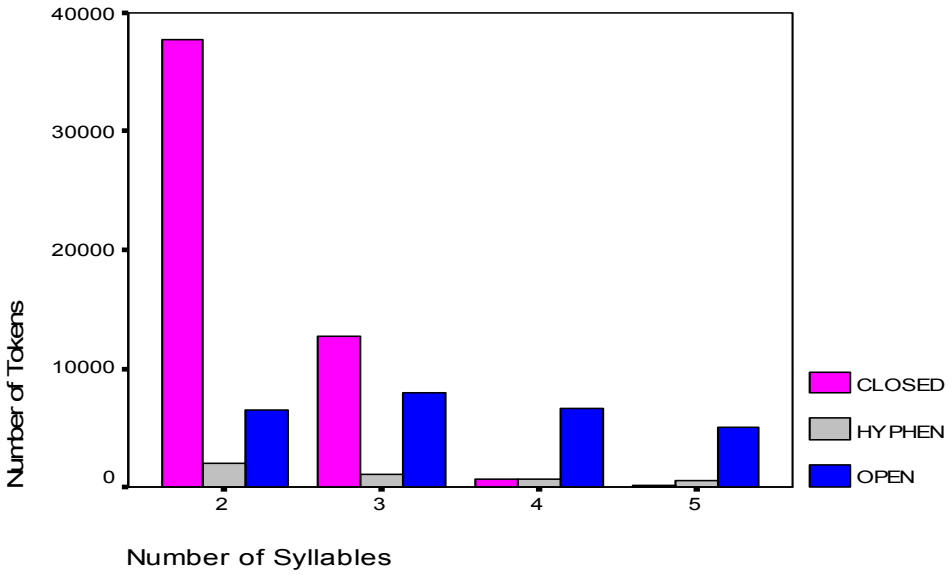


Figure 5.2: Number of Syllables for Three Forms (Compound Tokens)

The data reveal that 364 (51%) of the compounds in the 35+ list occurred in the closed form at least once. Of these 364 compounds, 260 are bisyllabic, 92 are trisyllabic, and 9 are tetrasyllabic. The remaining 3 contain five syllables and are enumerated in Table 5.6.

COMPOUND			FREQUENCY		
W1	W2	# of syll.	CLOSED	HYPHEN	OPEN
decision	**making**	**5**	**13**	**25**	**10**
policy	**maker**	**5**	**25**	**24**	**17**
suicide	**bomber**	**5**	**1**	**0**	**34**

Table 5.6: Closed compounds in 35+ list containing 5 syllables

Among those that contain five syllables, "suicidebomber" occurred only once in the closed form as opposed to its 34 open occurrences. This might suggest that the closed instance is an orthographic error. As for the other 2 five-syllable forms - "policymaker" and "decisionmaking", one could posit that the words "maker" and "making" are suffix–like. That is to say, these words might be destined to "suffixdom", as "hood" and "dom" before them. Bauer (1983: 36) makes a similar argument for "man" when used as a compound's second constituent. If "maker" and "making" were treated as suffixes, they should not be subject to compounding constraints.

A more plausible explanation might be related to the compound class. What is most obvious in this group is that all three of the five-syllable forms are synthetic or "verbal" compounds. That is, they are derived by adding the agentive "-er" or the gerundive "-ing" to verbs (Fabb, 1998: 68; Lieber 1983). Thus, it might follow that the derivational character of these compounds could exempt them once again from compounding constraints, allowing them to behave as derivatives with no definable

syllabic limits. Furthermore, since agentive and gerundive suffixes are both unstressed and neutral in terms of stress, the addition of one of these syllables to a stem only minimally alters the prosodic word.

The point to be made here is that there is strong evidence for a syllable limit in the choice between closed and open or hyphenated compounds, since closed forms seem to be constrained with respect to the number of syllables, and there appears to be no set limit on the number of syllables possible in nonconcatenated compounds. The following open and hyphenated forms found in the corpus attest to this:

7 syllables -- "cabinet secretary"
8 syllables -- "judiciary committee"
9 syllables -- "communications technology"
10 syllables -- "telecommunications analyst"
11 syllables -- "telecommunications infrastructure"
12 syllables -- "radiohalogenating biomolecule"
12 syllables -- "gadolinium-diethylenetriamine"

While it has been well-established that high frequency words tend to be short (Zipf 1935), and one might therefore suppose that what appears to be due to a syllable limit is simply a by-product of frequency effects, further analysis of the high frequency open compounds in this study indicated that the number of syllables in a compound does effect a bias in its form. Among the 35+ group there were 89 compounds which had 5 or more syllables and occurred in the open form but never in the closed form. Of these, there were 10 monosyllabic W1's and just 5 monosyllabic W2's; i.e., approximately 8%

of the 178 constituents were monosyllabic. By contrast, there were 40 constituents (22%) of four or five syllables within this subset.

5.3.2 Vowel sequences:

This variable was assessed in order to determine whether sequences of vowels were resistant to concatenation. Let us begin by explaining that we are referring here to "vowels" in both the orthographic and the phonological sense. Thus, a word such as "snow" would be counted as vowel final, and so would "state". Among the most frequent compound types, only 18 actually had a *vowel final-vowel initial* sequence (see Table 5.7). Six of the 18 occurred in the hyphenated form and all but one occurred in the open form. However, only three occurred in the concatenated form: "homeowner", "firearm", and "pineapple". Looking at these three counterexamples, one could posit that perhaps the constraint in fact is only phonological (since these three are phonologically #_C-V_#). However, the sample size for true V-V sequences is too small to substantiate such a claim. This is the case not just for this corpus, but in general, since vowel-final words are relatively few in number. In short, the presence of a vowel sequence across the internal boundary may factor into the equation, but on its own this variable is not very informative.

W1	W2	Closed	Hyphen	Open	Total
Orthographic Vowel sequence					
fire	arm	111	0	5	116
home	owner	67	0	12	79
life	insurance	0	2	40	42
line	item	0	42	12	54
love	affair	0	1	43	44
pine	apple	40	0	0	40
police	officer	0	0	125	125
price	increase	0	0	64	64
side	effect	0	7	80	87
state	income	0	0	67	67
state	university	0	0	41	41
trade	agreement	0	0	41	41
Phonological Vowel sequence					
academy	award	0	0	36	36
fbi	agent	0	0	54	54
law	enforcement	0	29	314	343
privacy	issue	0	0	40	40
security	adviser	0	0	37	37
tobacco	industry	0	9	72	81

Table 5.7: 35+ compounds with orthographic or phonological vowel sequence across morpheme boundary

5.3.3 Consonant sequences

There is, in general, no problem concatenating a word that ends in a consonant and a word that begins with a consonant. The data show multiple instances of such sequences: “businessman”, “landmark”, “earthquake”, etc. There is one exception to this, however. If the sequence involves identical consonants, the compound resists concatenation. Following the same procedure as for vowel sequences, both orthographic and phonological double consonants were identified (see Table 5.8). Among the 35+

compounds, there were 15 cases of orthographic doubling and an additional 9 cases of phonological doubling only. Twenty-two compound types occurred in the open form at least once, while only four occurred as closed compounds: “filmmaker”, roommate”, “teammate”, and “newsstand”, the latter two occurring exclusively in the closed form.

W1	W2	closed	hyphen	open	total
Orthographic Doubling					
breakfast	table	0	0	67	67
consumer	report	0	0	40	40
cross	section	0	0	36	36
domain	name	0	1	37	38
film	maker	122	1	1	124
gas	station	0	1	37	38
impeachment	trial	0	0	54	54
merger	review	0	1	44	45
news	stand	58	0	0	58
news	story	0	0	78	78
phantom	menace	0	0	38	38
quantum	mechanics	0	0	36	36
room	mate	122	0	1	123
spin	network	0	0	55	55
team	mate	144	0	0	144
Phonological Doubling only					
defense	secretary	0	0	49	49
house	speaker	0	0	68	68
ice	storm	0	0	48	48
phone	number	0	0	80	80
space	station	0	0	39	39
telephone	number	0	0	37	37
trade	deficit	0	1	75	76
welfare	reform	0	15	102	117
welfare	roll	0	0	36	36

Table 5.8: 35+ compounds with same consonant or phoneme across morpheme boundary

Three of the four closed forms had double “m-m” sequences. Perhaps the letter “m” and or the phoneme /m/ are exceptionally “alignable”. Or perhaps there is a lexical

explanation, as “mate” and “maker” are rather generic terms, almost suffix-like (cf., discussion on p.90). The one anomaly here is “newsstand”. But as was the case with the concatenated vowel sequences, the consonant sequence in “new**ss**tand” is orthographic and not phonological: /nuz.stænd/.

An additional “double consonant” test was applied to all compounds in the corpus. The results are summarized in Table 5.9 below. The data show a consistent pattern in terms of types and tokens. More importantly, these results reinforce the data from the 35+ set, showing that closed compounds resist concatenation of the same consonant.

	closed	**proportion of all closed**	**hyphen**	**proportion of all hyphen**	**open**	**proportion of all open**	**total**	**proportion of total**
TYPES	64	.012	457	.046	3574	.044	3977	.041
TOKENS	772	.010	644	.032	7774	.044	9190	.034

Table 5.9: Frequency and proportions of compounds with the same consonant or phoneme at internal boundaries

5.3.4 Compound Stress

Stress judgments were made on the compounds that appeared 35 or more times in the entire corpus. Compound stress was judged by three members of the CUNY Linguistics Program (including the author), who are speakers of General American English. Where possible, the *Cambridge Pronouncing Dictionary* (Jones 2003) was used to further validate judgments. Many closed compounds were listed in the dictionary, but closed or hyphenated forms frequently were not. Each judge assigned a value of 1 to

compounds determined to have main stress on the left component (i.e., "compound stress" or "CS"), and 0 to all others. Stress was finally assigned according to a best-of-three judgment. Total agreement among the three human judges for either CS or not CS was approximately 66%. Agreement on closed forms was the strongest at 79%.

The results of the stress judgments showed a definite correlation between closed forms and compound stress, as nearly all closed forms in the 35+ list were judged to have stress on W1. There were 18 that were assigned phrasal stress, some of which could be errors in terms of form and/or stress judgment. For example, "suicide bomber", with its stress pattern judged to be *sùicide bómber*, discussed previously, occurred only once in the closed form. The results showed that open compounds frequently have compound stress as well, with 370 out of 473 displaying this pattern. On the other hand, hyphenated forms appear to have a somewhat weaker preference for compound stress, as 166 hyphenated forms bore compound stress and 65 (about 28%) did not.

5.3.5 Summary of Phonological Effects: Feature Models

How well do the phonological variables predict the proportions of closed, hyphenated, and open compounds? Table 5.10 defines the five phonological variables used in the analysis. Note that the new feature added here, BSYLL, is a binary variant of SYLL.

Criterion Variables	
PCLOSED	Proportion of occurrences of the compound in closed form
PHYPHEN	Proportion of occurrences of the compound in hyphenated form
POPEN	Proportion of occurrences of the compound in open form
Phonological Features	
SYLL	Number of syllables in the compound
BSYLL	Binary feature based on the syllable count of less than 5 (=0) or greater than or equal to 5 (=1)
VSEQ	Presence (=1) or absence (=0) of a vowel sequence across internal lexical boundary of the compound - e.g., "polic**e** **o**fficer", "academ**y** **a**ward", "pin**ea**pple"
DOUBLEC	Presence (=1) or absence (=0) of a double consonant across internal boundary of the compound – e.g., "domai**n** **n**ame", "i**c**e **s**torm"
STRESS	Presence (=1) or absence (=0) of compound stress (i.e., main stress on the left component) for the compound

Table 5.10: Description of criterion variables and phonological features used in regression analyses

Correlation coefficients were computed to determine the strength of association between each of the phonological features (VSEQ, DOUBLEC, SYLL, BSYLL, and STRESS) and the proportions of the three orthographic forms of the compounds in the corpus (Biber, Conrad, and Reppen, 1998). The results in Table 5.11 indicate that the number of syllables is the strongest single predictor of the closed and open forms. The binary syllable feature and compound stress also show moderate to large effects, and compound stress is significant at the .01 level for all three forms.

Phonological Features	**P(closed)**	**P(hyphen)**	**P(open)**
VSEQ	-.100**	-.001	.103**
DOUBLEC	-.122**	-.049	.146**
SYLL	-.589**	.080*	.573**
BSYLL	-.357**	-.002	.369**
STRESS	.343**	-.304**	-.224**

Table 5.11: Bivariate Correlations for Phonological Features

* significant at the .05 level

** significant at the .01 level

Stepwise multiple linear regression was then used to combine the phonological features to predict the proportion of each orthographic form of compound. Table 5.12 shows the summary of the best regression model for closed compounds. The features SYLL, STRESS, BSYLL, and DOUBLEC (ordered by strongest to weakest predictor) accounted for 39.5% of the variance. An ANOVA showed the regression to be statistically significant ($F(4,702) = 116.151$, $p<.001$). (See Appendix B-2, Table 12a and b for the ANOVA table and the regression coefficients.). Note that the feature VSEQ is not included in the regression model as it failed to make a significant contribution.

Model Summary				
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**
	.631	.398	.395	.37579
Predictors: (Constant), SYLL, STRESS, BSYLL, DOUBLEC				
Dependent Variable: PCLOSED				

Table 5.12: Phonological feature model summary for closed forms

Table 5.13 is the regression summary for the proportion of hyphenated forms. The phonological features of STRESS and BSYLL accounted for just 9.5% of the variance, but this was statistically significant ($F(2,704) = 38.153$, $p < .001$). (See Appendix B-2, Table 13a and b for the ANOVA table and the regression coefficients.) A comparison with Table 5.12 shows a much weaker relationship between phonological features and hyphenation than between phonological features and closed forms.

Model Summary				
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**
	.313	.098	.095	.18966
Predictors: (Constant), STRESS, BSYLL				
Dependent Variable: PHYPHEN				

Table 5.13: Phonological feature model summary for hyphenated forms

The regression model summarized in Table 5.14 accounted for 34.3% of the variance in the proportion of open forms by combining the features SYLL, DOUBLEC, and BSYLL. An ANOVA showed that this was significant ($F(3,703) = 123.921$, $p<.001$). (See Appendix B-2, Table 14a and b for the ANOVA table and the regression coefficients.)

Model Summary				
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**
	.588	.346	.343	.38028
Predictors: (Constant), SYLL, DOUBLEC, BSYLL				
Dependent Variable: POPEN				

Table 5.14: Phonological feature model summary for open forms

The regression analyses thus indicate that, with the exception of the feature VSEQ, the phonological features considered here do in fact have some predictive power, particularly with respect to closed and open forms. Hyphenation seems to be only weakly accounted for.

5.4 Lexical features: Results and observations

As previously stated, the primary goal of this research was to identify the phonological motivations in the distribution of compound forms. In addition, the possible lexical effects on this distribution were acknowledged and explored. Furthermore, there appears to be some evidence that a degree of interplay between phonological and lexical features exists. For example, some connection between stress placement and certain lexical constituents was observed in the data. Though this was only apparent for a handful of constituents, the phenomenon is suggestive of a more general pattern, which will be explored through statistical measures.

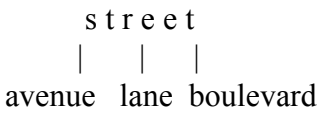
5.4.1 Illustrations of lexical effects on stress placement

In examining the stress patterns of compounds in the present study, it was observed that the word "end" seems to favor primary stress. In "END+user", "tail+END", and "rear+END**"** main stress is on "end" regardless of which constituent slot it fills. These compounds occur in either open or hyphenated forms, but are never concatenated. One counterexample to this pattern is the word "WEEKend", which has stress on "week". In this case, we have both the presence of compound stress and concatenation.

It was further observed from the data that the words "home" and "child" tend to resist main stress: child+STAR, child+PORN, child+SAFETY, home+RULE, home+FIELD, home+THEATER. The absence of compound stress in these examples is

accompanied by nonconcatenation. Counterexamples such as “HOME+owner” and “HOME+page” likewise show the opposite effect: the presence of compound stress and concatenation.

The issue of stress placement in these cases could perhaps be explained by what Burstein (1992) calls “broad reference nouns” (or BR nouns), as described in a previous chapter. Burstein characterizes words such as “street” and “school” as BR nouns, contrasting them with narrower terms such as “avenue”, “lane”, or “boulevard”, and “university” or “academy”. This could be illustrated in terms of a hierarchy, with the top level being the broadest and the lower level narrower:



The motivation for this distinction is that, as head nouns in a compound, “broad” and “narrow” reference nouns contrast in terms of stress – e.g., “WALL **Street**” vs. “Park **AVENUE**”. While Burstein (1992) focuses on BR head nouns, the idea that a semantically weak (i.e., broad) noun systematically avoids stress is compelling. In the “home” examples above we see that “home” avoids stress whether it is the head or nonhead. The word “child” follows the same pattern – compare GOD+child and child+STAR. Furthermore, if we plug in a narrow reference noun for “child” or “home”, the stress (arguably) shifts:

“TEEN+star”	“HOUSE+theater”
“BABY+safety”	“COLONY+rule”

Thus, it seems reasonable to suggest that words like “home” and “child” may be semantically weak forms compared to their adjoining nouns, and that as compound constituents, they avoid main stress. This explanation could possibly even account for the stress-initial counterexamples since “homeowner” pairs “home” with a semantically weaker, or at best, equally as weak constituent, since one could imagine numerous hyponyms for “owner”: restaurateur, partner, proprietress, lessor, landlord, etc. This scenario might equally apply to “weekend”.

Sports terms also tend to avoid compound stress: “center-FIELD”, “switch-HITTER”, “pinch-RUNNER”, and “home-PLATE”. Again we have the pattern of no compound stress and no concatenation of constituents.

In any case, if we accept that all of these examples are compounds, then we would have to acknowledge that neither compound stress nor concatenation is criterial in defining compounds. What is clear though, and what has been established, is that there is a strong link between compound stress and concatenation. So if stress can be lexically determined, one might posit that concatenation may also be lexically determined.

5.4.2 W1W2 analysis: Lexical Feature Models

In a second set of statistical tests, multiple regression analyses were computed, this time using the lexical features of compounds in the corpus as predictors and the proportions of closed, hyphenated, and open forms again as the criterion variables. This analysis used the general compounding “behavior” of W1 and W2 in the corpus to predict the

distribution of forms for W1+W2. For example, it made it possible to ask "If W1 appears in a large number of closed compound types, is it more likely that its W1+W2 form will be closed?", and also, "Is the total number of occurrences, the token frequency, of W1's closed forms a better predictor than its type frequency?"

Table 5.15 defines the variables used in this analysis:

Corpus Frequency	
W1CORPFR	Frequency of the first component (W1) in the corpus
W2CORPFR	Frequency of the second component (W2) in the corpus
TOTAL	Frequency of all forms of the W1+W2 compound in the corpus
Token Frequency in Compounds	
W1TOKFR	Token frequency (i.e., the total number of instances) of the first component in all compounds
W1CTOKFR	Token frequency (i.e., the total number of instances) of the first component in closed compounds – e.g., "web" occurs as W1 190 times in closed compounds
W1HTOKFR	Token frequency of the first component in hyphenated compounds
W1OTOKFR	Token frequency of the first component in open compounds
W2TOKFR	Token frequency of the second component in all compounds
W2CTOKFR	Token frequency of the second component in closed compounds – e.g., "site" occurs as W2 195 times in closed compounds
W2HTOKFR	Token frequency of the second component in hyphenated compounds
W2OTOKFR	Token frequency of the second component in open compounds
Type Frequency in Compounds	
W1TYPFR	Type frequency of a particular W1 in all compounds -- e.g., "web" occurs as W1 in 144 different compounds, across all forms (website, web-page, web search, etc.)
W1CTYPFR	Type frequency of a particular W1 among closed forms – e.g., "web" occurs as W1 in 17 different closed compounds (website, weblog, webmaster, etc.)
W1HTYPFR	Type frequency of a particular W1 among hyphenated forms
W1OTYPFR	Type frequency of a particular W1 among open forms
W2TYPFR	Type frequency of a particular W2 for all compounds
W2CTYPFR	Type frequency of a particular W2 among closed forms
W2HTYPFR	Type frequency of a particular W2 among hyphenated forms
W2OTYPFR	Type frequency of a particular W2 among open forms

Table 5.15: Description of lexical variables used in the regression analyses

For each of the frequency variables listed, the log frequency was also calculated. Log frequencies (e.g., W2CTYPLF, the logarithm of the number of closed compound types that W2 appears in) were included because studies suggest that subjective word frequency is at least roughly logarithmic (Balota, Pilotti, and Cortese 2001). In experiments by Balota et al., for example, it was shown that participants' estimates of how often they encounter specific words are highly correlated with the logarithm of the words' objective frequencies, as measured by corpus counts. This suggests that the perception of frequency is on a log scale, and that the log transforms of the frequency variables may be more reliable predictors of compound variant preferences.

As in the phonological analysis, correlation coefficients (see Table 5.16) were computed for lexical features to determine the strength of the relationship between a given feature and each criterion variable. The two strongest correlations for closed and open compounds were the log frequencies of the W2 in closed types and the W2 in closed tokens. The fact that closed forms showed a strong correlation with W2 type frequency was expected, since there clearly are W2's which favor closed forms – e.g., *man, house, way*, etc. Correlations in hyphenated forms were generally weaker than for open or closed forms.

Bivariate Correlations
N=707

	P(closed)	P(hyphen)	P(open)		P(closed)	P(hyphen)	P(open)
Corpus frequency				***Log (Corpus Frequency)***			
W1 CORPFR	-.158	-.024	.172	W1 CORPLF	-.259	-.015	.273
W2 CORPFR	.039	-.016	-.034	W2 CORPLF	-.062	-.181	.134
TOTAL	.208	-.033	-.200				
Token Frequency In Compounds				***Log (Token Frequency In Compounds)***			
W1 TOKFR	-.032	-.038	.049	W1 TOKLF	-.163	-.133	.224
W1 CTOKFR	.146	-.059	-.125	W1 CTOKLF	.388	-.140	-.340
W1 HTOKFR	-.100	.394	-.064	W1 HTOKLF	-.176	.125	.128
W1 OTOKFR	-.311	-.103	.364	W1 OTOKLF	-.387	-.173	.472
W2 TOKFR	.295	-.129	-.249	W2 TOKLF	.215	-.384	-.058
W2 CTOKFR	.399	-.106	-.366	W2 CTOKLF	**.657**	-.210	**-.587**
W2 HTOKFR	.019	-.010	-.016	W2 HTOKLF	.057	-.093	-.020
W2 OTOKFR	-.232	-.107	.285	W2 OTOKLF	-.204	-.304	.339
Type Frequency In Compounds				***Log (Type Frequency In Compounds)***			
W1 TYPFR	-.264	-.042	.289	W1 TYPLF	-.271	-.122	.331
W1 CTYPFR	.343	-.109	-.307	W1 CTYPLF	**.447**	-.132	**-.405**
W1 HTYPFR	-.109	.299	-.015	W1 HTYPLF	-.175	.098	.138
W1 OTYPFR	-.317	-.112	.374	W1 OTYPLF	-.346	-.176	.431
W2 TYPFR	.100	-.153	-.038	W2 TYPLF	.073	-.367	.081
W2 CTYPFR	.393	-.102	-.361	W2 CTYPLF	**.652**	-.193	**-.589**
W2 HTYPFR	.117	-.040	-.104	W2 HTYPLF	.095	-.121	-.047
W2 OTYPFR	-.116	-.136	.178	W2 OTYPLF	-.112	-.322	.253

Table 5.16: Correlations for lexical features
(minimum value for p<.01 is +/-.081, p<.05 is +/-.098)

The summary results of stepwise multiple regression analyses using these variables are shown in Tables 5.17-5.19. Additional tables of the ANOVAs and regression coefficients can be found in Appendix B-2.

As shown in Table 5.17, ten lexical features – W2CTOKLF, W1CTYPLF, W1OTOKLF, W2OTOKLF, W2CTYPLF, W1HTOKLF, W2TOKFR, W2HTOKFR,

TOTAL, and W2TYPLF - accounted for 66.5% of the variance for closed compounds. This was statistically significant (F(10,696) = 140.941 p<.001).

Model Summary				
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**
	.818	.669	.665	.27973
Predictors: (Constant), W2CTOKLF, W1CTYPLF, W1OTOKLF, W2OTOKLF, W2CTYPLF, W1HTOKLF, W2TOKFR, W2HTOKFR, TOTAL, W2TYPLF				
Dependent Variable: PCLOSED				

Table 5.17: Lexical feature model summary for closed forms

The results for hyphenated compounds are shown in Table 5.18. Here, the lexical features W1HTOKFR, W2TOKLF, W2HTOKLF, W2TYPFR, W2TYPLF, W1CTYPLF, W1HTYPLF, W1OTOKLF, and W2CORPFR accounted for 40.5% of the variance. An analysis of variation showed this to be significant, F(9,697) = 54.462 p<.001.

Model Summary				
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**
	.643	.413	.405	.15376
Predictors: (Constant), W1HTOKFR, W2TOKLF, W2HTOKLF, W2TYPFR, W2TYPLF, W1CTYPLF, W1HTYPLF, W1OTOKLF, W2CORPFR				
Dependent Variable: PHYPHEN				

Table 5.18: Lexical feature model summary for hyphenated forms

The results for open compounds are displayed in Table 5.19. The following features accounted for 67.7% of the variance in open forms: W2CTYPLF, W2OTOKLF, W1CTYPLF, W1OTOKLF, W1HTOKFR, W2CTOKLF, TOTAL, W2HTYPLF,

W2TOKLF, W2HTOKFR An analysis of variance also showed this to be significant, $F(10,696) = 149.151$, $p<.001$.

Model Summary				
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**
	.826	.682	.677	.26655
Predictors: (Constant), W2CTYPLF, W2OTOKLF, W1CTYPLF, W1OTOKLF, W1HTOKFR, W2CTOKLF, TOTAL, W2HTYPLF, W2TOKLF, W2HTOKFR				
Dependent Variable: POPEN				

Table 5.19: Lexical feature model summary for open forms

From the results of the regression analyses for lexical features, four generalizations can be made: (1) lexical features are better predictors of closed and open forms than of hyphenated forms, (2) log frequencies are better predictors overall, (3) both type and token frequencies are predictive, and (4) the second component of closed forms appears to play a significant role in the orthography of compounds.

5.5 Combining phonological and lexical features

Finally, stepwise regression analyses were computed using all of the features, i.e., both phonological and lexical.

5.5.1 Model Summaries: All features

Summaries of the models for closed, hyphenated and open forms, showing the combined effects of all features on determining the orthographic form of compounds, are provided in Tables 5.20, 5.21, and 5.22 below. ANOVAs and regression coefficients are in Appendix B-2.

Model Summary				
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**
	.826	.682	.677	.27448
Predictors: (Constant), W2CTOKLF, W1CTYPLF, W1OTOKLF, W2OTOKLF, STRESS, W2CTYPLF, W2HTOKFR, SYLL, BSYLL, W1HTOKLF, TOTAL				
Dependent Variable: PCLOSED				

Table 5.20: All-feature model summary for closed forms

Model Summary				
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**
	.655	.429	.420	.15190
Predictors: (Constant), W1HTOKFR, W2TOKLF, W2HTOKLF, STRESS, W2TYPFR, W2TYPLF, W1HTOKLF, W2OTYPFR, W1OTOKLF, W1CTYPLF, W2CORPFR				
Dependent Variable: PHYPHEN				

Table 5.21: All-feature model summary for hyphenated forms

Model Summary				
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**
	.831	.691	.685	.26324
Predictors: (Constant), W2CTYPLF, W2OTOKLF, W1CTYPLF, W1OTOKLF, SYLL, W1HTOKFR, TOTAL, W2HTYPFR, W2HTOKFR, W2CTOKLF, DOUBLEC, BSYLL				
Dependent Variable: POPEN				

Table 5.22: All-feature model summary for open forms

Tables 5.20 - 5.22 illustrate the interplay of features involved in concatenation decisions. In essence, no single feature determines whether a noun compound is concatenated or not. Rather, a cumulative effect is evidenced here. These models account for approximately 68% of the variance ($p<.001$) in the proportion of closed and open compounds, and 42% of the variance ($p<.001$) in hyphenated forms.

Both phonological and lexical features hold some predictive power. Lexical features, as a group, have a greater effect on the distribution of English compound orthography than do phonological features. But combining the two sets produces a slightly better model than the lexical model alone. Furthermore, in terms of individual features, the log frequency of the second component of closed compounds and the number of syllables in a compound appear to be the best predictors for closed and open forms.

5.5.2 Accuracy of the predictions

The regression model predictions for each compound in the 35+ list were computed and displayed in a table (see Appendix B, Table 2). An excerpt from this table, aligning the predicted proportion (i.e., PREDC, PREDH, PREDO) with the observed proportion of occurrences (PCLOSED, PHYPHEN, POPEN), is shown below in Table 5.23 for the purpose of discussion. To get an idea of the overall reliability of these predictions, let us look first at the word “policeman”. The predicted proportion of “policeman” is high compared to that of “police-man” or “police man”, since in the large majority of cases, when “man” occurs as the second component of a compound, that

compound is closed. In this corpus the actual percentage of occurrences in closed form for police+man was 99%. The corresponding prediction of the regression model was about 70%. Other compounds with "man" as the W2 showed even higher expected proportions for the closed form, and actual occurrences tended to match these predictions.

W1	W2	Closed	Hyphen	Open	Total	Proportion Closed	PredC	Proportion Hyphen	PredH	Proportion Open	PredO
child	abuse	0	5	54	59	.00	.027	.08	.146	.92	.819
drug	abuse	0	2	38	40	.00	.007	.05	.084	.95	.907
internet	access	0	3	51	54	.00	.000	.06	.000	.94	1.00
heart	attack	0	0	109	109	.00	.241	.00	.000	1.00	.748
self	esteem	0	81	5	86	.00	.142	.94	1.00	.06	.000
pine	apple	40	0	0	40	1.00	.481	0	.149	0	.395
sales	man	139	0	5	228	1.00	.912	.00	.000	.02	.069
chair	man	1284	0	0	1284	1.00	1.00	.00	.077	.00	.000
congress	man	159	0	0	159	1.00	.846	.00	.059	.00	.055
fisher	man	128	0	0	128	1.00	.855	.00	.120	.00	.000
police	man	152	0	1	153	.99	.704	.00	.008	.01	.252

Table 5.23: Illustrative examples of component closure predictions

Although the features used in this analysis showed substantial predictive power for closed forms, it was not always reliable. For example, the word "pine+apple", which occurred 40 times in the corpus, is always closed. Yet the prediction here was that it should occur closed only 48% of the time, as compared to the 39% expected rate for the open variant and 15% for the hyphenated form. This may be due in part to the effect of "pine tree" which occurred 19 times in the open form. Moreover, there is an internal-boundary vowel sequence in "pineapple".

An analysis of the results for open forms suggests that the component type is a fairly good predictor. Morphological "families" (de Jong et al., 2002) tended to produce

similar predictions. For example, compounds of the form noun+*system* (e.g., school system, operating system, computer system) show high expected proportions for nonconcatenation, and the predictions are supported by the data. Similarly, “child abuse” and ”drug abuse” (see Table 5.23) predictably favor the open form. In a few cases, where one component had a high type frequency but showed no strong concatenating bias, the predictions were not as successful. For example, “air conditioner” is predicted to have a 38% chance of occurring open, when in reality it occurred open 86% of the time. Furthermore, it was predicted that “air conditioner“ had a 38% chance of being closed, and in fact did not occur in the closed form at all. This discrepancy might be attributed to the fact that “air” also occurs in many high-frequency closed compounds – “airport”, “airplane”, etc., and underscores the fact that lexical features tend to outweigh phonological features.

Hyphenated forms were generally unpredictable using this approach. Most forms showed very low likelihoods for the hyphenation of two candidate components, except for “self” compounds (i.e., “self-esteem”, etc.), which showed consistently higher predictions. Proper nouns such as “Hewlett Packard” generally hovered around 50%, while the predictions for other predominantly hyphenated compounds such as “dot-com” and “price-fixing” were below 50%. Since the results for hyphenated forms in this case fail to explain the distribution of hyphenated compounds, it may be concluded that the particular feature set considered in this analysis does not adequately delimit the set of hyphenated compounds.

The results of the regression analysis show mostly accurate predictions, at least for the occurrence of open and closed forms. Moreover, they suggest that the frequency

of component types, particularly W2's, are linked to the orthographic form of a compound. However, phonological features also show significant effects. In fact, a combination of phonological and lexical features produces better results overall.

5.6 Free Variation

In a separate analysis, the set of free variants found in the entire corpus was considered. One of the problems in dealing with compounds has been to identify what actually constitutes a compound, other than concatenation. It is well established that, in English, compound nouns need not be concatenated. However, this "open" invitation serves to blur the line between compounds and phrases. Are all noun sequences compounds? Looking at forms that vary may at least provide us with a reliable basis on which to consider this question. The suggestion here is that if a noun sequence that is normally written open is occasionally or even mistakenly connected (i.e., by concatenation or hyphen), it is very likely that the sequence is a compound.

In the entire corpus, there were 167 compounds that varied over the three forms to a greater or lesser degree. In addition, 131 varied between hyphenated and closed forms, 1793 between hyphenated and open, and 741 between open and closed. Free variant distribution lists can be found in Appendix B, Tables 3a-d.

The closed-open-hyphenated group includes many words from popular culture including: "website", "webpage", "chatroom", "cellphone", and "facelift". The variation to some extent may be due to their "newness", meaning that the form has not yet been

standardized. Furthermore, this set includes many high frequency compounds, as 44 of the 167 (26%) are in the 35+ group.

The open-closed set likewise contains many high-frequency words such as "bookstore" and "bookshop". The number of syllables is limited for the most part to three. In addition, certain W2's appear in many compound types, for example, "house", "man", and "line". The latter would ordinarily be expected to occur in closed forms. The fact that lexical strings such as "farmhouse", "clothesline", and "taxman" also occur as "farm house", "clothes line", and "tax man" may underscore the notion that we do not need concatenation to connect lexical items. However, this does not preclude the idea that concatenation may serve some purpose.

In the hyphenated–closed set again there is a low syllable count. The most evident feature, however, is the large proportion of synthetic compounds: 23 with the agentive *-er* and 21 with the gerundive *–ing* suffix among the 131 types in this group. This type of compounding is highly productive and may be prone to hyphenation perhaps because of the "object – verb" structure. That is, many synthetic or verbal compounds may be interpreted as "reversed" verb phrases: a "trend-setter" sets trends, a "house-painter" paints houses, etc. So it seems plausible that hyphenation might indicate a reluctance to concatenate noncanonical lexical orderings. On the other hand, the frequency counts for the synthetic compounds were generally low. Of the 23 agentive forms, only 3 had a combined count of 35 or greater, and 14 had a combined count of just 2. Among the gerundive forms, only one was in the 35+ list, and 7 had a combined count of 2. So while one could posit some semantic motivation for variation in these cases, it seems more likely that frequency plays a greater role in the choice between closed and hyphenated *–er* and *–ing* forms. The

frequencies for the remainder of the closed-hyphenated set were also low overall, indicating that, where there is variation, the tendency to hyphenate, or perhaps simply the reluctance to concatenate, is linked to the particular compound's degree of usage.

As for the open-hyphenated variants, the only salient connection appears to be that many of the compounds contain five or more syllables. Other than this there are some "graphotactic" (Neijt 2002) preferences such as V-V or identical consonant sequences which resist concatenation. For example, "grab bag" and "grab-bag", but not "grabbag" are found in the corpus.

5.7 Etymology

Etymology was also considered but not explored deeply enough to conclusively determine its effects. Based upon a snapshot view, the possibility of component origin (e.g., Latin or Germanic) contributing to the form of the compound is not evident. From the small set of examples enumerated below, it appears that etymology plays at best a minimal role in the open/closed dichotomy. Native words are frequently short and monosyllabic ("foot", "house", "book"), while words of classical origin are more likely to be polysyllabic ("intelligence", "information", "independence") Thus, it is possible that one might observe a tendency for words with Germanic roots to concatenate when compounded. But there are also a substantial number of Latinate monosyllabic words which are (closed) compound components. So it would seem that the origin of a word does not significantly bias its behavior in compounding. The table below illustrates this:

COMPOUND	FORM FREQUENCY (closed/ hyphen/open)	W1 ORIGIN	W2 ORIGIN
airport	648 / 1 / 0	Latin	Latin
courthouse	104 / 0 / 0	Latin	Germanic
football	884 / 0 / 0	Germanic	Germanic
postcard	39 / 0 / 1	Latin	Greek
salesman	139 / 0 / 0	Germanic	Germanic
vineyard	96 / 0 / 0	Latin	Germanic

Table 5.24: Closed compound component origin

Rather than etymology, then, concatenation seems to be favored in these cases due to the length, i.e., the phonological length or metrical shape of the compound. Perhaps a more in-depth look at the role of etymology in compounding would reveal more substantial effects.

5.8 Linguistic Generalizations

Based upon an analysis of the data, the following generalizations can be made with respect to the surface form of English noun compounds:

1. Closed forms have a syllable limit.
2. Vowel-vowel sequences are avoided at internal lexical boundaries in closed forms.
3. Identical consonant sequences are avoided at the internal boundary in closed forms.
4. Closed forms follow word stress rules; open and hyphenated forms may or may not.
5. Hyphenated compounds frequently do not have compound stress.

6. Neither compound stress nor concatenation is criterial in defining compounds.

7. Compound stress may be a consequence of concatenation.

8. Closed compounds have, for the most part, binary foot structure.

9. The maximum number of feet in closed forms seems to be 2.

10. The components of compounds can sometimes predict whether a compound is closed, hyphenated, or open.

11. Compounds which vary in form provide a good foundation for examining and defining compounds.

12. A combination of lexical and phonological features is best for predicting the orthographic form of a noun compound.

Chapter Six

GENERAL CONCLUSIONS

The aim in this final chapter will be to tie the pieces together. Therefore, we will revisit the goals that were set forth in Chapter 1. A discussion of the findings relative to these goals and their possible applications will follow, as well as suggestions for further studies. But before reviewing these points, some ancillary but relevant observations will be considered.

6.1 English vs. Estonian compounds

The present study was not designed as a comparison of English and Estonian compounding. There were nevertheless some points of overlap between the two systems which merit some notice.

In general, the structure of English is vastly different from the structure of Estonian. So it is not surprising that compounds pattern differently in these two languages. Estonian does not have, for example, the orthographic variation present in English forms. Nor does it present the problem of figuring out whether a string is a compound or a syntactic phrase, since compounds are, as a rule, concatenated. English, on the other hand, does not require a decision about which form of the W1 to use. Yet

despite the disparate patterning, English and Estonian compounds have a few things in common. For example, they both have some variation in form based largely on phonological preferences, and they both place main stress on the left element in most cases. And while Estonian uses the genitive form in many compounds, true genitives as we know them are written separately, as in English: e.g., *Jaani maja* (John-gen.sg. + house → "John's house") or *inimeste elud* (person-gen.pl. + lives → "people's lives"). In addition, an interesting and odd parallel exists between the two: in Estonian compounds, nominative case is assigned to W1's that express the material or substance of W2 (e.g., *palkmaja* → "log cabin"); English compounds choose phrasal stress in the same cases (e.g., "silver dollar", "fur coat", etc.). In other words, the marked forms – nominative case for Estonian and open compounds with phrasal stress for English – are motivated by the same semantic class. If this is a coincidence, it is at least a remarkable one. In any case, a deeper investigation of this phenomenon is certainly warranted. The focus in this study was not the semantics of compounding; however, the interaction of semantics and phonology within the domain of morphology must be acknowledged.

Examining a morphological process, or any linguistic process, in two structurally distant languages can provide insights which may not be available from more homogeneous data sets. The similarities seem all the more important when they are not taken for granted.

6.2 Revisiting the goals of the study

The present study sought to determine the motivations for the choice of variants in Estonian and English noun compounds. In order to establish a principled answer to this, the following goal set was defined and targeted:

a. To show the extent to which free variation occurs in both English and Estonian noun compounds and to discover the conditions under which a compound's form may vary freely.
b. To clarify the role of phonology in compounding.
c. To account for the distribution of open/hyphenated/closed orthographic forms in English noun compounds and the distribution of genitive/nominative forms in Estonian noun compounds.
d. To work towards a greater understanding of what defines a compound.

6.2.1 a. The extent to which free variation occurs in both English and Estonian noun compounds and the conditions under which a compound's form may vary freely.

There is variation in Estonian compounding; however, the results of the current study show that "free" variation is less pervasive than expected. The two prevalent forms in Estonian, GC and NC, in some cases are defined by semantic constraints. Where they are not, the form may vary freely. The data suggest that this variation is allowable based

on the principle of foot isochrony. In addition, it was observed that new and foreign word forms tend to vary more than established forms. Within varying forms, however, there is a strong bias in favor of the genitive form. This appears to be phonologically motivated. Specifically, the tendency is driven by two common principles: ease of pronunciation and ease of perception.

In English, conversely, there is a great deal of variation in compound forms. In this case, however, the variation itself is not phonological[14]; instead, it is manifested in the orthography, but motivated to a great extent by the phonology. Resistance to concatenation frequently results from one or more phonological constraints. Words that can concatenate often do, but are by no means required to. The conclusions reached in the analysis of English compounds indicate that variation is relatively free. As in Estonian, where there is variation, one form is generally preferred over another.

6.2.2 b. The role of phonology in compounding: evidence of phonological constraints.

This research was borne from the hypothesis that variation preferences are phonologically determined. The results of the analysis of Estonian noun compounds (ch.4) support this hypothesis. The choice of genitive vs. nominative compounds is available when the two variants satisfy foot timing requirements. Furthermore, the group of compounds which may vary is constrained by phonotactics (i.e., the Consonant Cluster Constraint), in addition to a broader set of semantic and lexical constraints.

---

[14] There seems to be a degree of free variation in stress placement, particularly on English open compounds, but this has not been explored in the present study.

The English data also revealed the substantial role of phonology in determining the form of compounds. Specifically, it was found that the number of syllables, stress placement, and phonotactics all play a role in the realization of noun compounds as closed, hyphenated, or open. The number of syllables proved to be the strongest phonological predictor. While compound stress did show a significant effect, its correlation to closed forms was considerably lower than the syllable feature (.343 vs. -.589). This seems to point (albeit indirectly) to a relatively weak correlation between compounds and stress. Put another way, if compound stress were criterial for compounds then one would expect a rather high correlation between form and stress.

Discussions of compounding have focused on stress as the phonological feature most involved in the process. In terms of variation in compounding, however, the role of phonology extends beyond the notion of stress. For English, the syllable in fact plays a more critical role. For Estonian, foot isochrony seems to guide variation.

6.2.3 c. The distribution of open/hyphenated/closed orthographic forms in English noun compounds and genitive/nominative forms in Estonian noun compounds.

As stated previously, for the set of Estonian compounds that may alternate between genitive and nominative, the genitive is nearly always preferred. The reason for this preference may be due to general phonological principles such as ease of perception and ease of pronunciation since the difference between nominative and genitive in Estonian is often the addition of a syllable in the genitive. In cases where the nominative

ends in a consonant, the genitive variant effectively inserts a vowel. The vowel then forms a salient word boundary, possibly making decomposition of the compound easier. Or in other cases, the genitive vowel may facilitate pronunciation by avoiding a consonant sequence which might require greater articulatory effort.

While the present study primarily sought to show the influence of phonological properties on the form of noun compounds, the English data indicates both phonological and lexical effects. Statistical tests showed the strongest feature to be the W2 log frequencies for closed tokens and types. The next strongest feature was the number of syllables. This pattern was true for both closed and open forms. The regression models for open and closed compounds presented in Chapter 5 reveal the cumulative effect of phonological and lexical features on the orthography of English compounds.

Hyphenated forms showed generally weak correlations except with W2 log frequencies for all tokens and types, W2 log frequencies for open tokens and types, and stress. The results for the hyphenated form suggest that perhaps this group should be handled separately, i.e., with the exception of hyphenated forms that also occur as closed or open.

6.2.4 d. Towards a greater understanding of what defines a compound.

We have dealt here with the varying surface form of noun compounds in Estonian and in English. Constraints on the form and motivations for the variation have been

proposed. But what does any of this tell us about compounding in general? Does it delimit in any way the set of nominal compounds?

While a definitive characterization of compounds cannot be claimed in this work, the conclusions derived from the data may indeed shed some light on such a characterization. For example, the statistical analysis of variables in the English data, specifically the correlations, support the claim by Bauer (1983, 1998), ten Hacken (1999), and others that stress is not criterial in compounding. Moreover, the results of this research suggest that no single feature serves to delimit the set of compounds. In addition, the complementary patterning of correlations to closed vs. open compounds attests to a significant relationship between the two forms. Hyphenated forms, on the other hand, did not reveal any such relationship.

The existence of free variants in English rules out concatenation as a criterion for compound status, though concatenation of constituents certainly creates an unambiguous link between them. The decision to concatenate or not, may be determined, to some extent, by analogy. The statistics also point to this since there was a strong correlation between the W2 and the concatenative property of compounds. We might suppose then that novel compounds with "man" or "way" as the second component are likely to be concatenated, whereas "rescue call" might be chosen instead of "rescuecall" just because "house call", "conference call" and "phone call" are common.

The pattern of variation in Estonian may also be indicative of this phenomenon. For example, it was observed that many of the compounds that show substantial variation contain foreign constituents, so analogy may not always be an option, e.g., *veebsepp/veebisepp*. Where constituents are native and presumably familiar, the

preference for one form may be much stronger. In any event, determining the precise role of analogy in compound formation would probably be accomplished more appropriately by a psycholinguistic study.

6.3 Application to Current Theories

The generalizations that have been made here may be grounded in some of the field's existing theories. For example, using a constraint-based theory such as Optimality Theory (Kager 2001, McCarthy 2001), we can account for the variation in form evidenced in this study. For example, the phonological generalizations for determining the surface orthographic form of English compounds can be converted into constraints and ranked in accordance with the obtained correlations as follows:

SYLL > BSYLL > DOUBLEC > VSEQ > CSTRESS

The tableaux below illustrate how this ranking can predict the preferred output:

police+man	SYLL	BSYLL	CSTRESS	DOUBLEC	VSEQ
policeman	-	-	-	-	-
police-man	*	-	*!	-	-
police man	*	*!	*	-	-

Tableau 6.1(a)

police+officer	SYLL	BSYLL	CSTRESS	DOUBLEC	VSEQ
policeofficer	*!	*		-	*
police-officer	-	*	*!	-	*
police officer	-	-	*	-	-

Tableau 6.1(b)

police+department	SYLL	BSYLL	CSTRESS	DOUBLEC	VSEQ
policedepartment	*!	*	-	-	-
police-department	-	*	*!	-	-
police department	-	-	*	-	-

Tableau 6.1(c)

Figure 6.1: Sample OT Tableaux

"Policeman" actually occurs in the closed form 152 of 153 times in the corpus. "Police officer" occurs only in the open form (frequency =125), and "police department" occurs open 48 of 49 times. It is not being claimed here that this is the ideal scheme within an Optimality framework since admittedly not all compounds will fit in as neatly as these. Rather, the tableaux are presented here as an example of how this issue might be approached within OT.

6.4 Suggestions for further studies

So much work has been done on compounding; nevertheless, there are some unexplored areas which may provide useful insights. A few of these have already been mentioned: an investigation of possible parallels between the semantic properties of Estonian and English compounds, a psycholinguistic study of the use of analogy in compounding, and the possible effect of etymology on the distribution of compound forms.

Another topic which needs more attention is the role of perceptual isochrony in linguistic variation. Lehiste's (1977) investigation of isochrony is persuasive and is in

sync with other work (Balota, et al., 2001) which emphasizes the perceptual strength of relative (as opposed to exact) distances.

6.5 The value of a corpus study

The use of corpora for this study made it possible to examine the actual use of compounds, rather than their possible use. Furthermore, because of the orthographic considerations, there was no other efficient way to carry out such an analysis. There are of course some difficulties in the logistics: choosing an appropriate corpus, cleaning the data of errors, etc. But the benefits outweigh the problems.

The generative tradition may focus more on the creative competence of language users; nevertheless, statistical analyses of actual language use are hardly incompatible with traditional theoretical approaches. In fact, studies of language competence would be virtually meaningless without evidence of the corresponding linguistic performance.

Appendix A: Estonian Compound Data

TABLE 1: Estonian compound data (92 varying forms followed by remaining 624, sorted alphabetically) "nv"=no variation

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
aed vili	nom	7	4.1
alkoholi joove	gen	5	94.4
ameti isik	gen	14	99.4
ameti koht	gen	53	99.7
ameti mees	gen	18	83.5
ameti post	gen	9	82.4
ameti ühing	gen	16	98.6
arvamus avaldus	nom	5	17.9
diplomaadi kohver	gen	5	78.7
hommiku eine	gen	5	99.2
hommiku söök	gen	14	99.6
isik koosseis	nom	5	1.1
jumala teenistus	gen	9	99.6
käe kell	gen	9	98.6
kaksik vend	nom	5	1.9
käsk kiri	nom	18	3.3
kauba hall	gen	6	66.5
kaup mees	nom	20	6.7
keele kasutus	gen	5	99.4
keele oskus	gen	17	99.3
kella aeg	gen	21	98.9
kiik tool	nom	11	1.8
kindlus tunne	nom	11	6.9
klassi õde	gen	5	99.2
kommerts direktor	nom	5	0.8
kontserdi saal	gen	15	94.2
kool meister	nom	13	0.2
kooli maja	gen	44	99.9
kooli õpetaja	gen	11	99.5
kruusa tee	gen	12	98.5
kultuuri elu	gen	8	99.7
kultuuri lugu	gen	6	99.5
kultuuri maja	gen	24	99.9
kultuuri minister	gen	27	99.8
kultuuri osakond	gen	13	99.9
kunsti muuseum	gen	45	99.9

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
kuuli pilduja	gen	8	96.1
lahk heli	nom	7	1.3
liidu vabariik	gen	5	5.5
linna osa	gen	73	99.5
linna pea	gen	141	99.8
linna valitsus	gen	222	99.8
lõikus kuu	nom	5	2.4
looma aed	gen	28	98.7
märtsi kuu	gen	5	98.4
masina värk	gen	8	99.6
meele olu	gen	48	99.2
mere mees	gen	31	95.1
mere tuul	gen	6	95.7
mere vesi	gen	5	99.6
mets loom	nom	9	2.1
metsa mees	gen	5	92.4
mootori kütus	gen	10	97.7
mõtte käik	gen	9	97.2
muinas jutt	nom	19	1
nädala vahetus	gen	61	99.8
nime kiri	gen	136	99.9
õppe aasta	gen	17	99.9
õppe laen	gen	5	99.3
õppe töö	gen	19	99.8
õppe tool	gen	6	99.8
õppe tund	gen	8	99.5
paberi leht	gen	19	95.4
päeva kord	gen	23	99.9
peo laud	gen	5	59.3
piir joon	nom	13	14.9
piiri punkt	gen	8	99.7
piiri valve	gen	30	99.4
põranda alune	gen	5	98.6
ring sõit	nom	5	1.4
risti usk	gen	9	99.7
rong käik	nom	11	1.3
seaduse rikkumine	gen	15	98.6
seisu koht	gen	194	99.8
seisu kord	gen	18	98.7
süda linn	nom	15	2.4
sügis õhtu	nom	171	2.8

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
suve öö	gen	10	99.8
suve päev	gen	11	99.3
täht kuju	nom	7	0.2
talve aed	gen	10	99.3
tegu viis	nom	7	0.6
telefoni number	gen	17	90.6
tõe meeli	gen	8	26.7
toidu aine	gen	26	78.7
toom kirik	nom	23	2.2
tule kahju	gen	32	99.8
vabaduse kaotus	gen	22	95.5
vaguni saatja	gen	8	97.8
vanuse klass	gen	7	98.8
vastutus tunne	nom	6	0.3
vere soon	gen	10	73.1
ääre linn+0	gen	26	nv
aate mees	gen	5	nv
ahju suu	gen	5	nv
aia maa	gen	6	nv
aia maja	gen	5	nv
aia saadus	gen	5	nv
aja arvamine	gen	5	nv
aja järk	gen	5	nv
aja kiri	gen	54	nv
aja kirjandus	gen	101	nv
aja kirjanik	gen	126	nv
aja leht	gen	218	nv
aja loolane	gen	41	nv
aja lugu	gen	168	nv
aja pikendus	gen	9	nv
aja vahemik	gen	18	nv
aja viide	gen	11	nv
ajalehe paber	gen	7	nv
ajaloo õpetaja	gen	5	nv
ajaloo teadus	gen	6	nv
akna klaas	gen	20	nv
akna laud	gen	26	nv
akna raam	gen	5	nv
alg klass	nom	11	nv

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
alg kool	nom	7	nv
alus pesu	nom	5	nv
alus püks	nom	8	nv
ämbliku võrk	gen	8	nv
ameti asutus	gen	7	nv
ameti auto	gen	8	nv
ameti võim	gen	5	nv
ämma emand	gen	23	nv
äratus kell	nom	12	nv
arengu ruum	gen	11	nv
armu asi	gen	6	nv
armu kadedus	gen	8	nv
arsti teaduskond	gen	7	nv
asi tõend	nom	9	nv
asja ajamine	gen	19	nv
asja armastaja	gen	5	nv
asja mees	gen	16	nv
asja olu	gen	97	nv
asja tundja	gen	47	nv
augusti kuu	gen	5	nv
baari lett	gen	5	nv
banaani vaba riik	gen	5	nv
bilansi maht	gen	11	nv
bussi jaam	gen	11	nv
bussi juht	gen	31	nv
bussi peatus	gen	23	nv
bussi pilet	gen	6	nv
bussi raha	gen	6	nv
detsembri kuu	gen	6	nv
ehitus järk	nom	5	nv
ehitus materjal	nom	6	nv
ehitus töö	nom	17	nv
elamis luba	sh gen[1]	17	nv
elamis pind	sh gen	13	nv
elatus miinimum	nom	11	nv
elektri jaam	gen	46	nv
elektri liin	gen	6	nv
elektri võrk	gen	6	nv
enese haletsus	gen	5	nv

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
enese hinnang	gen	6	nv
enese kaitse	gen	5	nv
enese kindlus	gen	15	nv
enese tapja	gen	10	nv
enese tapp	gen	18	nv
enese teostus	gen	7	nv
enese tunne	gen	23	nv
filmi kunst	gen	6	nv
filmi täht	gen	5	nv
firma juht	gen	6	nv
gaasi balloon	gen	7	nv
haak nõel	nom	5	nv
hamba arst	gen	8	nv
hamba hari	gen	12	nv
hamba pasta	gen	14	nv
hammas ratas	nom	8	nv
haridus osakond	nom	6	nv
haua plaat	gen	7	nv
haua plats	gen	15	nv
heina maa	gen	26	nv*[2]
herne supp	gen	6	nv
hinge õhk	gen	8	nv
hinge põhi	gen	7	nv
hinge rahu	gen	6	nv
hinge tõmme	gen	10	nv
hommiku poolik	gen	11	nv
hoo aeg	gen	107	nv
hotelli tuba	gen	8	nv
hundi jaht	gen	11	nv
huule pulk	gen	38	nv
ilma jaam	gen	8	nv
isiku pärane	gen	5	nv
iste koht	nom	6	nv
jaama hoone	gen	12	nv
jaani päev	gen	7	nv
jahi mees	gen	31	nv
jahi püss	gen	5	nv
jahi seltskond	gen	6	nv
jala käija	gen	10	nv

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
jala laba	gen	5	nv
jala vari	gen	5	nv
jalg ratas	nom	69	nv
jalg rattur	nom	21	nv
jalutus käik	nom	16	nv
järje kindel	gen	14	nv
järje kord	gen	78	nv
järje pidevus	gen	12	nv
järve kallas	gen	6	nv
jõe kallas	gen	11	nv
jõe vesi	gen	6	nv
joogi vesi	gen	25	nv
jõu jaam	gen	8	nv
jõu katsumine	gen	5	nv
jõu pingutus	gen	11	nv
jõu varu	gen	11	nv
juhi luba	gen	9	nv
juht kiri	nom	18	nv
juht nöör	nom	6	nv
jumala sõna	gen	6	nv
justiits minister	nom	17	nv
jutu ajamine	gen	18	nv
juustu kera	gen	8	nv
kaardi mäng	gen	7	nv
kaardi pakk	gen	5	nv
käe käik	gen	15	nv
käe kott	gen	16	nv
kaera helves	gen	5	nv
kaitse liit	nom	24	nv
kammer orkester	nom	7	nv
karistus määr	nom	5	nv
karja maa	gen	24	nv
kartuli põld	gen	5	nv
karus nahk	nom	6	nv
käsi raamat	nom	13	nv
kassi poeg	gen	5	nv
kasvu hoone	gen	15	nv
katuse alune	gen	6	nv
kauba laev	gen	17	nv

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
kauba maja	gen	63	nv
kauba vahetus	gen	17	nv
kaubandus keskus	nom	6	nv
kaugus hüppaja	nom	5	nv
keeris torm	nom	7	nv
keldri korrus	gen	7	nv
kiidu sõna	gen	5	nv
kilp konn	nom	6	nv
kindlustus selts	nom	21	nv
kindral staap	nom	6	nv
king sepp	nom	6	nv
kiriku kell	gen	6	nv
kiriku õpetaja	gen	9	nv
kirja mees	gen	7	nv
kirja oskus	gen	7	nv
kirja saatja	gen	5	nv
kirja töö	gen	16	nv
kirja tükk	gen	9	nv
kirja vahetus	gen	9	nv
kirjandus kriitik	nom	5	nv
kirjandus preemia	nom	14	nv
kirjandus teadlane	nom	7	nv
kirjandus tegu	nom	5	nv
kirjandus teos	nom	5	nv
kirjutus laud	nom	47	nv
kirjutus masin	nom	8	nv
klaasi kild	gen	5	nv
klassi juhataja	gen	10	nv
klassi kaaslane	gen	6	nv
klassi ruum	gen	6	nv
klassi tuba	gen	5	nv
klassi vend	gen	8	nv
kleidi saba	gen	5	nv
kohtu istung	gen	10	nv
kohtu kulu	gen	5	nv
kohtu otsus	gen	31	nv
kohtu protsess	gen	19	nv
kohtu saal	gen	6	nv
kohvi tass	gen	21	nv

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
kõie tantsija	gen	5	nv
koja mees	gen	12	nv
kõnni tee	gen	39	nv
konservi karp	gen	9	nv
köögi laud	gen	7	nv
köögi uks	gen	5	nv
köögi vili	gen	8	nv
kooli aasta	gen	6	nv
kooli aed	gen	6	nv
kooli aeg	gen	11	nv
kooli direktor	gen	9	nv
kooli kott	gen	5	nv
kooli laps	gen	19	nv
kooli õpilane	gen	8	nv
kooli päev	gen	7	nv
kooli pink	gen	6	nv
kooli poiss	gen	20	nv
kooli põlv	gen	5	nv
kooli raha	gen	6	nv
kooli tee	gen	7	nv
kooli vend	gen	7	nv
koondus laager	nom	8	nv
korra lagedus	gen	7	nv
korra pidaja	gen	55	nv
korteri naaber	gen	5	nv
korteri peremees	gen	5	nv
korteri uks	gen	7	nv
kraadi klaas	gen	7	nv
kraani kauss	gen	12	nv
kraavi perv	nom	5	nv
kriminaal asi	nom	68	nv
kriminaal koodeks	nom	9	nv
kruusa auk	gen	5	nv
kujutlus pilt	nom	10	nv
kujutlus võime	nom	8	nv
kuld nokk	nom	5	nv
kuller kupp	nom	7	nv
kultuuri keskus	gen	26	nv
kultuuri ministeerium	gen	35	nv

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
kultuuri sündmus	gen	5	nv
kultuuri üritus	gen	5	nv
kuning riik	nom	17	nv
kunsti elu	gen	8	nv
kunsti kogu	gen	6	nv
kunsti kool	gen	16	nv
kunsti kriitik	gen	5	nv
kunsti teadlane	gen	6	nv
kunsti teos	gen	12	nv
küüs lauk	nom	6	nv
kuuse oks	gen	13	nv
laeva mees	gen	8	nv
lagi pea	nom	6	nv
lahingu väli	gen	13	nv
lahk arvamus	nom	5	nv
lapse laps	gen	16	nv
lapse põlv	gen	71	nv
lapse vanem	gen	27	nv
laske moon	gen	18	nv
laua jupp	gen	5	nv
laua kaaslane	gen	5	nv
laulu koor	gen	5	nv
laulu pidu	gen	11	nv
laulu väljak	gen	13	nv
lehe külg	gen	42	nv
lehe lugeja	gen	5	nv
leid laps	nom	6	nv
lennu jaam	gen	34	nv
lennu vägi	gen	7	nv
lennu väli	gen	38	nv
liiva rand	gen	7	nv*
lille pott	gen	5	nv
linna hall	gen	5	nv
linna jagu	gen	6	nv
linna kodanik	gen	17	nv
linna maja	gen	5	nv
linna müür	gen	5	nv
linna pilt	gen	8	nv
linna rahvas	gen	16	nv

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
linna volinik	gen	6	nv
linnu vabrik	gen	5	nv
lipu kandja	gen	5	nv
lipu varras	gen	6	nv
liu väli	gen	12	nv
loengu sari	gen	8	nv
looma kaitseselts	gen	5	nv
looma liha	gen	8	nv
looma tohter	gen	5	nv
lõpu tunnistus	gen	12	nv
lugemis saal	sh gen	6	nv
lugu pidamine	nom	15	nv
lume memm	gen	6	nv
lume sadu	gen	7	nv
lume torm	gen	7	nv
luua vars	gen	6	nv
maakonna keskus	gen	5	nv
maali kunst	gen	13	nv
madal rõhkkond	nom	13	nv
mäe tipp	gen	10	nv
magamis tuba	sh gen	13	nv
majandus kriis	nom	11	nv
majandus ministeerium	nom	28	nv
majandus teadlane	nom	5	nv
mälestus märk	nom	21	nv
mälestus päev	nom	6	nv
mängu asi	gen	42	nv
mängu automaat	gen	5	nv
mängu reegel	gen	7	nv
mängu ruum	gen	5	nv
männi salu	gen	5	nv*
märgu anne	gen	5	nv
märk sõna	nom	15	nv
meditsiini õde	gen	10	nv
meele härm	gen	5	nv
meele hea	gen	5	nv
meele heide	gen	14	nv
meele kindlus	gen	5	nv
meele lahutus	gen	15	nv

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
meele mürk	gen	6	nv
meele paha	gen	6	nv
meele vald	gen	10	nv
mees sugu	nom	5	nv
meeskonna liige	gen	5	nv
mere kool	gen	6	nv
mere laevandus	gen	5	nv
mere rand	gen	10	nv*
mere sõitja	gen	5	nv
mere vägi	gen	9	nv
mesi nädal	nom	5	nv
mets kits	nom	7	nv
mets siga	nom	7	nv
metsa talu	gen	5	nv
metsa tukk	gen	14	nv
metsa vend	gen	14	nv
mõisa hoone	gen	6	nv
mõõdu puu	gen	12	nv
mootor paat	nom	7	nv
mootor ratas	nom	21	nv
mõtte avaldus	gen	5	nv
mõtte kaaslane	gen	7	nv
mõtte laad	gen	10	nv
muster näide	nom	5	nv
müügi lett	gen	5	nv
muuseumi töötaja	gen	5	nv
naabri mees	gen	13	nv
naabri naine	gen	16	nv
naabri poiss	gen	5	nv
nädala lõpp	gen	10	nv
nädala päev	gen	11	nv
naeru alune	gen	5	nv
naeru pahvak	gen	7	nv
nahk tagi	nom	25	nv
nais kunstnik	sh gen	5	nv
nais sugu	sh gen	7	nv
näite mäng	gen	8	nv
näite ring	gen	5	nv
nalja mees	gen	6	nv

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
näo ilme	gen	15	nv
näo joon	gen	9	nv
näpu näide	gen	12	nv
näpu ots	gen	6	nv
neegri naine	gen	6	nv
niidi rull	gen	6	nv
noa haav	gen	7	nv
noa tera	gen	6	nv
novembri kuu	gen	6	nv
nuku teater	gen	9	nv
numbri märk	gen	8	nv
numbri tuba	gen	5	nv
õhu pall	gen	8	nv
õhu temperatuur	gen	18	nv
ohu tunne	gen	28	nv
okas traat	nom	5	nv
oktoobri kuu	gen	5	nv
okupatsiooni aeg	gen	6	nv
õlle tehas	gen	13	nv
omandi õigus	gen	9	nv
õmblus masin	nom	5	nv
õnnetus juhtum	nom	9	nv
oote ruum	gen	8	nv
õppe asutus	gen	14	nv
õppe hoone	gen	5	nv
õppe jõud	gen	54	nv
osakonna juhataja	gen	13	nv
ostu jõud	gen	6	nv
ots tarve	nom	26	nv
otsa sein	gen	8	nv
otsustus õigus	nom	5	nv
õuna puu	gen	19	nv*
paberi tükk	gen	7	nv
päeva palk	gen	6	nv
päeva pea	gen	12	nv
päeva raamat	gen	6	nv
päeva raha	gen	5	nv
päeva uudis	gen	9	nv
päeva valgus	gen	7	nv

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
pagasi ruum	gen	6	nv
päikese paiste	gen	7	nv
päikese tõus	gen	6	nv
päikese valgus	gen	12	nv
pais järv	nom	5	nv
palga lisa	gen	5	nv
palga päev	gen	6	nv
palga raha	gen	7	nv
paneel maja	nom	5	nv
panga arve	gen	7	nv
panga rööv	gen	5	nv
pargi pink	gen	6	nv
parlamendi liige	gen	5	nv
parlamendi saadik	gen	9	nv
pealis pind	nom	7	nv
peegel pilt	nom	6	nv
peig mees	nom	8	nv
peo pesa	gen	22	nv
perekonna elu	gen	5	nv
perekonna liige	gen	6	nv
perekonna nimi	gen	11	nv
piima nõu	gen	6	nv
piiri maa	gen	8	nv*
piiri valvur	gen	11	nv
pildi kast	gen	6	nv
pilli mees	gen	10	nv
pingi naaber	gen	6	nv
pipar kook	nom	6	nv
põhi kool	nom	145	nv
põhi osa	nom	83	nv
poiss mees	nom	6	nv
poja poeg	gen	7	nv
põllu maa	gen	8	nv*
põllu majandus	gen	59	nv
põllu mees	gen	32	nv
põllu töö	gen	5	nv
poole hoidja	nom	8	nv
post kaart	nom	7	nv
post kast	nom	22	nv

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
post kontor	nom	15	nv
posti maja	gen	5	nv
põue tasku	gen	15	nv
pritsi mees	gen	5	nv
puhke päev	gen	12	nv
puhke paik	gen	6	nv
püksi rihm	gen	12	nv
puna armee	nom	10	nv
puna pea	nom	5	nv
püstol kuulipilduja	nom	9	nv
raamatu kauplus	gen	5	nv
raamatu kogu	gen	107	nv
raamatu pidaja	gen	16	nv
raamatu pidamine	gen	22	nv
raamatu pood	gen	10	nv
raamatu riid	gen	11	nv
raamatu tarkus	gen	6	nv
rae koda	gen	35	nv
rahandus ministeerium	nom	49	nv
rahva hulk	gen	11	nv
rahva maja	gen	18	nv
rahvus kaaslane	nom	5	nv
rahvus kultuur	nom	8	nv
rahvus riik	nom	8	nv
ranna küla	gen	20	nv
ranna liiv	gen	10	nv
ratas tool	nom	6	nv
raud tee	nom	82	nv
raud tee jaam	nom	5	nv
raud tee tamm	nom	23	nv
raud teelane	nom	5	nv
rea mees	gen	5	nv
rehe tuba	gen	6	nv
reisi kulu	gen	5	nv
reisi laev	gen	5	nv
relva äri	gen	9	nv
riide kapp	gen	8	nv
riide tükk	gen	10	nv
riigi ametnik	gen	36	nv

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
riigi ettevõte	gen	17	nv
riigi juht	gen	10	nv
riigi kaitse	gen	21	nv
riigi kogu	gen	427	nv
riigi mees	gen	8	nv
riigi pea	gen	15	nv
riigi piir	gen	7	nv
riigi pööre	gen	5	nv
rind kere	nom	5	nv
ring kaitse	nom	6	nv
rinna hoidja	gen	8	nv
rinna vähk	gen	6	nv
rist sõna	nom	7	nv
rühma ülem	gen	9	nv
rukki lill	gen	9	nv
ruut meeter	nom	53	nv
saatuse kaaslane	gen	5	nv
sae puru	gen	9	nv
sajandi lõpp	gen	5	nv
sajandi vahetus	gen	5	nv
sea liha	gen	12	nv
seadus pära	nom	5	nv
seadus pärasus	nom	10	nv
seebi mull	gen	7	nv
selgitus töö	nom	5	nv
selja kott	gen	28	nv
selja tugi	gen	9	nv
seltsi mees	gen	59	nv
septembri kuu	gen	7	nv
siht asutus	nom	31	nv
siht koht	nom	6	nv
siili soeng	gen	5	nv
silma nurk	gen	9	nv
silma piir	gen	30	nv
silma pilk	gen	54	nv
silma vaade	gen	11	nv
sireli põõsas	gen	10	nv
sõja aeg	gen	5	nv
sõja laev	gen	13	nv

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
sõja mees	gen	9	nv
sõja väebaas	gen	6	nv
sõja väelane	gen	7	nv
sõja väeteenistus	gen	24	nv
sõja vägi	gen	83	nv
sõja vang	gen	5	nv
sõja veteran	gen	5	nv
söögi isu	gen	5	nv
söögi laud	gen	15	nv
söögi maja	gen	5	nv
söögi raha	gen	7	nv
sooja kraad	gen	5	nv
spordi laager	gen	5	nv
spordi võistlus	gen	5	nv
staabi ülem	gen	9	nv
süda öö	nom	19	nv
südame atakk	gen	5	nv
südame löök	gen	6	nv
südame tunnistus	gen	24	nv
sündmus koht	nom	16	nv
sünni maa	gen	6	nv
sünni päev	gen	65	nv
sünni päeva laps	gen	7	nv
suurus järk	nom	10	nv
suve ilm	gen	5	nv
suve laager	gen	5	nv
taeva keha	gen	7	nv
täht päev	nom	11	nv
takti tunne	gen	7	nv
tänava nurk	gen	17	nv
tantsu põrand	gen	5	nv
teatri maja	gen	6	nv
tegevus ala	nom	6	nv
tegevus kava	nom	10	nv
telefoni kõne	gen	24	nv
telefoni raamat	gen	5	nv
telefoni toru	gen	24	nv
tellis kivi	nom	15	nv
tiib hoone	nom	11	nv

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
tiku toos	gen	5	nv
tipp juht	nom	20	nv
toa nurk	gen	5	nv
tõe näosus	gen	40	nv
tõe otsing	gen	5	nv
toidu laud	gen	5	nv
toidu õli	gen	7	nv
toidu pood	gen	7	nv
toidu raha	gen	6	nv
tõkke puu	gen	5	nv
tolli ametnik	gen	12	nv
toone kurg	gen	6	nv
tõsi asi	nom	27	nv
trepi aste	gen	5	nv
trepi koda	gen	21	nv
trüki koda	gen	6	nv
trüki sõna	gen	5	nv
tsiviil isik	nom	8	nv
tugi tool	nom	36	nv
tule tõrje	gen	11	nv
tule tõrjuja	gen	16	nv
tuleviku plaan	gen	6	nv
tuli relv	nom	7	nv
turm tuli	nom	5	nv
turu hind	gen	18	nv
tütar laps	nom	49	nv
tuule hoog	gen	11	nv
tuuma relv	gen	11	nv
tüüri mees	gen	9	nv
ühend riik	nom	51	nv
ukse ava	gen	9	nv
ukse link	gen	5	nv
une nägu	gen	79	nv
usaldus väärsus	nom	9	nv
usu lahk	gen	5	nv
uudis himu	nom	33	nv
uurimis organ	sh gen	6	nv
uurimis töö	sh gen	15	nv
vaate aken	gen	9	nv

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
vaate mäng	gen	7	nv
vaate nurk	gen	10	nv
vaate pilt	gen	16	nv
vaate punkt	gen	6	nv
vaate väli	gen	12	nv
vabadus sõda	nom	14	nv
vabadus võitleja	nom	7	nv
vägi vald	nom	33	nv
vahi alune	gen	8	nv
vahi mees	gen	7	nv
vahu koor	gen	5	nv
vahu vein	gen	7	nv
valgus foor	nom	8	nv
väli politsei	nom	6	nv
valimis kampaania	sh gen	20	nv
valimis künnis	sh gen	9	nv
väljendus vahend	nom	7	nv
valla maja	gen	19	nv
valla vanem	gen	53	nv
vanadus pension	nom	11	nv
vande nõu	gen	12	nv
vande sõna	gen	5	nv
vangi kong	gen	5	nv
vangi laager	gen	8	nv
vangi maja	gen	6	nv
vangi valvur	gen	6	nv
vanni tuba	gen	40	nv
värava vaht	gen	12	nv
varju nimi	gen	8	nv
varju paik	gen	16	nv
varju teater	gen	8	nv
vee juga	gen	7	nv
vee kogu	gen	10	nv
vee toru	gen	5	nv
veebruari kuu	gen	5	nv
veini pudel	gen	14	nv
veo auto	gen	36	nv
vere jooks	gen	7	nv
vere rõhk	gen	11	nv

Appendix A, Table 1 (cont.)

NOUN COMPOUNDS (N=716)	W1 CASE in Tartu Corpus	Frequency in Tartu Corpus	Percent GC in Google search
vestlus kaaslane	nom	6	nv
vihma sadu	gen	13	nv
vihma vari	gen	13	nv
vihma vesi	gen	9	nv
viina pits	gen	5	nv
viina pudel	gen	8	nv
vilja puu	gen	6	nv*
vint püss	nom	10	nv
võidu käik	gen	8	nv
võimu mees	gen	12	nv
võimu võitlus	gen	7	nv
võla kiri	gen	16	nv
voodi koht	gen	5	nv
võõra maalane	gen	7	nv
võrk pall	nom	7	nv

1. "sh. gen" = "shortened genitive"
2. "*" nominative form is found only as a surname. (i.e., Heinmaa, Liivrand, Mändsalu, Merirand, Õunpuu, Piirmaa, Põldmaa, and Vilipuu)

Appendix B: English Compound Data

Table 1: English Data for most frequent compounds (35+): relative frequencies and phonological variables (type count = 708*; "CS"=compound stress)

W1	W2	CLOSED	HYPHEN	OPEN	TOTAL	VSEQ	doubleC	SYLL	CS
news	paper	2506	0	3	2509	0	0	3	1
net	work	2009	0	0	2009	0	0	2	1
sun	day	1873	0	0	1873	0	0	2	1
week	end	1351	0	2	1353	0	0	2	1
chair	man	1284	0	0	1284	0	0	2	1
base	ball	1175	0	0	1175	0	0	2	1
head	line	994	0	1	995	0	0	2	1
news	week	965	0	1	966	0	0	2	1
john	son	936	0	0	936	0	0	2	1
chatter	box	930	0	0	930	0	0	3	1
foot	ball	884	0	0	884	0	0	2	1
front	page	11	818	11	840	0	0	2	0
holly	wood	808	0	0	808	0	0	3	1
air	line	767	0	4	771	0	0	2	1
back	ground	768	0	0	768	0	0	2	1
guide	line	764	0	0	764	0	0	2	1
health	care	70	152	517	739	0	0	2	1
credit	card	0	42	646	688	0	0	3	1
jack	son	673	0	0	673	0	0	2	1
web	site	149	6	515	670	0	0	2	1
air	port	648	1	0	649	0	0	2	1
cover	story	0	0	615	615	0	0	4	1
law	suit	535	0	3	538	0	0	2	1
basket	ball	499	0	0	499	0	0	3	1
stock	market	0	26	435	461	0	0	3	1
bed	room	430	0	1	431	0	0	2	1
world	war	0	0	418	418	0	0	2	0
cow	boy	391	0	0	391	0	0	2	1
tax	payer	374	1	12	387	0	0	3	1
death	penalty	0	16	353	369	0	0	4	1
market	place	358	0	11	369	0	0	3	1
frame	work	354	0	1	355	0	0	2	1
law	enforcement	0	29	314	343	1	0	4	1
air	craft	336	0	0	336	0	0	2	1
house	hold	332	0	0	332	0	0	2	1
interest	rate	0	23	307	330	0	0	3	1
life	time	321	3	6	330	0	0	2	1
green	span	329	0	0	329	0	0	2	1
data	base	300	1	27	328	0	0	3	1
op	ed	0	318	0	318	0	0	2	0
share	holder	312	0	0	312	0	0	3	1
break	fast	311	0	0	311	0	0	2	1
birth	day	305	0	0	305	0	0	2	1

Appendix B, Table 1 (cont.)

W1	W2	CLOSED	HYPHEN	OPEN	TOTAL	VSEQ	doubleC	SYLL	CS
head	quarter	286	0	0	286	0	0	3	1
home	land	285	0	0	285	0	0	2	1
life	style	269	1	12	282	0	0	2	1
work	shop	276	0	1	277	0	0	2	1
fly	trap	276	0	0	276	0	0	2	1
girl	friend	269	0	6	275	0	0	2	1
bath	room	269	0	0	269	0	0	2	1
air	plane	264	0	0	264	0	0	2	1
class	room	259	0	1	260	0	0	2	1
quarter	back	257	0	0	257	0	0	3	1
work	place	211	0	45	256	0	0	2	1
school	system	0	0	256	256	0	0	3	1
north	west	247	2	1	250	0	0	2	0
law	maker	249	0	0	249	0	0	3	1
culture	box	241	0	0	241	0	0	3	1
pipe	line	240	0	1	241	0	0	2	1
robin	son	239	0	0	239	0	0	3	1
scot	land	238	0	0	238	0	0	2	1
health	insurance	0	7	229	236	0	0	4	1
income	tax	0	16	213	229	0	0	3	1
air	force	3	1	224	228	0	0	2	1
business	man	223	0	5	228	0	0	3	1
earth	quake	224	0	0	224	0	0	2	1
day	time	210	0	7	217	0	0	2	1
gun	control	0	8	209	217	0	0	3	1
stock	price	0	5	212	217	0	0	2	1
child	care	8	36	168	212	0	0	2	1
bloom	berg	210	0	0	210	0	0	2	1
water	gate	210	0	0	210	0	0	3	1
pay	roll	204	0	0	204	0	0	2	1
rail	road	201	0	2	203	0	0	2	1
day	care	2	38	157	197	0	0	2	1
tax	cut	0	4	193	197	0	0	2	1
phone	call	0	1	192	193	0	0	2	1
north	east	186	1	3	190	0	0	2	0
master	piece	189	0	0	189	0	0	3	1
sand	stone	189	0	0	189	0	0	2	1
ash	croft	186	0	0	186	0	0	2	1
cell	phone	23	19	137	179	0	0	2	1
self	regulation	0	177	2	179	0	0	5	0
war	fare	178	0	0	178	0	0	2	1
south	east	175	2	0	177	0	0	2	0
press	conference	0	1	174	175	0	0	3	1
book	store	162	0	11	173	0	0	2	1
text	book	169	0	4	173	0	0	2	1
boy	friend	162	0	3	165	0	0	2	1

Appendix B, Table 1 (cont.)

W1	W2	CLOSED	HYPHEN	OPEN	TOTAL	VSEQ	doubleC	SYLL	CS
campaign	finance	0	25	137	162	0	0	4	0
video	tape	148	0	13	161	0	0	4	1
congress	man	159	0	0	159	0	0	3	1
consumer	protection	0	1	154	155	0	0	5	0
family	member	0	0	155	155	0	0	4	1
block	buster	153	1	0	154	0	0	3	1
hand	gun	137	1	16	154	0	0	2	1
note	book	154	0	0	154	0	0	2	1
pea	nut	153	0	0	153	0	0	2	1
police	man	152	0	1	153	0	0	3	1
daytime	phone	0	0	151	151	0	0	3	0
land	mark	151	0	0	151	0	0	2	1
stock	option	0	20	128	148	0	0	3	1
white	water	147	0	0	147	0	0	3	1
back	lash	145	0	0	145	0	0	2	1
grocery	store	0	1	144	145	0	0	3	1
team	mate	144	0	0	144	0	1	2	1
oak	land	140	0	0	140	0	0	2	1
lime	stone	138	1	0	139	0	0	2	1
sales	man	139	0	0	139	0	0	2	1
information	technology	0	9	128	137	0	0	5	0
lap	top	136	0	0	136	0	0	2	1
court	room	132	0	3	135	0	0	2	1
back	yard	133	0	1	134	0	0	2	0
pay	check	134	0	0	134	0	0	2	1
school	district	0	1	133	134	0	0	3	1
sea	food	132	0	2	134	0	0	2	1
south	west	132	0	1	133	0	0	2	0
side	walk	128	0	1	129	0	0	2	1
tobacco	company	0	4	125	129	0	0	5	1
fisher	man	128	0	0	128	0	0	3	1
ware	house	128	0	0	128	0	0	2	1
insurance	company	0	0	127	127	0	0	5	1
wrong	doing	126	0	0	126	0	0	3	1
heart	disease	0	3	122	125	0	0	3	1
police	officer	0	0	125	125	1	0	5	1
air	strike	105	0	19	124	0	0	2	1
film	maker	122	1	1	124	0	1	3	1
rail	way	124	0	0	124	0	0	2	1
thanks	giving	124	0	0	124	0	0	3	0
room	mate	122	0	1	123	0	1	2	1
air	conditioning	0	37	85	122	0	0	5	1
motor	cycle	122	0	0	122	0	0	4	1
air	way	121	0	0	121	0	0	2	1
fire	fighter	120	0	1	121	0	0	4	1
news	hour	118	0	3	121	0	0	2	1

Appendix B, Table 1 (cont.)

W1	W2	CLOSED	HYPHEN	OPEN	TOTAL	VSEQ	doubleC	SYLL	CS
court	yard	120	0	0	120	0	0	2	1
monday	morning	0	1	119	120	0	0	4	0
wild	life	120	0	0	120	0	0	2	1
copy	right	119	0	0	119	0	0	3	1
country	side	119	0	0	119	0	0	3	1
death	row	0	14	105	119	0	0	2	0
home	work	117	0	2	119	0	0	2	1
capital	punishment	0	3	115	118	0	0	5	0
breast	cancer	0	6	111	117	0	0	3	1
welfare	reform	0	15	102	117	0	1	4	1
work	force	69	0	48	117	0	0	2	1
fire	arm	111	0	5	116	1	0	3	1
view	point	113	0	3	116	0	0	2	1
wheel	chair	112	1	3	116	0	0	2	1
sun	set	113	0	2	115	0	0	2	1
competition	policy	0	0	114	114	0	0	5	0
consent	order	0	0	114	114	0	0	4	1
summer	time	81	0	33	114	0	0	3	1
trade	mark	113	0	0	113	0	0	2	1
sales	tax	0	5	107	112	0	0	2	1
market	power	0	1	110	111	0	0	4	1
gold	berg	110	0	0	110	0	0	2	1
gold	man	110	0	0	110	0	0	2	1
news	day	96	0	14	110	0	0	2	1
stand	point	109	0	1	110	0	0	2	1
heart	attack	0	0	109	109	0	0	3	1
saturday	night	0	3	106	109	0	0	4	0
district	court	0	0	107	107	0	0	3	0
judgment	column	0	0	107	107	0	0	4	1
state	department	0	0	107	107	0	0	4	1
cock	tail	106	0	0	106	0	0	2	1
house	republican	0	0	106	106	0	0	5	1
capital	gang	0	0	105	105	0	0	4	0
class	mate	104	0	0	104	0	0	2	1
court	house	104	0	0	104	0	0	2	1
dinner	party	0	0	104	104	0	0	4	1
green	house	104	0	0	104	0	0	2	1
loop	hole	102	0	2	104	0	0	2	1
coca	cola	0	103	0	103	0	0	4	0
family	reunion	0	0	103	103	0	0	5	0
market	share	2	4	97	103	0	0	3	1
air	pollution	0	0	102	102	0	0	4	1
gate	way	102	0	0	102	0	0	2	1
land	lord	102	0	0	102	0	0	2	1
thai	land	102	0	0	102	0	0	2	1
coast	line	100	0	1	101	0	0	2	1

Appendix B, Table 1 (cont.)

W1	W2	CLOSED	HYPHEN	OPEN	TOTAL	VSEQ	doubleC	SYLL	CS
east	coast	0	1	99	100	0	0	2	0
foot	note	100	0	0	100	0	0	2	1
table	spoon	100	0	0	100	0	0	3	1
key	board	99	0	0	99	0	0	2	1
desk	top	97	0	1	98	0	0	2	1
enforcement	action	0	0	98	98	0	0	5	1
government	official	0	0	98	98	0	0	5	0
peace	process	0	0	98	98	0	0	3	1
fire	work	97	0	0	97	0	0	3	1
golf	course	0	1	96	97	0	0	2	1
play	wright	97	0	0	97	0	0	2	1
stock	holder	96	0	1	97	0	0	3	1
brand	name	0	40	56	96	0	0	2	1
college	student	0	0	96	96	0	0	4	1
main	frame	96	0	0	96	0	0	2	1
straw	berry	96	0	0	96	0	0	3	1
vine	yard	96	0	0	96	0	0	2	1
water	front	96	0	0	96	0	0	3	1
club	house	95	0	0	95	0	0	2	1
news	conference	0	0	95	95	0	0	3	1
senate	majority	0	0	95	95	0	0	5	0
world	series	0	0	95	95	0	0	3	0
side	bar	93	0	1	94	0	0	2	1
fund	raiser	9	78	6	93	0	0	3	1
wave	length	90	1	1	92	0	0	2	1
back	drop	91	0	0	91	0	0	2	1
green	berg	91	0	0	91	0	0	2	1
path	way	91	0	0	91	0	0	2	1
press	release	0	1	89	90	0	0	3	1
department	store	0	6	83	89	0	0	4	1
ground	troop	0	1	88	89	0	0	2	1
sun	time	0	89	0	89	0	0	2	0
tv	show	0	0	89	89	0	0	3	1
drug	use	0	1	87	88	0	0	2	1
paper	work	82	0	5	87	0	0	3	1
robinson	patman	0	87	0	87	0	0	5	0
side	effect	0	7	80	87	1	0	3	1
box	office	0	31	55	86	0	0	3	1
self	esteem	0	81	5	86	0	0	3	0
gun	man	85	0	0	85	0	0	2	1
land	fill	85	0	0	85	0	0	2	1
spot	light	84	0	1	85	0	0	2	1
bore	hole	84	0	0	84	0	0	2	1
cock	pit	84	0	0	84	0	0	2	1
front	pager	0	84	0	84	0	0	3	0
hill	side	84	0	0	84	0	0	2	1

Appendix B, Table 1 (cont.)

W1	W2	CLOSED	HYPHEN	OPEN	TOTAL	VSEQ	doubleC	SYLL	CS
prescription	drug	0	3	81	84	0	0	4	0
sun	shine	83	1	0	84	0	0	2	1
air	bag	9	8	66	83	0	0	2	1
boy	scout	0	0	83	83	0	0	2	1
drug	company	0	2	81	83	0	0	4	1
homeland	security	0	6	77	83	0	0	5	0
reform	party	0	0	83	83	0	0	4	1
school	teacher	28	0	55	83	0	0	3	1
terrorist	attack	0	0	83	83	0	0	5	1
birth	control	0	21	61	82	0	0	3	1
court	decision	0	0	82	82	0	0	4	1
drive	way	81	0	1	82	0	0	2	1
food	stamp	0	15	67	82	0	0	2	1
home	page	5	0	77	82	0	0	2	1
monday	night	0	0	82	82	0	0	3	0
crime	rate	0	0	81	81	0	0	2	1
tea	spoon	81	0	0	81	0	0	2	1
tobacco	industry	0	9	72	81	1	0	5	1
executive	director	0	0	80	80	0	0	5	0
head	ache	80	0	0	80	0	0	2	1
patter	son	80	0	0	80	0	0	3	1
phone	number	0	0	80	80	0	1	3	1
family	value	0	10	69	79	0	0	4	0
george	town	79	0	0	79	0	0	2	1
graduate	student	0	0	79	79	0	0	5	1
home	owner	67	0	12	79	1	0	3	1
power	plant	6	2	71	79	0	0	3	1
role	model	0	1	78	79	0	0	3	1
news	story*	0	0	78	78	0	1	3	1
pass	port	78	0	0	78	0	0	2	1
peace	keeping	76	2	0	78	0	0	3	1
drug	test	0	2	75	77	0	0	2	1
law	school	0	4	73	77	0	0	2	1
port	land	77	0	0	77	0	0	2	1
campaign	contribution	0	2	74	76	0	0	5	0
east	wood	76	0	0	76	0	0	2	1
ground	water	66	5	5	76	0	0	3	1
property	right	0	2	74	76	0	0	4	1
trade	deficit	0	1	75	76	0	1	4	1
man	son	75	0	0	75	0	0	2	1
property	tax	0	0	75	75	0	0	4	1
savings	account	0	1	74	75	0	0	4	1
screen	play	75	0	0	75	0	0	2	1
video	game	0	11	64	75	0	0	4	1
community	service	0	4	70	74	0	0	5	0
god	father	74	0	0	74	0	0	3	1

Appendix B, Table 1 (cont.)

W1	W2	CLOSED	HYPHEN	OPEN	TOTAL	VSEQ	doubleC	SYLL	CS
hand	writing	74	0	0	74	0	0	3	1
home	town	55	0	19	74	0	0	2	0
web	page	2	6	65	73	0	0	2	1
baby	boomer	0	3	69	72	0	0	4	1
ball	park	65	0	7	72	0	0	2	1
bill	board	72	0	0	72	0	0	2	1
life	sentence	0	1	71	72	0	0	3	0
privacy	policy	0	0	72	72	0	0	5	1
baby	sitter	37	2	32	71	0	0	4	1
league	baseball	0	1	70	71	0	0	3	1
merger	guideline	0	0	71	71	0	0	4	0
mile	stone	71	0	0	71	0	0	2	1
play	boy	71	0	0	71	0	0	2	1
star	telegram	0	70	1	71	0	0	4	0
state	law	0	0	71	71	0	0	2	0
butter	fly	70	0	0	70	0	0	3	1
nether	land	70	0	0	70	0	0	3	1
play	ground	69	0	1	70	0	0	2	1
radio	station	0	0	70	70	0	0	5	1
self	defense	0	65	5	70	0	0	3	0
shot	gun	67	0	3	70	0	0	2	1
enforcement	agency	0	0	69	69	0	0	5	1
home	run	1	0	68	69	0	0	2	0
stair	case	69	0	0	69	0	0	2	1
story	line	5	0	64	69	0	0	3	1
sun	light	69	0	0	69	0	0	2	1
war	time	67	1	1	69	0	0	2	1
band	width	68	0	0	68	0	0	2	1
cook	book	68	0	0	68	0	0	2	1
hip	hop	0	65	3	68	0	0	2	1
house	speaker	0	0	68	68	0	1	3	0
safe	guard	68	0	0	68	0	0	2	1
tax	break	0	0	68	68	0	0	2	1
war	crime	2	10	56	68	0	0	2	1
breakfast	table	0	0	67	67	0	1	4	1
card	board	67	0	0	67	0	0	2	1
country	music	0	4	63	67	0	0	4	0
mail	box	62	0	5	67	0	0	2	1
main	land	67	0	0	67	0	0	2	1
movie	star	0	10	57	67	0	0	2	1
state	income	0	0	67	67	1	0	3	0
time	table	67	0	0	67	0	0	3	1
afl	cio	0	66	0	66	0	0	5	0
master	card	60	0	6	66	0	0	3	1
policy	maker	25	24	17	66	0	0	5	1
saturday	morning	0	1	65	66	0	0	5	0

Appendix B, Table 1 (cont.)

W1	W2	CLOSED	HYPHEN	OPEN	TOTAL	VSEQ	doubleC	SYLL	CS
screen	writer	66	0	0	66	0	0	3	1
tax	rate	0	1	65	66	0	0	2	1
cover	package	0	0	65	65	0	0	4	1
front	runner	7	58	0	65	0	0	3	1
grass	root	32	25	8	65	0	0	2	0
interest	group	0	1	64	65	0	0	3	1
nick	name	65	0	0	65	0	0	2	1
security	council	0	0	65	65	0	0	5	1
software	company	0	1	64	65	0	0	5	1
sun	beam	65	0	0	65	0	0	2	1
friday	night	0	0	64	64	0	0	3	0
letter	man	64	0	0	64	0	0	3	1
peace	talk	0	0	64	64	0	0	2	1
price	increase	0	0	64	64	1	0	3	0
tax	credit	0	0	64	64	0	0	3	1
volley	ball	62	0	2	64	0	0	3	1
eye	brow	63	0	0	63	0	0	2	1
sea	side	63	0	0	63	0	0	2	1
soap	opera	0	5	58	63	0	0	3	1
belt	way	62	0	0	62	0	0	2	1
center	piece	62	0	0	62	0	0	3	1
hewlett	packard	0	62	0	62	0	0	4	0
online	service	0	0	62	62	0	0	4	0
type	writer	62	0	0	62	0	0	3	1
work	station	61	0	1	62	0	0	3	1
auto	maker	55	2	4	61	0	0	4	1
half	hour	0	1	60	61	0	0	2	0
night	club	57	0	4	61	0	0	2	1
red	skin	61	0	0	61	0	0	2	1
sunday	morning	0	11	50	61	0	0	4	0
wall	paper	57	0	4	61	0	0	3	1
wednesday	morning	0	0	61	61	0	0	4	0
blood	pressure	0	2	58	60	0	0	3	1
monopoly	power	0	0	60	60	0	0	5	1
oil	company	0	0	60	60	0	0	4	1
science	fiction	0	9	51	60	0	0	4	0
thursday	night	0	1	59	60	0	0	3	0
west	coast	0	0	60	60	0	0	2	0
winter	time	18	0	42	60	0	0	3	1
child	abuse	0	5	54	59	0	0	3	1
competition	law	0	0	59	59	0	0	5	1
credit	union	0	0	59	59	0	0	4	1
dot	com	0	57	2	59	0	0	2	1
hotel	room	0	1	58	59	0	0	3	1
nobel	prize	0	0	59	59	0	0	3	0
operating	system	0	2	57	59	0	0	5	1

Appendix B, Table 1 (cont.)

W1	W2	CLOSED	HYPHEN	OPEN	TOTAL	VSEQ	doubleC	SYLL	CS
seat	belt	7	4	48	59	0	0	2	1
story	telling	59	0	0	59	0	0	4	1
task	force	1	1	57	59	0	0	2	1
battle	field	57	1	0	58	0	0	3	1
campaign	reform	0	14	44	58	0	0	4	0
child	rearing	0	23	35	58	0	0	3	1
christmas	time	7	0	51	58	0	0	3	1
consent	agreement	0	0	58	58	0	0	5	1
french	man	58	0	0	58	0	0	2	1
government	agency	0	0	58	58	0	0	5	0
hall	mark	58	0	0	58	0	0	2	1
news	stand	58	0	0	58	0	1	2	1
night	life	53	0	5	58	0	0	2	1
price	fixing	0	42	16	58	0	0	3	1
star	trek	0	0	58	58	0	0	2	1
base	line	49	1	7	57	0	0	2	1
finger	print	57	0	0	57	0	0	3	1
hormone	replacement	0	2	55	57	0	0	5	1
missile	defense	0	6	51	57	0	0	4	1
news	media	0	0	57	57	0	0	4	1
paper	back	55	0	2	57	0	0	3	1
death	sentence	0	2	54	56	0	0	3	1
house	wife	56	0	0	56	0	0	2	1
main	stream	56	0	0	56	0	0	2	1
party	line	3	15	38	56	0	0	3	1
self	interest	0	55	1	56	0	0	3	0
spread	sheet	55	1	0	56	0	0	2	1
supply	side	0	52	4	56	0	0	3	1
tax	reform	0	3	53	56	0	0	3	1
war	head	56	0	0	56	0	0	2	1
water	fall	56	0	0	56	0	0	3	1
birth	place	55	0	0	55	0	0	2	1
health	plan	0	0	55	55	0	0	2	1
rod	man	55	0	0	55	0	0	2	1
snap	shot	55	0	0	55	0	0	2	1
spin	network	0	0	55	55	0	1	3	1
tax	revenue	0	0	55	55	0	0	4	1
bull	pen	53	0	1	54	0	0	2	1
fbi	agent	0	0	54	54	1	0	5	1
fire	wall	52	0	2	54	0	0	3	1
hen	man	54	0	0	54	0	0	2	1
impeachment	trial	0	0	54	54	0	1	5	1
internet	access	0	3	51	54	0	0	5	1
line	item	0	42	12	54	1	0	3	1
merger	enforcement	0	0	54	54	0	0	5	0
phone	company	0	1	53	54	0	0	4	1

Appendix B, Table 1 (cont.)

W1	W2	CLOSED	HYPHEN	OPEN	TOTAL	VSEQ	doubleC	SYLL	CS
race	track	51	0	3	54	0	0	2	1
back	bone	53	0	0	53	0	0	2	1
cable	tv	0	9	44	53	0	0	4	0
labor	party	0	0	53	53	0	0	4	1
media	company	0	0	53	53	0	0	5	1
news	magazine	25	0	28	53	0	0	4	1
sky	scraper	53	0	0	53	0	0	3	1
tax	increase	0	0	53	53	0	0	3	1
west	side	6	0	47	53	0	0	2	0
world	class	0	49	4	53	0	0	2	0
check	book	46	0	6	52	0	0	2	1
computer	industry	0	2	50	52	0	0	5	1
consumer	welfare	0	0	52	52	0	0	5	0
fairy	tale	9	15	28	52	0	0	3	1
hill	top	52	0	0	52	0	0	2	1
iran	contra	0	52	0	52	0	0	4	0
peter	son	52	0	0	52	0	0	3	1
power	house	52	0	0	52	0	0	3	1
stein	berg	52	0	0	52	0	0	2	1
budget	deal	0	0	51	51	0	0	3	1
budget	deficit	0	1	50	51	0	0	5	0
computer	science	0	1	50	51	0	0	5	0
double	day	51	0	0	51	0	0	3	1
glass	man	51	0	0	51	0	0	2	1
internet	service	0	0	51	51	0	0	5	1
state	government	0	0	51	51	0	0	4	0
whistle	blower	20	31	0	51	0	0	4	1
approval	rating	0	0	50	50	0	0	5	1
chap	man	50	0	0	50	0	0	2	1
computer	system	0	0	50	50	0	0	5	1
day	light	49	0	1	50	0	0	2	1
drug	dealer	0	0	50	50	0	0	3	1
fore	front	50	0	0	50	0	0	2	1
graduate	school	0	3	47	50	0	0	4	1
health	problem	0	0	50	50	0	0	3	1
honey	moon	50	0	0	50	0	0	3	1
law	firm	0	0	50	50	0	0	2	1
sci	fi	0	50	0	50	0	0	2	1
show	case	50	0	0	50	0	0	2	1
weapons	inspection	0	15	35	50	0	0	5	1
bank	account	0	1	48	49	0	0	3	1
bat	man	49	0	0	49	0	0	2	1
bell	curve	0	2	47	49	0	0	2	1
class	action	0	32	17	49	0	0	3	0
configuration	space	0	0	49	49	0	0	5	1
defense	secretary	0	0	49	49	0	1	5	1

Appendix B, Table 1 (cont.)

W1	W2	CLOSED	HYPHEN	OPEN	TOTAL	VSEQ	doubleC	SYLL	CS
farm	house	46	0	3	49	0	0	2	1
half	year	0	1	48	49	0	0	2	0
news	letter	44	0	5	49	0	0	3	1
police	department	0	1	48	49	0	0	5	1
slap	stick	49	0	0	49	0	0	2	1
sound	track	47	0	2	49	0	0	2	1
sunday	night	0	1	48	49	0	0	3	0
bench	mark	48	0	0	48	0	0	2	1
blue	print	48	0	0	48	0	0	2	1
camp	site	37	0	11	48	0	0	2	1
decision	making	13	25	10	48	0	0	5	1
estate	tax	0	5	43	48	0	0	3	1
gene	therapy	0	1	47	48	0	0	4	1
ice	cream	0	13	35	48	0	0	2	1
ice	storm	0	0	48	48	0	1	2	1
pop	culture	0	10	38	48	0	0	3	1
sea	gram	48	0	0	48	0	0	2	1
senate	race	0	1	47	48	0	0	3	1
tooth	paste	48	0	0	48	0	0	2	0
word	processor	0	1	47	48	0	0	4	1
air	traffic	0	6	41	47	0	0	3	1
ball	game	29	0	18	47	0	0	2	1
body	guard	47	0	0	47	0	0	3	1
business	leader	0	0	47	47	0	0	4	1
cash	flow	0	1	46	47	0	0	2	1
drug	problem	0	0	47	47	0	0	3	1
horse	back	47	0	0	47	0	0	2	0
man	kind	47	0	0	47	0	0	2	1
merger	case	0	0	47	47	0	0	3	1
school	year	0	0	47	47	0	0	2	1
subject	matter	0	0	47	47	0	0	4	0
friday	morning	0	0	46	46	0	0	4	0
plastic	bag	0	0	46	46	0	0	3	0
speed	way	46	0	0	46	0	0	2	1
band	wagon	41	0	4	45	0	0	3	1
business	unit	0	2	43	45	0	0	4	1
capital	gain	0	3	42	45	0	0	4	0
coast	guard	0	0	45	45	0	0	2	1
corner	stone	45	0	0	45	0	0	3	1
defense	lawyer	0	0	45	45	0	0	3	1
football	game	0	0	45	45	0	0	3	1
gas	mileage	0	0	45	45	0	0	3	1
gold	water	45	0	0	45	0	0	3	1
lemon	ade	45	0	0	45	0	0	3	0
mccain	feingold	0	45	0	45	0	0	4	0
merger	review	0	1	44	45	0	1	4	1

Appendix B, Table 1 (cont.)

W1	W2	CLOSED	HYPHEN	OPEN	TOTAL	VSEQ	doubleC	SYLL	CS
run	way	45	0	0	45	0	0	2	1
sail	boat	39	0	6	45	0	0	2	1
administration	official	0	0	44	44	0	0	5	1
base	man	44	0	0	44	0	0	2	1
campaign	fund	0	5	39	44	0	0	3	0
election	year	0	15	29	44	0	0	4	1
foot	print	44	0	0	44	0	0	2	1
hatch	waxman	0	44	0	44	0	0	3	0
love	affair	0	1	43	44	1	0	3	1
post	office	0	0	44	44	0	0	3	1
show	room	44	0	0	44	0	0	2	1
side	line	44	0	0	44	0	0	2	1
sky	line	44	0	0	44	0	0	2	1
well	stone	44	0	0	44	0	0	2	1
wood	land	44	0	0	44	0	0	2	1
yellow	stone	44	0	0	44	0	0	3	1
air	wave	43	0	0	43	0	0	2	1
art	work	37	0	6	43	0	0	2	1
budget	surplus	0	0	43	43	0	0	4	0
emergency	room	0	4	39	43	0	0	5	1
fire	place	42	0	1	43	0	0	3	1
foot	step	43	0	0	43	0	0	2	1
lip	stick	43	0	0	43	0	0	2	1
middle	man	42	0	1	43	0	0	3	1
news	group	42	0	1	43	0	0	2	1
sex	life	0	0	43	43	0	0	2	1
state	park	0	0	43	43	0	0	2	0
stock	pile	43	0	0	43	0	0	2	1
training	camp	0	2	41	43	0	0	3	1
use	net	43	0	0	43	0	0	2	1
world	economy	0	0	43	43	0	0	5	0
age	group	0	0	42	42	0	0	2	1
life	insurance	0	2	40	42	1	0	4	1
man	slaughter	42	0	0	42	0	0	3	1
opinion	poll	0	0	42	42	0	0	4	1
pass	word	42	0	0	42	0	0	2	1
profit	margin	0	1	41	42	0	0	4	1
school	age	0	23	19	42	0	0	2	1
staff	member	0	0	42	42	0	0	3	1
war	criminal	0	1	41	42	0	0	4	1
aluminum	can	0	0	41	41	0	0	5	0
background	check	0	0	41	41	0	0	3	1
bow	man	41	0	0	41	0	0	2	1
capitol	hill	0	0	41	41	0	0	4	0
clinton	gore	0	41	0	41	0	0	3	0
growth	rate	0	1	40	41	0	0	2	1

Appendix B, Table 1 (cont.)

W1	W2	CLOSED	HYPHEN	OPEN	TOTAL	VSEQ	doubleC	SYLL	CS
investment	bank	0	0	41	41	0	0	4	1
lunch	time	28	1	12	41	0	0	2	1
senate	republican	0	0	41	41	0	0	5	1
state	university	0	0	41	41	1	0	5	1
trade	agreement	0	0	41	41	1	0	4	1
wind	mill	41	0	0	41	0	0	2	1
air	liner	40	0	0	40	0	0	3	1
ballot	box	0	2	38	40	0	0	3	1
bed	time	38	0	2	40	0	0	2	1
consumer	report	0	0	40	40	0	1	5	0
defense	department	0	0	40	40	0	0	5	1
dough	nut	40	0	0	40	0	0	2	1
drug	abuse	0	2	38	40	0	0	3	1
egypt	air	40	0	0	40	0	0	3	0
folk	lore	40	0	0	40	0	0	2	1
horror	story	0	0	40	40	0	0	4	1
house	majority	0	0	40	40	0	0	5	0
independence	day	0	0	40	40	0	0	4	1
internet	company	0	0	40	40	0	0	5	1
pine	apple	40	0	0	40	1	0	3	1
pocket	book	31	0	9	40	0	0	3	1
post	card	39	0	1	40	0	0	2	1
privacy	issue	0	0	40	40	1	0	5	1
show	time	40	0	0	40	0	0	2	1
space	time	10	30	0	40	0	0	2	0
sweat	shop	39	1	0	40	0	0	2	1
tax	dollar	0	0	40	40	0	0	3	1
tobacco	settlement	0	3	37	40	0	0	5	1
touch	down	40	0	0	40	0	0	2	1
treasury	secretary	0	0	40	40	0	0	5	1
tv	station	0	0	40	40	0	0	4	1
world	leader	0	0	40	40	0	0	3	0
abortion	right	0	15	24	39	0	0	4	1
blood	clot	0	0	39	39	0	0	2	1
bull	market	0	3	36	39	0	0	3	1
cable	television	0	5	34	39	0	0	5	0
chat	room	4	2	33	39	0	0	2	1
defense	contractor	0	1	38	39	0	0	5	1
education	system	0	0	39	39	0	0	5	1
eye	ball	38	0	1	39	0	0	2	1
eye	witness	39	0	0	39	0	0	3	0
government	regulation	0	0	39	39	0	0	5	0
gun	fire	39	0	0	39	0	0	2	1
jury	duty	0	0	39	39	0	0	4	1
labor	cost	0	1	38	39	0	0	3	1
land	slide	39	0	0	39	0	0	2	1

Appendix B, Table 1 (cont.)

W1	W2	CLOSED	HYPHEN	OPEN	TOTAL	VSEQ	doubleC	SYLL	CS
life	span	7	2	30	39	0	0	2	1
night	line	39	0	0	39	0	0	2	1
race	relation	0	2	37	39	0	0	4	1
resale	price	0	0	39	39	0	0	3	1
river	side	39	0	0	39	0	0	3	1
security	number	0	0	39	39	0	0	5	1
space	station	0	0	39	39	0	1	3	1
star	buck	39	0	0	39	0	0	2	1
track	record	0	0	39	39	0	0	3	1
tv	news	0	1	38	39	0	0	3	0
wood	stock	39	0	0	39	0	0	2	1
black	mail	38	0	0	38	0	0	2	1
body	part	0	0	38	38	0	0	3	1
bounty	hunter	0	0	38	38	0	0	4	1
cable	company	0	0	38	38	0	0	5	1
computer	program	0	0	38	38	0	0	5	0
district	attorney	0	0	38	38	0	0	5	0
domain	name	0	1	37	38	0	1	3	1
gas	station	0	1	37	38	0	1	3	1
hall	way	38	0	0	38	0	0	2	1
mail	delivery	0	1	37	38	0	0	5	1
parent	company	0	0	38	38	0	0	5	1
percentage	point	0	0	38	38	0	0	4	1
phantom	menace	0	0	38	38	0	1	4	0
price	tag	2	2	34	38	0	0	2	1
stone	man	38	0	0	38	0	0	2	1
term	limit	0	1	37	38	0	0	3	1
town	house	31	0	7	38	0	0	2	1
world	view	26	1	11	38	0	0	2	1
air	space	34	0	3	37	0	0	2	1
arms	race	0	0	37	37	0	0	2	1
back	pack	37	0	0	37	0	0	2	1
baseball	game	0	0	37	37	0	0	4	1
bed	rock	37	0	0	37	0	0	2	1
book	club	0	1	36	37	0	0	2	1
book	review	0	4	33	37	0	0	3	1
bristol	myer	0	37	0	37	0	0	4	0
cheer	leader	37	0	0	37	0	0	3	1
consumer	education	0	1	36	37	0	0	5	0
consumer	privacy	0	1	36	37	0	0	5	0
drug	store	21	0	16	37	0	0	2	1
drug	user	0	0	37	37	0	0	3	1
drug	war	0	2	35	37	0	0	2	1
house	manager	0	0	37	37	0	0	4	1
investment	banker	0	0	37	37	0	0	5	0
market	definition	0	1	36	37	0	0	5	1

Appendix B, Table 1 (cont.)

W1	W2	CLOSED	HYPHEN	OPEN	TOTAL	VSEQ	doubleC	SYLL	CS
movie	goer	35	0	2	37	0	0	2	1
net	working	37	0	0	37	0	0	3	1
news	organization	0	0	37	37	0	0	5	1
plea	bargain	0	5	32	37	0	0	3	1
red	wood	37	0	0	37	0	0	2	1
search	engine	0	0	37	37	0	0	3	1
security	adviser	0	0	37	37	1	0	5	1
space	shuttle	0	0	37	37	0	0	3	1
suicide	bombing	0	3	34	37	0	0	5	0
telephone	number	0	0	37	37	0	1	5	1
trust	fund	0	1	36	37	0	0	2	1
wednesday	night	0	1	36	37	0	0	3	0
week	day	36	0	1	37	0	0	2	1
white	head	37	0	0	37	0	0	2	1
work	week	6	0	31	37	0	0	2	1
academy	award	0	0	36	36	1	0	5	0
air	bus	36	0	0	36	0	0	2	1
air	conditioner	0	5	31	36	0	0	5	0
apparel	industry	0	0	36	36	0	0	5	1
balance	sheet	0	6	30	36	0	0	3	1
business	person	15	0	21	36	0	0	4	1
city	council	0	0	36	36	0	0	4	0
court	case	0	0	36	36	0	0	2	1
cross	section	0	0	36	36	0	1	3	1
earnings	report	0	0	36	36	0	0	4	1
east	side	4	1	31	36	0	0	2	0
farm	land	36	0	0	36	0	0	2	1
government	intervention	0	0	36	36	0	0	5	0
hind	sight	36	0	0	36	0	0	2	1
labor	secretary	0	0	36	36	0	0	5	1
labor	union	0	3	33	36	0	0	4	1
product	market	0	0	36	36	0	0	4	1
quantum	mechanics	0	0	36	36	0	1	5	0
rate	hike	0	0	36	36	0	0	2	1
road	side	34	0	2	36	0	0	2	1
rock	star	0	2	34	36	0	0	2	1
school	choice	0	2	34	36	0	0	2	0
sports	car	0	2	34	36	0	0	2	1
time	period	0	1	35	36	0	0	3	1
tribune	herald	0	36	0	36	0	0	4	0
welfare	roll	0	0	36	36	0	1	3	1
bear	market	0	0	35	35	0	0	3	1
bombing	campaign	0	0	35	35	0	0	4	1
boot	leg	35	0	0	35	0	0	2	1
camp	ground	30	0	5	35	0	0	2	1
college	football	0	0	35	35	0	0	4	0

Appendix B, Table 1 (cont.)

W1	W2	CLOSED	HYPHEN	OPEN	TOTAL	VSEQ	doubleC	SYLL	CS
committee	chairman	0	0	35	35	0	0	4	0
curb	side	33	0	2	35	0	0	2	1
fair	way	35	0	0	35	0	0	2	1
fire	man	35	0	0	35	0	0	3	1
flash	back	35	0	0	35	0	0	2	1
gamma	ray	0	26	9	35	0	0	3	1
health	initiative	0	0	35	35	0	0	5	1
hitch	cock	35	0	0	35	0	0	2	1
internet	stock	0	1	34	35	0	0	4	1
lawn	mower	5	0	30	35	0	0	3	1
majority	leader	0	0	35	35	0	0	5	0
oil	price	0	1	34	35	0	0	2	1
party	leader	0	0	35	35	0	0	4	0
school	board	1	3	31	35	0	0	2	1
slaughter	house	30	5	0	35	0	0	3	1
space	ship	34	0	1	35	0	0	2	1
suicide	bomber	1	0	34	35	0	0	5	0
time	frame	4	1	30	35	0	0	2	1
unemployment	rate	0	0	35	35	0	0	5	1
weight	loss	0	15	20	35	0	0	2	1
welfare	state	0	6	29	35	0	0	3	1

*”news story” was among the 35+ group, but not included in the regression analysis due to a programming error.

Table 2: Regression Model predictions for 707 compounds (sorted by frequency)

W1	W2	TOTAL	Proportion Closed	Proportion Hyphen	Proportion Open	PRED Closed	PRED Hyphen	PRED Open
news	paper	2509	.9988	00	12	.61769	-.07388	.48826
net	work	2009	1 00	00	00	1.15395	.04205	-.08231
sun	day	1873	1 00	00	00	.88748	.07375	.07968
week	end	1353	.9985	00	15	.54133	.04006	.45833
chair	man	1284	1 00	00	00	1.04147	.07665	-.10706
base	ball	1175	1 00	00	00	.98651	-.08272	.04969
head	line	995	.9990	00	10	1.02536	.06228	-.07539
news	week	966	.9990	00	10	.63448	-.02919	.40894
john	son	936	1 00	00	00	1 863	-.01577	-.05883
chatter	box	930	1 00	00	00	.91569	.04879	.05226
foot	ball	884	1 00	00	00	1.09773	-.10484	-.04993
front	page	840	.0131	.9738	.0131	.34665	.32163	.32350
holly	wood	808	1 00	00	00	1.11653	-.01613	-.18237
air	line	771	.9948	00	52	.85922	.04643	.10821
back	ground	768	1 00	00	00	1.02545	-.04365	.03116
guide	line	764	1 00	00	00	.84565	.10059	.05637
health	care	739	.0947	.2057	.6996	.36692	.02960	.68248
credit	card	686	00	.0612	.9388	.32703	-.02751	.71782
jack	son	673	1 00	00	00	1.22198	.04216	-.34769
web	site	670	.2224	90	.7687	.66479	-.02930	.40313
air	port	649	.9985	15	00	.92987	.03756	.05183
cover	story	615	00	00	1 00	.22594	-.08051	.85598
law	suit	538	.9944	00	56	.60989	-.03762	.44053
basket	ball	499	1 00	00	00	.82200	-.04724	.14274
stock	market	461	00	.0564	.9436	.35169	.05181	.66042
bed	room	431	.9977	00	23	.99850	- 611	.03177
world	war	418	00	00	1 00	.09491	.10195	.74834
cow	boy	391	1 00	00	00	.88047	.01355	.14933
tax	payer	387	.9664	26	.0310	.15734	.23760	.68612
death	penalty	369	00	.0434	.9566	.05969	.10074	.84026
market	place	369	.9702	00	.0298	.35791	-.05104	.68903
frame	work	355	.9972	00	28	.83977	.02546	.18056
law	enforcement	343	00	.0845	.9155	.19756	-.06312	.79341
air	craft	336	1 00	00	00	.94093	.01421	.03883
house	hold	332	1 00	00	00	.81377	-.02391	.22577
interest	rate	330	00	.0697	.9303	.15216	.02117	.84848
life	time	330	.9727	91	.0182	.64181	.10006	.29004
green	span	329	1 00	00	00	.86122	- 117	.21750
data	base	328	.9146	30	.0823	.37659	.03556	.57502
op	ed	318	00	1 00	00	.29747	.37284	.36146
share	holder	312	1 00	00	00	.75658	.02696	.23563
break	fast	311	1 00	00	00	.67492	.20212	.16677
birth	day	305	1 00	00	00	.67519	-.01483	.31389

Table 2 (cont.)

W1	W2	TOTAL	Proportion Closed	Proportion Hyphen	Proportion Open	PRED Closed	PRED Hyphen	PRED Open
head	quarter	286	1 00	00	00	.72013	.11891	.17196
home	land	285	1 00	00	00	1.12681	-.14399	-.05502
life	style	282	.9539	35	.0426	.45317	.08685	-.20616
work	shop	277	.9964	00	36	.82444	-.04017	.24947
fly	trap	276	1 00	00	00	.74439	.05499	.21782
girl	friend	275	.9782	00	.0218	.55740	196	.44362
bath	room	269	1 00	00	00	.89268	.01260	.11231
air	plane	264	1 00	00	00	.66872	.02882	.32821
class	room	260	.9962	00	38	.65824	.02921	.32711
quarter	back	257	1 00	00	00	.84966	.04133	.05036
school	system	256	00	00	1 00	.30658	.05105	.75642
work	place	256	.8242	00	.1758	.80249	-.09771	.28924
north	west	250	.9880	80	40	.69897	.22829	.06419
law	maker	249	1 00	00	00	.70469	.02371	.26385
culture	box	241	1 00	00	00	.54254	-.03267	.47661
pipe	line	241	.9959	00	41	.81041	.06221	.11230
robin	son	239	1 00	00	00	.99201	457	-.10023
scot	land	238	1 00	00	00	1.02677	.02549	-.14117
health	insurance	236	00	.0297	.9703	-.06184	.11961	.91104
income	tax	229	00	.0699	.9301	-.11024	.03319	1.03620
air	force	228	.0132	44	.9825	.63239	- 535	.39837
business	man	228	.9781	00	.0219	.65558	310	.30842
earth	quake	224	1 00	00	00	.68558	.16284	.19230
day	time	217	.9677	00	.0323	.59635	.16991	.25833
gun	control	217	00	.0369	.9631	.18197	.03131	.73004
stock	price	217	00	.0230	.9770	.24536	.02382	.72126
child	care	212	.0377	.1698	.7925	.23825	.12917	.72521
bloom	berg	210	1 00	00	00	1.02849	.11351	-.23619
water	gate	210	1 00	00	00	.97053	-.03459	.05684
pay	roll	204	1 00	00	00	.67693	-.02423	.36731
rail	road	203	.9901	00	99	.56217	.02375	.43000
day	care	197	.0102	.1929	.7970	.46076	.10934	.52357
tax	cut	197	00	.0203	.9797	.39226	-.03140	.64166
phone	call	193	00	52	.9948	.31297	- 673	.68899
north	east	190	.9789	53	.0158	.67606	.15180	.12998
master	piece	189	1 00	00	00	.62121	- 935	.40399
sand	stone	189	1 00	00	00	1.20410	-.09692	-.16255
ash	croft	186	1 00	00	00	.94759	.16013	-.11711
cell	phone	179	.1285	.1061	.7654	.51067	- 931	.52122
self	regulation	179	00	.9888	.0112	627	.78901	.26344
war	fare	178	1 00	00	00	.74377	- 874	.27550
south	east	177	.9887	.0113	00	.73917	.09243	.12464
press	conference	175	00	57	.9943	.26448	-.05409	.76592
book	store	173	.9364	00	.0636	.57884	494	.42994
text	book	173	.9769	00	.0231	.69771	.04460	.29320
boy	friend	165	.9818	00	.0182	.63166	-.06456	.43563

Table 2 (cont.)

W1	W2	TOTAL	Proportion Closed	Proportion Hyphen	Proportion Open	PRED Closed	PRED Hyphen	PRED Open
campaign	finance	162	00	.1543	.8457	-.14402	.19159	.89234
video	tape	161	.9193	00	.0807	.33427	.02275	.63196
congress	man	159	1 00	00	00	.84635	.05932	.05537
consumer	protection	155	00	65	.9935	-.13950	.11696	1.02365
family	member	155	00	00	1 00	.01499	-.03883	1 467
block	buster	154	.9935	65	00	.72179	.16318	.12274
hand	gun	154	.8896	65	.1039	.78021	.02930	.22096
note	book	154	1 00	00	00	.81100	.05303	.16832
pea	nut	153	1 00	00	00	.89633	.05413	.06223
police	man	153	.9935	00	65	.70387	830	.25163
daytime	phone	151	00	00	1 00	.30059	.15506	.56551
land	mark	151	1 00	00	00	.92146	-.03592	.12352
stock	option	148	00	.1351	.8649	.19625	329	.74584
white	water	147	1 00	00	00	.94122	- 105	.06186
back	lash	145	1 00	00	00	1.04553	.11328	-.13450
grocery	store	145	00	69	.9931	.40327	.01713	.60438
team	mate	144	1 00	00	00	.71996	- 202	.40733
oak	land	140	1 00	00	00	1.10122	-.12575	-.05590
lime	stone	139	.9928	72	00	.98155	-.02357	-.02470
sales	man	139	1 00	00	00	.91182	- 798	.06879
information	technology	137	00	.0657	.9343	-.25371	.10977	1.08841
lap	top	136	1 00	00	00	1.01573	- 992	-.02433
court	room	135	.9778	00	.0222	.69638	-.04236	.35983
back	yard	134	.9925	00	75	1.05460	-.01911	-.10012
pay	check	134	1 00	00	00	.56855	-.04979	.47205
school	district	134	00	75	.9925	.25102	058	.71967
sea	food	134	.9851	00	.0149	.75802	-.01026	.25036
south	west	133	.9925	00	75	.77783	.17590	.02104
side	walk	129	.9922	00	78	.83588	- 860	.19265
tobacco	company	129	00	.0310	.9690	-.15300	.06398	1.14707
fisher	man	128	1 00	00	00	.85473	.12025	-.01751
ware	house	128	1 00	00	00	.95607	.10940	-.02870
insurance	company	127	00	00	1 00	-.09154	.04985	1.10959
wrong	doing	126	1 00	00	00	.52138	.18955	.27302
heart	disease	125	00	.0240	.9760	.29664	.01123	.66774
police	officer	125	00	00	1 00	.06584	-.06304	.95400
air	strike	124	.8468	00	.1532	.42708	.02439	.51622
film	maker	124	.9839	81	81	.73214	.01166	.37196
rail	way	124	1 00	00	00	.95769	-.10239	.09109
thanks	giving	124	1 00	00	00	.36996	.31103	.28343
room	mate	123	.9919	00	81	.72251	.01498	.38418
air	conditionin	122	00	.3033	.6967	.44868	.13279	.49033
motor	cycle	122	1 00	00	00	.33325	- 024	.64306
air	way	121	1 00	00	00	1.06385	-.09630	-.01578
fire	fighter	121	.9917	00	83	.63020	-.01575	.36543
news	hour	121	.9752	00	.0248	.40116	-.02224	.60509

Table 2 (cont.)

W1	W2	TOTAL	Proportion Closed	Proportion Hyphen	Proportion Open	PRED Closed	PRED Hyphen	PRED Open
court	yard	120	1 00	00	00	.78078	-.06615	.26125
monday	morning	120	00	83	.9917	-.24165	.14259	1.07180
wild	life	120	1 00	00	00	.92608	.09608	.02635
copy	right	119	1 00	00	00	.39834	.07664	.54015
country	side	119	1 00	00	00	.59463	.04268	.42983
death	row	119	00	.1176	.8824	.42843	.17882	.40867
home	work	119	.9832	00	.0168	.93499	-.02729	.13364
capital	punishment	118	00	.0254	.9746	-.07689	.21531	.85525
breast	cancer	117	00	.0513	.9487	.17375	.05617	.73081
welfare	reform	117	00	.1282	.8718	-.20302	.06786	1.24095
work	force	117	.5897	00	.4103	.53796	-.05536	.50735
fire	arm	116	.9569	00	.0431	.67248	.01962	.29319
view	point	116	.9741	00	.0259	.71103	.04885	.30984
wheel	chair	116	.9655	86	.0259	.58814	.12506	.27418
sun	set	115	.9826	00	.0174	.96457	.12210	-.07410
competit	policy	114	00	00	1 00	-.22844	.09560	1.10791
consent	order	114	00	00	1 00	-.04852	.05614	.95591
summer	time	114	.7105	00	.2895	.45272	.10503	.45205
trade	mark	113	1 00	00	00	.51077	-.03954	.53583
sales	tax	112	00	.0446	.9554	.20369	-.02204	.79775
market	power	111	00	90	.9910	.26804	326	.73830
gold	berg	110	1 00	00	00	1.11151	-.02324	-.18231
gold	man	110	1 00	00	00	1.07622	- 455	-.10078
news	day	110	.8727	00	.1273	.81641	-.10754	.26491
stand	point	110	.9909	00	91	.84736	.12771	.08667
heart	attack	109	00	00	1 00	.24160	-.05546	.74830
saturday	night	109	00	.0275	.9725	-.03048	.09648	.88430
district	court	107	00	00	1 00	.15727	.06895	.74140
judgment	column	107	00	00	1 00	-.01898	.09105	.82593
state	department	107	00	00	1 00	-.06436	-.07999	1.08319
cock	tail	106	1 00	00	00	1.08912	.11710	-.22635
house	republican	106	00	00	1 00	.28573	-.03905	.75583
capital	gang	105	00	00	1 00	.02658	.16450	.78074
class	mate	104	1 00	00	00	.74810	.02322	.22271
court	house	104	1 00	00	00	.77107	-.01520	.28706
dinner	party	104	00	00	1 00	.01563	-.02111	.94313
green	house	104	1 00	00	00	1.11889	-.07046	133
loop	hole	104	.9808	00	.0192	.80891	.06767	.14425
coca	cola	103	00	1 00	00	.12827	.40344	.45812
family	reunion	103	00	00	1 00	-.16527	.22153	.94825
market	share	103	.0194	.0388	.9417	.24742	- 621	.72439
air	pollution	102	00	00	1 00	.30185	.04453	.61134
gate	way	102	1 00	00	00	1.17110	-.07950	-.14729
land	lord	102	1 00	00	00	.78674	.05363	.20177
thai	land	102	1 00	00	00	.96278	-.06654	.01793
coast	line	101	.9901	00	99	.70762	.02339	.25479

Table 2 (cont.)

W1	W2	TOTAL	Proportion Closed	Proportion Hyphen	Proportion Open	PRED Closed	PRED Hyphen	PRED Open
east	coast	100	00	.0100	.9900	.37669	.12198	.46636
foot	note	100	1 00	00	00	.78176	049	.24861
table	spoon	100	1 00	00	00	.65722	.11109	.23159
key	board	99	1 00	00	00	1 520	-.04511	.08924
desk	top	98	.9898	00	.0102	.79270	.01603	.16757
enforcement	action	98	00	00	1 00	-.05669	.02970	1.08150
government	official	98	00	00	1 00	-.31058	323	1.23870
peace	process	98	00	00	1 00	.11921	-.03642	.88228
fire	work	97	1 00	00	00	.94392	- 697	.09403
golf	course	97	00	.0103	.9897	.29515	-.02985	.72227
play	wright	97	1 00	00	00	.85842	.12923	.06836
stock	holder	97	.9897	00	.0103	.67839	.01224	.31607
brand	name	96	00	.4167	.5833	.43296	.05713	.53429
college	student	96	00	00	1 00	-.21014	.01551	1.13822
main	frame	96	1 00	00	00	.72447	.15052	.13422
straw	berry	96	1 00	00	00	.65591	.07280	.24690
vine	yard	96	1 00	00	00	1.02174	.04603	-.10023
water	front	96	1 00	00	00	.86402	.02621	.15492
club	house	95	1 00	00	00	.67433	.04760	.31646
news	conference	95	00	00	1 00	.41679	-.07014	.63189
senate	majority	95	00	00	1 00	-.14503	.06624	1.03886
world	series	95	00	00	1 00	-.06872	.10856	.86398
side	bar	94	.9894	00	.0106	.81553	-.01008	.21628
fund	raiser	93	.0968	.8387	.0645	.40423	.20328	.43816
wave	length	92	.9783	.0109	.0109	.47212	.09567	.40774
back	drop	91	1 00	00	00	.98824	.01047	.04053
green	berg	91	1 00	00	00	1.24962	-.07505	-.26350
path	way	91	1 00	00	00	1.05366	-.05495	-.06016
press	release	90	00	.0111	.9889	.21477	.01576	.74931
department	store	89	00	.0674	.9326	.28401	.01641	.70736
ground	troop	89	00	.0112	.9888	.34396	.02881	.57291
sun	time	89	00	1 00	00	.84082	.14347	-.01439
tv	show	89	00	00	1 00	.14919	.01338	.84315
drug	use	88	00	.0114	.9886	.17214	.04003	.77356
paper	work	87	.9425	00	.0575	.75458	-.05219	.33178
robinson	patman	87	00	1 00	00	.29741	.52690	.27980
side	effect	87	00	.0805	.9195	.45487	-.04571	.56432
box	office	86	00	.3605	.6395	.15494	.08644	.72839
self	esteem	86	00	.9419	.0581	.14267	1.17012	-.13007
gun	man	85	1 00	00	00	.98405	-.01073	- 257
land	fill	85	1 00	00	00	.86604	.11699	.02794
spot	light	85	.9882	00	.0118	.80507	-.01495	.21249
bore	hole	84	1 00	00	00	.86208	.10703	.04706
cock	pit	84	1 00	00	00	.84361	.08434	.07321
front	pager	84	00	1 00	00	.20538	1.01624	-.07373
hill	side	84	1 00	00	00	1.04131	.03955	- 465

Table 2 (cont.)

W1	W2	TOTAL	Proportion Closed	Proportion Hyphen	Proportion Open	PRED Closed	PRED Hyphen	PRED Open
prescription	drug	84	00	.0357	.9643	-.12348	.10605	.92902
sun	shine	84	.9881	.0119	00	.96062	.23896	-.18867
air	bag	83	.1084	.0964	.7952	.76778	043	.27385
boy	scout	83	00	00	1 00	.38050	.06193	.55832
drug	company	83	00	.0241	.9759	-.12158	.05668	1.08498
homeland	security	83	00	.0723	.9277	-.08727	.21794	.80837
reform	party	83	00	00	1 00	-.15593	-.01547	1.10815
school	teacher	83	.3373	00	.6627	.20615	-.01406	.77465
terrorist	attack	83	00	00	1 00	-.03553	-.05426	1.06526
birth	control	82	00	.2561	.7439	.17871	.03702	.72061
court	decision	82	00	00	1 00	.07360	-.04851	.90054
drive	way	82	.9878	00	.0122	.96603	-.07544	.04977
food	stamp	82	00	.1829	.8171	.52364	.11074	.42559
home	page	82	.0610	00	.9390	.12645	.08083	.84438
monday	night	82	00	00	1 00	.01200	.10159	.84097
crime	rate	81	00	00	1 00	.12614	.10045	.80615
tea	spoon	81	1 00	00	00	.83958	.09175	.07351
tobacco	industry	81	00	.1111	.8889	-.21192	.04803	1.15316
executive	director	80	00	00	1 00	-.21011	.12985	1.07109
head	ache	80	1 00	00	00	1.02214	.09122	-.11654
patter	son	80	1 00	00	00	1.01560	.02327	-.15098
phone	number	80	00	00	1 00	.15759	-.01170	.99464
family	value	79	00	.1266	.8734	-.36160	.08494	1.17845
george	town	79	1 00	00	00	.92374	.10503	- 725
graduate	student	79	00	00	1 00	-.04333	.02785	1 465
home	owner	79	.8481	00	.1519	.53184	-.03545	.50896
power	plant	79	.0759	.0253	.8987	.49971	-.04671	.54510
role	model	79	00	.0127	.9873	.11844	.06036	.79029
news	story	78	00	00	1 00	.33532	-.07948	.85920
pass	port	78	1 00	00	00	.96632	.06397	-.07053
peace	keeping	78	.9744	.0256	00	.62410	.03426	.34080
drug	test	77	00	.0260	.9740	.28394	.03053	.70742
law	school	77	00	.0519	.9481	.35660	.02264	.61182
port	land	77	1 00	00	00	1.10787	-.13041	-.05906
campaign	contribution	76	00	.0263	.9737	-.14093	.12364	.94661
east	wood	76	1 00	00	00	1.04778	-.10233	-.06222
ground	water	76	.8684	.0658	.0658	.70550	479	.28223
property	right	76	00	.0263	.9737	.21822	.01163	.81086
trade	deficit	76	00	.0132	.9868	-.04703	.04118	1.12026
man	son	75	1 00	00	00	1.09709	-.05187	-.14715
property	tax	75	00	00	1 00	-.22521	.02131	1.16402
savings	account	75	00	.0133	.9867	-.02771	- 124	.98121
screen	play	75	1 00	00	00	.57602	.02220	.41774
communit	service	74	00	.0541	.9459	-.15120	.15434	1.01721
god	father	74	1 00	00	00	.54004	.06727	.36167
hand	writing	74	1 00	00	00	.90080	.10443	-.01260

Table 2 (cont.)

W1	W2	TOTAL	Proportion Closed	Proportion Hyphen	Proportion Open	PRED Closed	PRED Hyphen	PRED Open
home	town	74	.7432	00	.2568	.83599	.02143	.12658
video	game	74	00	.1351	.8649	.34447	-.01304	.69697
web	page	73	.0274	.0822	.8904	.08014	.05415	.92562
baby	boomer	72	00	.0417	.9583	.34386	.21828	.49172
ball	park	72	.9028	00	.0972	.56229	.05720	.41979
bill	board	72	1 00	00	00	.91006	.01711	.11470
life	sentence	72	00	.0139	.9861	.14440	.09004	.70059
privacy	policy	72	00	00	1 00	-.15292	.01952	1.15094
baby	sitter	71	.5211	.0282	.4507	.38511	.18101	.48102
league	baseball	71	00	.0141	.9859	.11127	.04470	.81973
merger	guideline	71	00	00	1 00	-.20396	.08622	1 688
mile	stone	71	1 00	00	00	.97511	-.06354	.02534
play	boy	71	1 00	00	00	.97444	-.04733	.09987
star	telegram	71	00	.9859	.0141	.39266	.35115	.32658
state	law	71	00	00	1 00	.18528	.03984	.76511
butter	fly	70	1 00	00	00	.93191	.08401	-.02754
nether	land	70	1 00	00	00	1.06205	-.06614	-.09066
play	ground	70	.9857	00	.0143	.93558	-.09067	.12394
radio	station	70	00	00	1 00	.52728	-.06221	.59220
self	defense	70	00	.9286	.0714	-.09210	.88807	.24578
shot	gun	70	.9571	00	.0429	.57371	.06807	.38243
enforcement	agency	69	00	00	1 00	-.07907	-.03274	1.12273
home	run	69	.0145	00	.9855	.38590	.17171	.40747
stair	case	69	1 00	00	00	.72176	-.01917	.34772
story	line	69	.0725	00	.9275	.67383	.04247	.25280
sun	light	69	1 00	00	00	1.03063	.04876	-.07636
war	time	69	.9710	.0145	.0145	.67191	.08642	.26445
band	width	68	1 00	00	00	.70777	.15769	.16868
cook	book	68	1 00	00	00	.87976	.04272	.10747
hip	hop	68	00	.9559	.0441	.37204	.22890	.45819
house	speaker	68	00	00	1 00	.33852	.06883	.67339
safe	guard	68	1 00	00	00	.82184	.07374	.11770
tax	break	68	00	00	1 00	.40904	924	.58464
war	crime	68	.0294	.1471	.8235	.35820	-.02259	.62311
breakfast	table	67	00	00	1 00	.28584	.08507	.76548
card	board	67	1 00	00	00	.74501	-.01004	.30880
country	music	67	00	.0597	.9403	-.09587	.09141	.88991
mail	box	67	.9254	00	.0746	.73333	-.06033	.31251
main	land	67	1 00	00	00	1.26183	-.02139	-.33370
movie	star	67	00	.1493	.8507	.55744	.24845	.25701
state	income	67	00	00	1 00	- 984	.05252	.87647
time	table	67	1 00	00	00	.33785	.05361	.61500
afl	cio	66	00	1 00	00	.29741	.52690	.27980
master	card	66	.9091	00	.0909	.44332	.04142	.58611
policy	maker	66	.3788	.3636	.2576	.59260	-.01142	.43655
saturday	morning	66	00	.0152	.9848	-.20308	.13816	1.08265

Table 2 (cont.)

W1	W2	TOTAL	Proportion Closed	Proportion Hyphen	Proportion Open	PRED Closed	PRED Hyphen	PRED Open
screen	writer	66	1 00	00	00	.56821	809	.45299
tax	rate	66	00	.0152	.9848	.21948	.02910	.79467
cover	package	65	00	00	1 00	.11474	-.13467	.97587
front	runner	65	.1077	.8923	00	.40949	.75197	-.06002
grass	root	65	.4923	.3846	.1231	.29351	.19171	.45955
interest	group	65	00	.0154	.9846	.12676	.09275	.84391
nick	name	65	1 00	00	00	.43390	.15562	.44269
security	council	65	00	00	1 00	.01226	-.04113	1 773
software	company	65	00	.0154	.9846	-.13093	.04876	1.14911
sun	beam	65	1 00	00	00	.67468	.17513	.17195
friday	night	64	00	00	1 00	.07793	.06647	.81564
letter	man	64	1 00	00	00	.76891	.06116	.12705
peace	talk	64	00	00	1 00	.42741	-.02017	.58925
price	increase	64	00	00	1 00	-.11863	.07389	.91231
tax	credit	64	00	00	1 00	.03802	255	.90742
volley	ball	64	.9688	00	.0313	.80610	.03061	.05553
eye	brow	63	1 00	00	00	.82534	.20461	- 009
sea	side	63	1 00	00	00	1.08708	-.02285	.01317
soap	opera	63	00	.0794	.9206	.39252	.15086	.46604
belt	way	62	1 00	00	00	.95805	-.05133	.03202
center	piece	62	1 00	00	00	.72604	899	.27278
hewlett	packard	62	00	1 00	00	.25691	.50706	.29683
online	service	62	00	00	1 00	-.22233	.12734	1.07114
type	writer	62	1 00	00	00	.72306	-.01513	.32672
work	station	62	.9839	00	.0161	.52425	-.06645	.54801
auto	maker	61	.9016	.0328	.0656	.70186	.02287	.25234
half	hour	61	00	.0164	.9836	.39669	.13584	.44502
night	club	61	.9344	00	.0656	.57922	-.05807	.48053
red	skin	61	1 00	00	00	1.07895	.11706	-.18251
sunday	morning	61	00	.1803	.8197	-.28879	.09368	1.12604
wall	paper	61	.9344	00	.0656	.59807	-.07183	.40111
wednesday	morning	61	00	00	1 00	-.22528	.15870	1.03114
blood	pressure	60	00	.0333	.9667	.28860	- 851	.66760
monopoly	power	60	00	00	1 00	.34866	066	.70697
oil	company	60	00	00	1 00	-.02964	.05787	.98940
science	fiction	60	00	.1500	.8500	.02098	.17576	.76184
thursday	night	60	00	.0167	.9833	.11349	.09411	.75366
west	coast	60	00	00	1 00	.49701	.06531	.42319
winter	time	60	.3000	00	.7000	.59273	.04736	.37604
child	abuse	59	00	.0847	.9153	.02712	.14692	.81871
competit	law	59	00	00	1 00	.10072	.01238	.95037
credit	union	59	00	00	1 00	-.11781	.07719	1.03050
dot	com	59	00	.9661	.0339	.57686	.40907	.13270
hotel	room	59	00	.0169	.9831	.52026	843	.47114
nobel	prize	59	00	00	1 00	-.01866	.16364	.79032
operating	system	59	00	.0339	.9661	.15346	.07503	.90805

Table 2 (cont.)

W1	W2	TOTAL	Proportion Closed	Proportion Hyphen	Proportion Open	PRED Closed	PRED Hyphen	PRED Open
seat	belt	59	.1186	.0678	.8136	.63684	.07126	.29920
story	telling	59	1 00	00	00	.47577	.16693	.38850
task	force	59	.0169	.0169	.9661	.51310	066	.50288
battle	field	58	.9828	.0172	00	.83796	.02735	.14616
campaign	reform	58	00	.2414	.7586	-.30096	.13752	1.10101
child	rearing	58	00	.3966	.6034	.25859	.37989	.48089
christmas	time	58	.1207	00	.8793	.30816	.10319	.59727
consent	agreement	58	00	00	1 00	-.04324	-.08172	1.07770
french	man	58	1 00	00	00	.91769	.01776	.03550
government	agency	58	00	00	1 00	-.29606	.05121	1.20815
hall	mark	58	1 00	00	00	.94347	.03522	.01914
news	stand	58	1 00	00	00	.79122	-.06644	.40037
night	life	58	.9138	00	.0862	.77912	-.01727	.29420
price	fixing	58	00	.7241	.2759	.26789	.32530	.52495
star	trek	58	00	00	1 00	.57076	.27260	.21956
base	line	57	.8596	.0175	.1228	.86200	.02752	.09508
finger	print	57	1 00	00	00	.72146	.09221	.19951
hormone	replacement	57	00	.0351	.9649	.10834	.13289	.78846
missile	defense	57	00	.1053	.8947	-.13555	.11649	1.03340
news	media	57	00	00	1 00	.27323	-.02611	.70044
paper	back	57	.9649	00	.0351	.91418	-.10346	.11773
death	sentence	56	00	.0357	.9643	.04578	.02001	.89791
house	wife	56	1 00	00	00	.55614	.07299	.36192
main	stream	56	1 00	00	00	.89777	.09610	.01314
party	line	56	.0536	.2679	.6786	.48614	.03254	.45043
self	interest	56	00	.9821	.0179	-.08632	.87522	.22577
spread	sheet	56	.9821	.0179	00	.54478	.14073	.35489
supply	side	56	00	.9286	.0714	.52299	.07658	.43090
tax	reform	56	00	.0536	.9464	-.03911	.03241	1 980
war	head	56	1 00	00	00	.99894	-.01142	.01389
water	fall	56	1 00	00	00	.86030	-.01212	.17076
birth	place	55	1 00	00	00	.64859	-.03685	.37337
health	plan	55	00	00	1 00	.04394	.02259	.87313
rod	man	55	1 00	00	00	.89130	.08667	-.01315
snap	shot	55	1 00	00	00	.78398	.12025	.11396
spin	network	55	00	00	1 00	- 526	.08180	1.01643
tax	revenue	55	00	00	1 00	-.03388	398	.97248
bull	pen	54	.9815	00	.0185	.72722	.08650	.22794
fbi	agent	54	00	00	1 00	.10429	- 965	.87951
fire	wall	54	.9630	00	.0370	.75904	- 739	.25984
hen	man	54	1 00	00	00	1.14405	.08085	-.25999
impeachment	trial	54	00	00	1 00	-.02385	-.01512	1.20060
internet	access	54	00	.0556	.9444	-.10763	-.03496	1.11560
line	item	54	00	.7778	.2222	.27481	-.02956	.69634
merger	enforcement	54	00	00	1 00	-.27073	.11990	1.17601
phone	company	54	00	.0185	.9815	-.08380	.01572	1.09066

Table 2 (cont.)

W1	W2	TOTAL	Proportion Closed	Proportion Hyphen	Proportion Open	PRED Closed	PRED Hyphen	PRED Open
race	track	54	.9444	00	.0556	.48474	.02461	.49459
back	bone	53	1 00	00	00	1.08081	-.05221	-.04104
cable	tv	53	00	.1698	.8302	-.11365	.14081	.86045
labor	party	53	00	00	1 00	-.25925	365	1.18881
media	company	53	00	00	1 00	-.01559	.03982	1.04104
news	magazine	53	.4717	00	.5283	.23081	-.06709	.79340
sky	scraper	53	1 00	00	00	.65355	.21462	.18703
tax	increase	53	00	00	1 00	- 025	-.06807	.98300
west	side	53	.1132	00	.8868	.88103	.06185	.08313
world	class	53	00	.9245	.0755	.06792	.09622	.80359
check	book	52	.8846	00	.1154	.96931	.02173	.04033
computer	industry	52	00	.0385	.9615	-.12043	.01991	1.09955
consumer	welfare	52	00	00	1 00	-.18118	.16953	1 351
fairy	tale	52	.1731	.2885	.5385	.58048	.11586	.30371
hill	top	52	1 00	00	00	1.10768	-.02384	-.10833
iran	contra	52	00	1 00	00	.28478	.54757	.22143
peter	son	52	1 00	00	00	1.05788	514	-.17617
power	house	52	1 00	00	00	.82186	-.03123	.24407
stein	berg	52	1 00	00	00	1.03474	.06462	-.20414
budget	deal	51	00	00	1 00	-.03998	-.03925	1.01031
budget	deficit	51	00	.0196	.9804	-.12486	.12951	.98364
computer	science	51	00	.0196	.9804	-.01840	.09214	.86647
double	day	51	1 00	00	00	.74747	.10985	.08806
glass	man	51	1 00	00	00	.91861	.03405	.01277
internet	service	51	00	00	1 00	-.09900	.06562	1.09786
state	government	51	00	00	1 00	-.13003	.08018	.98554
whistle	blower	51	.3922	.6078	00	.38877	.30388	.35073
approval	rating	50	00	00	1 00	.20437	.08603	.74083
chap	man	50	1 00	00	00	1.04935	.12381	-.21203
computer	system	50	00	00	1 00	.02059	.05563	1.02186
day	light	50	.9800	00	.0200	.81308	.01828	.17317
drug	dealer	50	00	00	1 00	.02468	.05718	.90397
fore	front	50	1 00	00	00	1.14505	.11130	-.21535
graduate	school	50	00	.0600	.9400	625	.03523	.93564
health	problem	50	00	00	1 00	-.05725	.07283	.97384
honey	moon	50	1 00	00	00	.55336	.26681	.23108
law	firm	50	00	00	1 00	.25189	-.08078	.74951
sci	fi	50	00	1 00	00	.42395	.49801	.14076
show	case	50	1 00	00	00	.84768	-.08073	.28738
weapons	inspection	50	00	.3000	.7000	238	.10448	.94004
bank	account	49	00	.0204	.9796	.26473	-.04065	.73449
bat	man	49	1 00	00	00	1.11609	.03470	-.18196
bell	curve	49	00	.0408	.9592	.51983	.03157	.40019
class	action	49	00	.6531	.3469	.04170	.02214	.83575
configur	space	49	00	00	1 00	.48246	.05537	.52575
defense	secretary	49	00	00	1 00	-.01250	-.04168	1.18406

Table 2 (cont.)

W1	W2	TOTAL	Proportion Closed	Proportion Hyphen	Proportion Open	PRED Closed	PRED Hyphen	PRED Open
farm	house	49	.9388	00	.0612	.87569	- 342	.16673
half	year	49	00	.0204	.9796	.23139	.15449	.56236
news	letter	49	.8980	00	.1020	.44662	-.01728	.56351
police	department	49	00	.0204	.9796	301	-.06691	1.04051
slap	stick	49	1 00	00	00	.84439	.07231	.07219
sound	track	49	.9592	00	.0408	.69256	-.02774	.34502
sunday	night	49	00	.0204	.9796	-.04532	.12791	.87176
bench	mark	48	1 00	00	00	.82842	106	.17001
blue	print	48	1 00	00	00	1.03616	.04212	-.05352
camp	site	48	.7708	00	.2292	.53475	-.01209	.53070
decision	making	48	.2708	.5208	.2083	.65130	.09898	.27988
estate	tax	48	00	.1042	.8958	-.06652	.04129	1.11827
gene	therapy	48	00	.0208	.9792	.22896	.07538	.67639
ice	cream	48	00	.2708	.7292	.19102	.13930	.68056
ice	storm	48	00	00	1 00	.47527	.06386	.60506
pop	culture	48	00	.2083	.7917	.18797	.05261	.72210
school	year	48	00	00	1 00	.38957	.01128	.59465
senate	race	48	00	.0208	.9792	.03900	-.01119	.95759
tooth	paste	48	1 00	00	00	.54337	.24526	.27253
word	processor	48	00	.0208	.9792	.12217	.13423	.66846
air	traffic	47	00	.1277	.8723	.35380	.02112	.58652
ball	game	47	.6170	00	.3830	.55329	.02733	.46294
body	guard	47	1 00	00	00	.51249	-.02949	.52776
business	leader	47	00	00	1 00	.17899	-.04857	.88562
cash	flow	47	00	.0213	.9787	.50515	.04549	.45114
drug	problem	47	00	00	1 00	-.06250	- 023	1.04277
horse	back	47	1 00	00	00	1.09367	-.03061	-.12829
man	kind	47	1 00	00	00	.65210	.16056	.14360
merger	case	47	00	00	1 00	.29303	-.03187	.78154
subject	matter	47	00	00	1 00	- 232	.03406	.89017
friday	morning	46	00	00	1 00	-.21011	.09516	1.08727
greenhouse	gas	46	00	.2174	.7826	.23499	.42557	.41467
plastic	bag	46	00	00	1 00	.29239	.04727	.66593
speed	way	46	1 00	00	00	.97540	-.09558	.06014
band	wagon	45	.9111	00	.0889	.56342	.13171	.33564
business	unit	45	00	.0444	.9556	-.04313	-.02279	1.01732
capital	gain	45	00	.0667	.9333	-.22121	.11369	1 741
coast	guard	45	00	00	1 00	.77737	.01392	.21379
corner	stone	45	1 00	00	00	.98382	-.03680	-.02388
defense	lawyer	45	00	00	1 00	.02307	-.03421	.96268
football	game	45	00	00	1 00	.18935	.02146	.84431
gas	mileage	45	00	00	1 00	.30950	.23812	.49878
gold	water	45	1 00	00	00	.85159	-.01962	.16367
lemon	ade	45	1 00	00	00	.41145	.25130	.28391
mccain	feingold	45	00	1 00	00	.19862	.55952	.29492
merger	review	45	00	.0222	.9778	-.16838	.03397	1.22363

Table 2 (cont.)

W1	W2	TOTAL	Proportion Closed	Proportion Hyphen	Proportion Open	PRED Closed	PRED Hyphen	PRED Open
run	way	45	1 00	00	00	.96508	-.02374	- 664
sail	boat	45	.8667	00	.1333	.84469	.07378	.09988
administrati	official	44	00	00	1 00	-.03325	- 699	1 986
base	man	44	1 00	00	00	1.04353	.01266	-.08849
campaign	fund	44	00	.1136	.8864	-.23169	.11591	1.03580
election	year	44	00	.3409	.6591	-.03783	.04797	.95924
foot	print	44	1 00	00	00	.96769	155	.05821
hatch	waxman	44	00	1 00	00	.40150	.46946	.20017
love	affair	44	00	.0227	.9773	.26769	- 205	.68588
post	office	44	00	00	1 00	.29528	.07587	.62706
show	room	44	1 00	00	00	1 001	-.07216	.08732
side	line	44	1 00	00	00	.87907	408	.09864
sky	line	44	1 00	00	00	1.01634	.03766	-.07409
well	stone	44	1 00	00	00	1.26636	648	-.34412
wood	land	44	1 00	00	00	1.21758	-.12223	-.18328
yellow	stone	44	1 00	00	00	1.14277	.01461	-.23643
air	wave	43	1 00	00	00	.66533	.06415	.26903
art	work	43	.8605	00	.1395	.74966	.03293	.25048
budget	surplus	43	00	00	1 00	-.12234	.14678	.90941
emergency	room	43	00	.0930	.9070	.49406	-.04221	.58894
fire	place	43	.9767	00	.0233	.76372	-.07900	.29634
foot	step	43	1 00	00	00	.89157	.04931	.11798
lip	stick	43	1 00	00	00	.74514	.05897	.18130
middle	man	43	.9767	00	.0233	.97527	.07383	-.09559
news	group	43	.9767	00	.0233	.54971	- 629	.50792
sex	life	43	00	00	1 00	.58243	.03451	.43428
state	park	43	00	00	1 00	.34530	.07723	.59162
stock	pile	43	1 00	00	00	.45869	.09849	.47711
training	camp	43	00	.0465	.9535	-.02836	.06615	.93023
use	net	43	1 00	00	00	.66981	.04261	.26934
world	economy	43	00	00	1 00	-.10393	.11636	.94499
age	group	42	00	00	1 00	.34068	.10767	.61832
life	insurance	42	00	.0476	.9524	.11353	.01081	.84584
man	slaughter	42	1 00	00	00	.54336	.34534	.22763
opinion	poll	42	00	00	1 00	-.04540	.07567	.93930
pass	word	42	1 00	00	00	.78954	.05341	.15693
profit	margin	42	00	.0238	.9762	.03144	.13083	.81728
school	age	42	00	.5476	.4524	.39980	-.01411	.61251
staff	member	42	00	00	1 00	.24065	-.05612	.80259
war	criminal	42	00	.0238	.9762	.30189	.06596	.60139
aluminum	can	41	00	00	1 00	.15742	.10132	.71124
background	check	41	00	00	1 00	.41881	-.02744	.60683
bow	man	41	1 00	00	00	1.13618	.03092	-.19828
capitol	hill	41	00	00	1 00	.38258	.14297	.45180
clinton	gore	41	00	1 00	00	.20375	.45566	.26240
growth	rate	41	00	.0244	.9756	.22035	.07245	.74573

Table 2 (cont.)

W1	W2	TOTAL	Proportion Closed	Proportion Hyphen	Proportion Open	PRED Closed	PRED Hyphen	PRED Open
investment	bank	41	00	00	1 00	.21074	.06012	.74660
lunch	time	41	.6829	.0244	.2927	.63433	.14111	.24052
senate	republican	41	00	00	1 00	-.07686	-.03031	1.10906
state	university	41	00	00	1 00	.26970	.08231	.71635
trade	agreement	41	00	00	1 00	-.11833	-.09596	1.11907
wind	mill	41	1 00	00	00	.81254	.04047	.20402
air	liner	40	1 00	00	00	.78103	.05425	.17418
ballot	box	40	00	.0500	.9500	.60450	-.04969	.41126
bed	time	40	.9500	00	.0500	.84062	.11595	.05805
consumer	report	40	00	00	1 00	-.30065	.03081	1.33050
defense	department	40	00	00	1 00	-.05457	-.07305	1.11111
dough	nut	40	1 00	00	00	.88260	.06901	.04561
drug	abuse	40	00	.0500	.9500	719	.08411	.90651
egypt	air	40	1 00	00	00	.31844	.28117	.33791
folk	lore	40	1 00	00	00	.52822	.20112	.28875
horror	story	40	00	00	1 00	-.07865	.04331	1 407
house	majority	40	00	00	1 00	.14715	.01240	.80221
independence	day	40	00	00	1 00	.41866	-.03708	.57135
internet	company	40	00	00	1 00	-.18085	.04409	1.20120
pine	apple	40	1 00	00	00	.48106	.14929	.39557
pocket	book	40	.7750	00	.2250	.55610	.07382	.38649
post	card	40	.9750	00	.0250	.62851	889	.44531
privacy	issue	40	00	00	1 00	-.16001	.05535	1.15240
show	time	40	1 00	00	00	.83807	.05614	.12790
space	time	40	.2500	.7500	00	.51274	.15133	.31141
sweat	shop	40	.9750	.0250	00	.82361	-.01673	.21591
tax	dollar	40	00	00	1 00	.10400	.06944	.82544
tobacco	settlement	40	00	.0750	.9250	-.07681	.02171	1.03307
touch	down	40	1 00	00	00	.82607	.17400	- 905
treasury	secretary	40	00	00	1 00	.03669	- 051	.96594
tv	station	40	00	00	1 00	.04160	.01839	.94512
world	leader	40	00	00	1 00	.13629	.08012	.76404
abortion	right	39	00	.3846	.6154	.21016	.02674	.78727
blood	clot	39	00	00	1 00	.56839	.30301	.25548
bull	market	39	00	.0769	.9231	.60347	.07245	.38925
cable	television	39	00	.1282	.8718	-.04240	.12677	.87470
chat	room	39	.1026	.0513	.8462	.79434	195	.21269
defense	contractor	39	00	.0256	.9744	.12356	.08096	.84636
education	system	39	00	00	1 00	.08176	.03261	1.03562
eye	ball	39	.9744	00	.0256	.97876	-.05746	-.01855
eye	witness	39	1 00	00	00	.26902	.12691	.50174
government	regulation	39	00	00	1 00	-.24169	.15247	1.13535
gun	fire	39	1 00	00	00	.76785	-.05189	.30047
jury	duty	39	00	00	1 00	618	.08626	.90157
labor	cost	39	00	.0256	.9744	-.05896	.05699	.99117
land	slide	39	1 00	00	00	.75572	.08478	.17744

Table 2 (cont.)

W1	W2	TOTAL	Proportion Closed	Proportion Hyphen	Proportion Open	PRED Closed	PRED Hyphen	PRED Open
life	span	39	.1795	.0513	.7692	.77066	.01344	.24168
night	line	39	1 00	00	00	.89018	- 600	.09877
race	relation	39	00	.0513	.9487	.10136	.05868	.79977
resale	price	39	00	00	1 00	.13364	.06904	.77650
river	side	39	1 00	00	00	.90271	.01182	.15281
security	number	39	00	00	1 00	.07946	-.02526	.99934
space	station	39	00	00	1 00	.44835	340	.68845
star	buck	39	1 00	00	00	.78394	.15631	.07987
track	record	39	00	00	1 00	.27839	.01556	.67917
wood	stock	39	1 00	00	00	.81905	093	.21931
black	mail	38	1 00	00	00	.78616	.05699	.20423
body	part	38	00	00	1 00	.18373	.02466	.73778
bounty	hunter	38	00	00	1 00	.41594	.18039	.39898
cable	company	38	00	00	1 00	-.02843	.07007	1.02056
computer	program	38	00	00	1 00	-.22599	.11974	1.15231
district	attorney	38	00	00	1 00	-.09519	.02654	1.01683
domain	name	38	00	.0263	.9737	.42972	.11923	.63017
gas	station	38	00	.0263	.9737	.39598	- 865	.75271
hall	way	38	1 00	00	00	1.10115	-.02698	-.13900
mail	delivery	38	00	.0263	.9737	.18500	.07195	.74764
parent	company	38	00	00	1 00	475	.11272	.94056
percentage	point	38	00	00	1 00	.50950	.04284	.49987
phantom	menace	38	00	00	1 00	.13206	.31809	.64789
price	tag	38	.0526	.0526	.8947	.39696	.14506	.49958
stone	man	38	1 00	00	00	1.08102	-.02183	-.08715
term	limit	38	00	.0263	.9737	.11910	243	.83861
town	house	38	.8158	00	.1842	.78700	374	.24703
tv	news	38	00	.0263	.9737	-.28765	.12231	1.06807
world	view	38	.6842	.0263	.2895	.36936	.06529	.58734
air	space	37	.9189	00	.0811	.68024	- 134	.35562
arms	race	37	00	00	1 00	.12440	.09426	.76058
back	pack	37	1 00	00	00	.97246	-.02659	.06764
baseball	game	37	00	00	1 00	.07285	.03086	.94221
bed	rock	37	1 00	00	00	.80458	.07823	.11148
book	club	37	00	.0270	.9730	.65412	-.02577	.51928
book	review	37	00	.1081	.8919	.28396	.02367	.65782
bristol	myer	37	00	1 00	00	.25691	.50706	.29683
cheer	leader	37	1 00	00	00	.54433	.06177	.39268
consumer	education	37	00	.0270	.9730	-.25081	.09221	1.10872
consumer	privacy	37	00	.0270	.9730	-.20844	.05442	1.07594
drug	store	37	.5676	00	.4324	.32995	.02468	.67020
drug	user	37	00	00	1 00	541	.05251	.92944
drug	war	37	00	.0541	.9459	.13316	-.03867	.87566
house	manager	37	00	00	1 00	.15905	-.06343	.85847
investment	banker	37	00	00	1 00	-.03019	.28849	.77284
market	definition	37	00	.0270	.9730	.15416	.15946	.75778

Table 2 (cont.)

W1	W2	TOTAL	Proportion Closed	Proportion Hyphen	Proportion Open	PRED Closed	PRED Hyphen	PRED Open
movie	goer	37	.9459	00	.0541	.64739	.36809	- 071
net	working	37	1 00	00	00	.79080	.06340	.13562
news	organization	37	00	00	1 00	.31282	-.06938	.75499
plea	bargain	37	00	.1351	.8649	.16536	.18746	.66076
red	wood	37	1 00	00	00	1.41759	-.05993	-.48048
search	engine	37	00	00	1 00	.14760	.02441	.78830
security	adviser	37	00	00	1 00	.04226	.01967	.95295
space	shuttle	37	00	00	1 00	.36240	.24584	.44536
suicide	bombing	37	00	.0811	.9189	.13677	.11643	.75697
telephone	number	37	00	00	1 00	.06251	- 419	1.12153
trust	fund	37	00	.0270	.9730	.32710	.02011	.63754
Wednesday	night	37	00	.0270	.9730	.04697	.10202	.80891
week	day	37	.9730	00	.0270	.72771	-.08598	.33096
white	head	37	1 00	00	00	1.17507	- 958	-.16520
work	week	37	.1622	00	.8378	.83281	-.08522	.19913
academy	award	36	00	00	1 00	.05458	.11940	.74853
air	bus	36	1 00	00	00	.64918	.05357	.33284
air	conditioner	36	00	.1389	.8611	.37643	.32140	.38344
apparel	industry	36	00	00	1 00	-.20708	.11973	1.07515
balance	sheet	36	00	.1667	.8333	.53446	.04073	.44117
city	council	36	00	00	1 00	-.03526	.04807	.87295
court	case	36	00	00	1 00	.59658	-.05499	.50951
cross	section	36	00	00	1 00	.26432	.07242	.75323
earnings	report	36	00	00	1 00	-.07179	-.04884	1.05629
east	side	36	.1111	.0278	.8611	.71666	.11475	.19528
farm	land	36	1 00	00	00	1.05565	-.12641	-.01625
government	intervention	36	00	00	1 00	-.14210	.16379	.97519
hind	sight	36	1 00	00	00	.79338	.21305	.01921
labor	secretary	36	00	00	1 00	-.15864	.02001	1.13542
labor	union	36	00	.0833	.9167	-.07452	.03254	1.02241
product	market	36	00	00	1 00	022	.07799	.97584
quantum	mechanics	36	00	00	1 00	-.04072	.25616	.90963
rate	hike	36	00	00	1 00	.32463	.07807	.62320
road	side	36	.9444	00	.0556	.88885	.03143	.15013
rock	star	36	00	.0556	.9444	.66855	.01017	.37327
school	choice	36	00	.0556	.9444	.19336	.06109	.67336
sports	car	36	00	.0556	.9444	.32579	.02990	.68786
time	period	36	00	.0278	.9722	.19679	.06153	.69237
tribune	herald	36	00	1 00	00	.12892	.41193	.45978
welfare	roll	36	00	00	1 00	.31759	.03913	.74958
bear	market	35	00	00	1 00	.41587	.14267	.49999
bombing	campaign	35	00	00	1 00	-.06926	518	1 481
boot	leg	35	1 00	00	00	.64115	.07007	.28213
camp	ground	35	.8571	00	.1429	.73705	-.04318	.26787
college	football	35	00	00	1 00	-.25979	.11230	1.05326
committee	chairman	35	00	00	1 00	-.01858	.11614	.83808

Table 2 (cont.)

W1	W2	TOTAL	Proportion Closed	Proportion Hyphen	Proportion Open	PRED Closed	PRED Hyphen	PRED Open
curb	side	35	.9429	00	.0571	.91874	.13244	.01415
fair	way	35	1 00	00	00	1.24424	-.06902	-.23532
fire	man	35	1 00	00	00	1.04494	977	-.09431
flash	back	35	1 00	00	00	1.15981	-.07689	-.14118
gamma	ray	35	00	.7429	.2571	.35113	.23509	.40521
health	initiative	35	00	00	1 00	.01620	.07922	.89397
hitch	cock	35	1 00	00	00	.97793	.22149	-.21848
internet	stock	35	00	.0286	.9714	.21133	- 360	.82012
lawn	mower	35	.1429	00	.8571	.25021	.20453	.58317
majority	leader	35	00	00	1 00	.12777	.07360	.81371
oil	price	35	00	.0286	.9714	.18217	.02141	.78601
party	leader	35	00	00	1 00	.06951	.06525	.83339
school	board	35	.0286	.0857	.8857	.84659	-.07715	.26824
slaughter	house	35	.8571	.1429	00	.81513	.04809	.16722
space	ship	35	.9714	00	.0286	.77002	.03050	.22620
suicide	bomber	35	.0286	00	.9714	-.04698	.23536	.83775
time	frame	35	.1143	.0286	.8571	.59614	.09322	.33147
unemployment	rate	35	00	00	1 00	.19031	.05024	.83071
weight	loss	35	00	.4286	.5714	.14444	.06397	.74984
welfare	state	35	00	.1714	.8286	.01000	.11777	.84991

Tables 3a-d: Free variants with total frequency < 35 (sorted alphabetically)

Table 3a - closed/hyphen/open variants

W1	W2	CLOSED	HYPHEN	OPEN	W1	W2	CLOSED	HYPHEN	OPEN
air	time	8	1	16	home	schooling	2	1	15
air	fare	13	1	5	home	school	1	2	7
air	base	1	5	11	horse	race	1	3	13
air	flow	5	2	2	ink	jet	26	1	2
art	house	1	15	1	land	owner	32	1	1
ball	cap	1	1	1	land	mine	2	4	21
beam	line	4	1	3	light	year	3	5	4
belly	dancer	2	1	1	lymph	node	3	1	5
birth	weight	1	2	3	mail	room	1	1	1
blah	blah	1	7	13	meat	space	2	1	1
boot	strap	10	1	2	middle	class	1	14	8
bottom	line	1	16	7	money	maker	8	1	5
brain	wave	1	1	2	money	man	1	4	3
bubble	gum	2	1	4	movie	maker	4	2	4
car	pool	11	12	3	odds	maker	4	1	1
car	wash	1	1	4	office	holder	5	4	2
center	stage	1	1	2	oil	field	6	2	19
chip	maker	1	1	5	peace	time	15	1	1
cinder	block	5	2	6	people	person	1	2	2
coffee	shop	1	1	13	phone	book	3	1	28
coffee	house	11	1	2	play	time	4	1	6
cruise	ship	1	2	19	policy	making	3	2	2
cry	baby	1	1	1	pop	star	1	3	28
cup	holder	1	2	4	pot	shot	6	1	1
deal	maker	4	4	10	power	sharing	1	1	1
decision	maker	3	13	12	punch	line	4	1	29
dime	store	2	2	1	race	horse	7	1	6
dirt	bag	1	1	2	radio	frequency	7	4	4
dose	fractionation	1	1	1	rain	forest	13	1	18
drum	roll	3	1	1	right	hand	1	3	10
end	game	12	1	4	right	wing	1	9	4
end	point	8	1	2	roller	coaster	1	14	15
eye	glass	20	1	3	salt	water	1	1	10
face	lift	3	1	6	sea	water	21	2	2
feed	bag	1	1	1	sea	view	1	1	2
ferry	boat	2	2	1	shock	wave	5	1	10
film	school	1	2	2	sky	dome	6	1	1
finger	nail	24	1	1	slime	ball	1	1	1
finger	tip	20	1	1	soul	mate	2	1	8
fire	power	5	1	2	space	walk	1	1	3
fire	bombing	3	2	1	spy	plane	1	1	4
fire	starter	1	1	1	stage	coach	7	1	1
fist	fight	8	1	1	sugar	cane	22	2	10
flame	thrower	3	1	1	taste	maker	2	1	3

Table 3a (cont.)

W1	W2	CLOSED	HYPHEN	OPEN	W1	W2	CLOSED	HYPHEN	OPEN
flex	time	4	2	8	test	drive	4	3	3
flip	flop	1	28	3	ticker	tape	1	1	3
flow	rate	8	1	2	ticket	holder	2	2	3
food	service	1	2	6	time	line	14	1	6
fortune	teller	4	1	1	trench	coat	2	2	6
front	line	15	12	3	trip	wire	4	1	2
fun	house	5	4	2	truck	driver	1	1	30
fund	raising	23	2	2	turtle	neck	18	1	1
gas	oil	1	2	2	voice	mail	8	3	18
glass	house	3	1	5	waste	water	14	3	2
guest	house	6	1	4	water	skiing	2	6	10
gun	maker	2	2	6	water	jet	2	1	4
half	time	22	1	2	water	ski	1	1	4
head	dress	10	3	1	way	station	1	1	1
heat	stroke	2	1	3	wine	tasting	2	1	1
hit	man	2	1	4	wing	nut	1	1	1
home	state	1	4	27	work	study	1	5	4
home	video	1	2	17	work	site	1	1	2

Table 3b – closed/hyphenated variants

W1	W2	CLOSED	HYPHEN	W1	W2	CLOSED	HYPHEN
spider	man	1	24	shadow	boxing	2	1
bell	south	23	1	shirt	tail	2	1
gun	shot	20	2	shoe	string	2	1
life	saving	10	10	sinker	ball	2	1
black	hole	15	2	top	gun	1	2
eye	lid	13	1	war	horse	2	1
ground	breaking	12	2	war	making	2	1
pin	point	13	1	axe	man	1	1
zig	zag	11	2	base	load	1	1
fiber	optic	1	11	belly	dancing	1	1
top	level	1	11	bench	scale	1	1
bobble	head	5	6	blood	sucker	1	1
bird	watcher	9	1	color	field	1	1
butt	head	1	9	day	tripper	1	1
half	life	1	9	deal	breaker	1	1
court	side	8	1	earth	mover	1	1
rear	view	5	4	fish	eye	1	1
ship	building	6	3	flat	top	1	1
house	cleaning	7	1	glamour	puss	1	1
rag	tag	5	3	glass	ceramic	1	1
sight	seer	7	1	glow	worm	1	1
chit	chat	6	1	goat	skin	1	1
lock	step	6	1	gum	ball	1	1

Table 3b (cont.)

W1	W2	CLOSED	HYPHEN	W1	W2	CLOSED	HYPHEN
nit	picking	6	1	hand	wringer	1	1
thumb	sucker	4	3	head	shaking	1	1
bogey	man	5	1	hop	head	1	1
campaign	gate	5	1	house	painter	1	1
party	goer	4	2	jack	boot	1	1
path	breaking	4	2	kick	boxing	1	1
peace	making	5	1	king	maker	1	1
shanty	town	3	3	land	holding	1	1
teen	age	2	4	latch	key	1	1
trend	setter	3	3	man	child	1	1
color	blindness	4	1	map	maker	1	1
motor	car	3	2	money	lender	1	1
pig	pen	2	3	night	clubbing	1	1
top	dog	4	1	path	breaker	1	1
will	power	4	1	poll	taker	1	1
bed	sheet	3	1	pudding	stone	1	1
horse	racing	2	2	rear	guard	1	1
info	tech	1	3	rise	time	1	1
jet	setting	1	3	river	bed	1	1
pack	horse	3	1	run	time	1	1
wing	spread	3	1	sales	lady	1	1
bar	stool	2	1	scandal	world	1	1
beach	goer	2	1	seat	mate	1	1
black	body	2	1	slave	holding	1	1
eye	shade	2	1	stream	flow	1	1
fox	hunting	1	2	strike	breaker	1	1
gall	bladder	2	1	stripe	domain	1	1
hand	knit	2	1	thermal	hydraulics	1	1
hour	glass	2	1	thirty	nine	1	1
lattice	work	2	1	thirty	three	1	1
lunch	pail	1	2	time	saving	1	1
pay	master	2	1	title	holder	1	1
piggy	back	2	1	track	goer	1	1
pom	pom	2	1	twenty	nine	1	1
self	quenching	1	2	water	colour	1	1
service	woman	2	1	wave	packet	1	1

Table 3c – hyphenated/open variants

W1	W2	HYPHEN	OPEN	W1	W2	HYPHEN	OPEN
abortion	clinic	3	10	life	science	2	5
acceptance	speech	1	9	life	support	3	13
accounting	fraud	1	3	lifetime	achievement	1	7
acid	deposition	1	1	light	bulb	1	12
acid	solution	1	6	light	green	1	1

Table 3c (cont.)

W1	W2	HYPHEN	OPEN	W1	W2	HYPHEN	OPEN
action	adventure	3	3	light	plane	1	2
action	figure	2	14	light	rail	2	1
action	film	1	7	lightning	rod	1	2
actor	politician	1	1	line	balancing	1	2
actors	register	1	1	line	drawing	1	1
address	book	1	10	lip	service	1	19
address	change	1	2	liquor	industry	1	5
age	distribution	1	8	litmus	test	2	22
aids	conference	1	20	living	room	1	23
aids	prevention	2	3	load	balancing	1	1
aids	research	2	24	loan	type	1	1
aids	vaccine	1	7	lobbying	law	1	1
air	balloon	1	5	locker	room	3	7
air	charter	1	1	love	fest	2	3
air	defense	1	12	love	letter	2	18
air	gun	1	1	love	note	1	2
air	polluting	1	1	love	triangle	1	3
air	power	1	24	lump	sum	4	10
air	pressure	1	2	lunch	hour	3	13
air	quality	1	18	lunch	table	1	2
air	raid	2	8	luxury	good	1	4
air	transport	1	7	luxury	object	1	1
aircraft	carrier	1	19	lynch	mob	3	5
airline	security	1	2	machine	age	1	2
airplane	security	1	1	machine	gun	7	22
airport	bookstore	1	4	mail	order	6	9
alarm	system	1	3	majority	minority	1	1
alkali	metal	1	5	mammary	cell	1	6
alternative	fuel	1	2	man	bite	2	1
alternative	medicine	2	3	man	machine	4	1
ambassador	nominee	1	1	man	woman	1	2
amino	acid	1	25	manor	house	1	1
amusement	park	2	12	market	bubble	1	1
anger	management	2	1	market	cap	3	14
animal	right	5	6	market	penetration	1	2
anion	exchange	2	4	market	person	1	1
apartheid	era	5	2	market	rate	2	5
apartment	block	1	2	market	research	3	16
appeals	court	1	10	market	rise	1	1
applause	meter	1	1	marriage	penalty	1	31
area	code	1	3	mass	circulation	8	1
arms	control	6	28	mass	destruction	1	24
arms	reduction	1	3	mass	market	16	6
arms	smuggling	1	1	mass	marketing	1	1
army	navy	1	1	mass	media	1	7
art	book	1	1	mass	murder	2	13
art	collection	1	5	mass	murderer	1	12
art	film	1	4	mass	spectrometry	3	1
art	history	1	21	mass	transit	3	4
art	object	4	5	master	race	1	1
art	rock	1	2	match	point	1	7
art	school	1	4	mating	game	1	2
art	sculpture	4	1	mccarthy	era	1	1
art	sensation	4	2	mean	field	2	1
art	supply	1	1	mean	square	2	1
art	world	2	30	meat	grinder	1	2
arts	section	2	3	media	art	1	2
assault	weapon	3	15	media	boy	1	2
assembly	line	7	23	media	bubble	1	1
assembly	room	1	3	media	business	3	11
asset	freeze	1	4	media	critic	1	6
asylum	seeker	4	5	media	criticism	3	2
atm	fee	1	5	media	mogul	1	9

Table 3c (cont.)

W1	W2	HYPHEN	OPEN	W1	W2	HYPHEN	OPEN
attack	dog	2	1	media	player	1	3
attention	deficit	5	4	media	savvy	4	3
attention	span	2	16	media	type	1	4
auto	accident	1	13	medicare	reform	1	4
auto	emission	1	1	medicare	tax	1	2
auto	insurance	1	8	medicine	man	1	12
auto	safety	2	1	medium	term	1	1
auto	theft	4	1	member	nation	1	6
award	winner	5	1	memory	chip	3	10
axe	murderess	1	1	memory	play	1	1
baby	boom	1	17	merchandise	trade	1	4
baby	seat	1	1	merger	mania	4	6
bachelor	party	1	1	meson	exchange	1	1
back	end	3	1	message	passing	1	2
back	pain	1	1	metal	ion	1	2
backscattering	channeling	1	1	metal	matrix	1	2
bah	humbug	1	1	metal	oxide	2	1
bail	bond	1	1	methanol	gasoline	1	1
bail	jumper	3	1	micro	device	1	3
banana	peel	1	2	middle	school	1	6
band	aid	4	1	milk	bottle	1	4
bank	merger	1	7	milk	carton	2	12
bank	president	1	1	mind	state	1	1
bank	vault	1	3	minimum	wage	16	2
bargain	basement	8	3	mining	industry	1	1
bargain	hunter	1	5	minority	language	1	8
bargain	price	1	5	minority	shareholder	1	1
base	layer	1	2	mint	julep	1	1
basis	point	2	4	miracle	cure	1	5
batting	practice	1	7	mirror	image	2	11
battle	cry	1	5	missile	guidance	3	1
beach	volleyball	3	7	missile	launching	1	1
bean	counter	1	7	missile	technology	1	9
beauty	contest	1	9	mob	mentality	1	1
beauty	pageant	4	4	mob	movie	1	8
bedtime	story	1	1	model	system	1	3
beer	belly	1	1	money	abuse	3	4
beer	snob	1	1	money	laundering	11	4
bench	press	1	2	money	loser	1	3
beta	decay	2	3	money	management	2	6
bible	belt	1	8	money	transfer	2	3
biotechnology	industry	2	5	mood	music	1	2
birthday	cake	1	4	morale	booster	1	1
birthday	suit	1	1	morning	drive	1	1
blast	furnace	2	1	morning	line	6	2
blood	clotting	1	2	mortgage	interest	1	6
blood	pool	3	1	mosquito	control	1	1
blood	sugar	2	1	mother	child	6	1
blood	supply	1	5	mother	love	1	1
blood	vessel	1	13	motion	picture	4	18
blue	ribbon	1	1	motor	scooter	1	2
boarding	school	1	7	motorcycle	gang	1	1
body	bag	1	6	mountain	bike	1	10
boiling	water	1	1	mouse	click	2	7
bomb	kit	1	1	movie	actress	1	2
bond	market	2	28	movie	career	4	1
bone	marrow	2	10	movie	celebrity	1	1
boo	boo	8	2	movie	character	4	3
boogie	board	1	1	movie	critic	1	10
book	chat	1	1	movie	director	1	5
book	length	4	1	movie	industry	1	12
book	reader	1	2	movie	mother	1	1
book	release	4	1	movie	movie	4	3

Table 3c (cont.)

W1	W2	HYPHEN	OPEN	W1	W2	HYPHEN	OPEN
book	stack	1	1	movie	taste	4	1
book	writer	1	1	movie	theater	1	28
boom	year	1	6	movie	ticket	9	10
border	control	1	2	multibillion	dollar	9	2
boston	area	6	4	multimillion	dollar	18	5
boundary	layer	1	2	mumbo	jumbo	5	7
box	cutter	1	2	murder	charge	1	6
brain	cell	3	5	murder	mystery	1	9
brain	development	1	8	muscle	relaxant	1	1
brass	band	1	4	music	biz	1	1
brass	tack	1	2	music	blood	4	1
breakfast	cereal	2	5	music	candle	4	1
breast	augmentation	1	1	music	hall	2	4
breast	reduction	1	1	music	library	1	2
broadband	access	1	3	music	network	6	1
broadcast	network	3	23	music	ray	3	1
broadway	style	1	1	music	spirit	1	1
bubble	chamber	1	1	music	store	2	9
buddy	buddy	1	1	music	time	4	1
budget	balancer	1	2	music	video	1	11
budget	balancing	6	2	name	brand	4	12
budget	buster	2	2	nanny	state	2	6
budget	cutting	2	1	nasa	ame	1	1
buffet	style	2	2	nation	state	18	1
bulletin	board	1	15	needle	exchange	7	16
bumper	sticker	1	11	nerve	damage	1	3
bus	station	2	5	nerve	gas	3	30
business	class	4	6	nest	egg	1	11
business	cycle	2	18	network	news	1	25
business	hour	1	3	network	paradigm	2	1
business	news	1	11	network	tv	2	11
business	process	1	8	neutron	activation	3	4
business	savvy	1	1	neutron	capture	1	4
business	school	4	11	neutron	diffraction	1	3
business	service	2	6	news	channel	1	7
business	side	1	5	news	development	1	10
business	system	1	5	news	gatherer	1	1
butterfly	wing	1	1	news	peg	1	2
cabbage	patch	4	4	newspaper	magnate	1	1
cabinet	level	9	1	night	shift	1	1
cable	car	2	9	night	vision	1	1
cable	modem	2	6	nitrogen	containing	1	1
cable	news	2	10	nobel	laureate	2	19
cafeteria	style	1	1	nose	cone	1	1
calendar	year	1	7	nose	ring	1	5
caller	id	2	5	november	december	1	1
campaign	expenditure	1	1	numbers	racket	1	1
campaign	law	1	4	nursing	home	3	20
campaign	manager	3	21	obedience	school	1	2
campaign	money	1	17	ocean	liner	1	2
campaign	season	1	3	ocean	view	1	3
campaign	trail	1	21	office	product	1	11
cancer	drug	1	15	office	supply	2	19
cancer	research	1	4	oh	radical	1	1
candy	store	1	4	oil	bearing	1	1
canon	builder	1	1	oil	drilling	1	3
capital	budget	2	2	oil	future	1	2
capital	crime	1	13	oil	pipeline	3	6
capital	market	1	21	oil	recovery	1	2
car	bomb	1	9	oil	refinery	1	8
car	bombing	2	1	oil	revenue	1	4
car	crash	1	17	oil	rig	2	3
car	insurance	4	18	oil	shipment	1	1

Table 3c (cont.)

W1	W2	HYPHEN	OPEN	W1	W2	HYPHEN	OPEN
car	rental	1	2	oil	shock	1	3
car	tax	1	3	olive	green	2	1
car	theft	4	11	opening	day	1	7
carbon	dioxide	1	33	opening	night	1	2
carbon	monoxide	1	9	opening	weekend	1	10
card	catalog	1	2	opinion	maker	4	5
care	package	2	4	opposition	party	1	24
career	choice	1	7	orange	vest	1	3
cargo	area	1	1	ore	formation	1	1
carpet	bombing	1	6	oscar	night	2	4
case	study	1	29	overtime	pay	1	1
cash	cow	1	13	oxygen	ion	1	2
cash	crop	2	2	package	delivery	1	3
cash	management	1	2	pain	relief	2	2
cash	register	2	15	paint	factory	1	1
cash	welfare	1	1	pair	transfer	1	1
casino	boat	1	1	pakistan	border	4	3
casino	gambling	1	9	pan	frying	2	1
cast	iron	3	4	paper	pusher	1	3
cat	person	1	8	paperback	book	1	3
cat	scan	1	1	parent	home	1	1
category	killer	1	2	parent	teacher	2	1
cation	radical	1	1	paris	match	1	5
cattle	breeder	1	1	particle	size	1	3
celebrity	murder	1	1	party	planner	1	1
celebrity	wedding	1	1	passenger	jet	1	4
celebrity	worship	1	1	passenger	plane	1	8
cell	type	1	3	passenger	side	2	3
cement	kiln	1	1	patent	holder	2	14
center	field	2	2	path	integral	2	2
cereal	box	1	5	patient	care	1	10
chain	letter	1	6	pattern	matching	4	1
chain	link	8	1	pattern	recognition	4	2
chain	mail	1	4	paycheck	protection	3	9
chain	reaction	1	10	payroll	tax	1	25
chair	rail	2	3	pc	maker	1	9
chamber	music	1	16	pc	type	1	2
channel	surfing	2	1	peak	form	1	1
charity	ball	1	2	peanut	butter	1	15
charm	school	2	2	penny	ante	1	4
charter	revision	1	3	pension	fund	1	23
checkout	line	2	1	people	man	1	1
cheese	fondue	1	1	pep	rally	3	3
chicago	area	2	4	pepper	spray	1	9
child	advocacy	1	1	period	piece	1	4
child	child	1	2	personality	cult	1	2
child	development	1	17	pest	control	1	5
child	health	1	2	pet	owner	1	10
child	porn	2	1	phase	change	1	1
child	pornography	1	11	phase	encoding	2	1
child	prodigy	1	2	phase	space	1	9
child	protection	1	1	phoenix	area	1	1
child	safety	2	5	phone	line	1	29
child	star	1	4	phone	service	1	21
child	support	11	23	phone	sex	2	10
child	welfare	5	8	phone	user	1	6
childhood	education	1	1	phony	baloney	6	2
chocolate	chip	2	6	photo	essay	1	17
christmas	tree	1	25	photo	op	6	13
church	burning	1	2	photo	opportunity	3	2
church	state	10	2	photon	counter	4	1
cigarette	advertising	1	1	physician	hospital	2	1
cigarette	packaging	1	1	picket	fence	1	3

Table 3c (cont.)

W1	W2	HYPHEN	OPEN	W1	W2	HYPHEN	OPEN
cigarette	tax	6	13	picture	postcard	1	2
cinderella	type	1	2	piece	part	3	1
circuit	board	1	2	pig	production	1	1
circuit	breaker	4	27	pilot	plant	1	10
city	dweller	1	4	pilot	training	1	3
city	state	12	2	pinch	hitter	1	2
city	town	1	1	pinch	runner	1	2
class	warfare	3	13	pit	bull	1	20
clay	court	1	1	pizza	box	1	2
climate	change	2	16	place	name	1	2
clinton	era	4	5	plane	wave	1	1
coal	gasification	1	6	plants	plant	1	1
coal	mine	1	8	plasma	spraying	1	1
coal	mining	2	6	plate	glass	1	3
coal	seam	1	1	plea	bargaining	3	14
coal	tar	1	2	pledge	week	1	1
cobalt	molybdenum	1	1	plumbing	supply	1	1
cocaine	use	1	10	pocket	protector	1	2
cocktail	party	1	20	point	size	1	1
code	name	1	1	poison	gas	2	6
coffee	table	4	19	police	brutality	4	22
coin	shop	1	1	police	news	1	1
coin	toss	1	1	police	state	1	8
coke	bottle	1	6	policy	tax	1	1
college	age	5	9	policy	wonk	2	9
college	basketball	3	9	polka	dot	3	4
college	tuition	4	14	pollution	control	1	5
collision	avoidance	4	2	pool	type	3	1
colon	cancer	1	16	pop	art	3	3
color	center	1	1	pop	history	1	1
color	coordination	1	1	pop	music	2	30
column	inch	2	4	pop	pilgrim	4	1
combination	drug	1	1	pop	pop	2	1
command	line	1	1	pop	psychology	1	6
community	building	1	1	pop	singer	1	5
community	policing	1	2	poppy	seed	2	6
community	source	3	1	population	control	4	3
computer	book	1	5	pork	barrel	8	8
computer	chip	1	19	pork	barreling	1	1
computer	expert	1	8	postage	stamp	2	6
computer	glitch	1	7	poster	girl	1	5
computer	magazine	1	2	pot	smoker	1	1
computer	nerd	1	2	potty	mouth	1	2
computer	programming	1	4	poverty	level	1	6
computer	service	1	5	power	generating	2	2
computer	simulation	1	19	power	grid	1	8
computer	software	2	16	power	law	3	18
con	artist	1	5	power	nerd	1	1
concentration	camp	6	28	power	play	1	5
confederate	flag	1	13	power	politics	1	2
confidence	booster	1	1	prayer	breakfast	1	2
confidence	builder	1	1	prep	school	6	6
conflict	resolution	1	1	prep	schooler	1	1
congress	senate	1	1	preschool	age	2	2
consensus	builder	1	2	pressure	drop	3	4
consensus	building	1	1	price	cutting	2	1
conspiracy	theorist	1	11	price	setter	1	1
conspiracy	theory	2	32	price	setting	2	1
consumer	electronics	3	10	print	media	1	10
consumer	good	1	21	prion	protein	1	1
consumer	market	1	20	prison	house	1	1
consumer	product	3	29	privacy	right	2	1
consumer	research	1	2	prize	winner	3	5

Table 3c (cont.)

W1	W2	HYPHEN	OPEN	W1	W2	HYPHEN	OPEN
consumer	right	1	1	problem	child	1	1
consumer	software	1	1	problem	solver	1	1
contact	group	1	1	product	development	1	10
convenience	store	1	13	product	liability	6	9
convention	center	1	7	production	cycle	1	2
cookie	cutter	5	4	production	line	2	19
cooling	tower	1	4	production	rate	1	3
copper	indium	1	1	profit	seeking	1	1
copper	wire	1	2	program	length	1	1
copyright	infringement	3	5	property	ownership	1	1
core	curriculum	1	9	protease	inhibitor	1	3
corner	store	1	3	protest	movement	1	3
correspondence	school	1	1	provider	service	2	3
cosmetic	surgery	1	10	pug	dog	1	1
cost	advantage	1	1	pulitzer	prize	1	22
cost	benefit	22	2	pulp	fiction	3	1
cost	containment	6	5	pumpkin	carving	1	1
cost	effectiveness	4	4	punch	card	1	10
cost	saving	1	23	punditry	duty	1	1
cost	savings	1	9	punitive	damage	1	9
cotton	field	1	4	punk	rock	3	6
country	club	1	24	puppet	master	1	1
country	house	2	11	putt	putt	2	2
country	rocker	1	1	quality	assurance	1	14
coupon	book	1	2	quantum	field	1	7
cow	egg	2	3	quantum	gravity	1	10
crack	growth	1	1	quantum	state	1	1
crash	test	3	4	quarter	acre	2	5
credit	information	1	3	quarter	century	21	10
credit	rating	2	21	quarter	inch	5	5
credit	repair	1	2	quarter	mile	2	2
credit	screen	1	2	quarter	million	3	3
crew	fatigue	1	1	quarter	point	10	2
crib	speech	1	3	queen	size	2	3
crime	control	1	9	question	mark	2	7
crime	lab	2	1	quotation	mark	1	11
crime	prevention	2	3	race	car	3	6
crime	reduction	1	6	race	consciousness	2	3
crime	scene	2	12	race	weekend	1	1
crime	spree	1	1	radiation	chemical	3	5
crime	stopper	1	1	radiation	therapy	1	5
cross	reactivity	1	1	raft	boy	1	1
crowd	pleaser	1	3	rain	delay	1	9
crown	ether	1	2	ranch	hand	1	1
cruise	missile	4	18	ranch	style	1	7
crystal	field	1	1	rap	music	1	12
ct	scan	1	12	rat	choice	1	4
cue	card	1	1	rate	price	1	3
currency	crisis	1	5	ratings	war	1	3
currency	exchange	1	6	ray	diffraction	1	27
customer	feedback	1	2	razor	wire	1	1
customer	service	5	26	reaction	rate	1	5
customs	service	3	7	reactor	year	2	1
cutting	edge	7	4	reader	discussion	4	1
damage	control	2	9	reader	response	1	1
dance	band	1	1	reagan	era	10	3
dance	music	1	5	reality	distortion	1	1
data	analysis	1	4	rear	end	3	2
data	collection	1	9	record	company	2	13
data	dependency	1	1	record	holder	1	1
data	mining	2	11	record	industry	1	4
data	processing	1	1	record	player	1	4
data	recorder	1	2	record	store	1	2

Table 3c (cont.)

W1	W2	HYPHEN	OPEN	W1	W2	HYPHEN	OPEN
data	storage	4	5	redox	reaction	1	1
date	night	1	1	refrigerator	lock	1	2
date	rape	2	5	religion	column	1	1
day	affair	1	2	removal	court	3	3
day	basis	5	2	rent	control	3	20
day	labor	1	1	rent	regulation	1	3
day	trader	1	29	replenishment	manufacturer	1	1
death	camp	1	7	restrictor	plate	5	4
death	squad	3	10	retail	apparel	19	11
debt	interest	1	3	retail	format	1	1
debt	reduction	1	9	retail	outlet	1	12
debt	service	1	4	retail	store	1	12
decay	scheme	1	1	return	mail	1	1
defense	industry	2	29	revenue	enhancement	1	1
deficit	hawk	3	1	rib	eye	1	1
deficit	reduction	7	23	ribbon	cutting	4	1
delay	line	1	1	risk	assessment	2	5
demand	side	4	6	risk	aversion	3	2
denver	area	1	1	risk	management	2	1
depression	era	16	1	risk	reduction	1	1
deputy	secretary	1	15	risk	sharing	5	3
desktop	software	1	2	rna	protein	1	1
development	cost	1	3	road	race	1	3
diet	drug	1	6	road	rage	1	7
diet	pill	1	16	road	trip	6	8
diffraction	pattern	1	2	rock	history	1	6
dining	room	1	18	rock	music	1	15
dinner	theater	1	5	rock	opera	1	2
dirt	road	1	14	rock	world	1	1
disability	study	1	1	role	modeling	2	1
disaster	aid	1	1	role	play	1	8
disaster	area	1	6	romance	novel	2	6
disaster	movie	1	7	room	service	1	5
disaster	relief	2	9	room	temperature	4	14
disaster	response	1	2	root	canal	1	4
discount	rate	1	6	rubber	stamp	3	4
discount	store	1	8	rule	breaker	4	1
disk	drive	1	4	ruling	class	1	2
dna	test	1	30	rush	hour	5	12
document	shredding	1	1	safety	deposit	1	2
dog	bite	3	4	safety	pin	1	4
dog	name	3	2	safety	valve	1	3
dollar	value	1	12	salary	cap	9	20
dorm	room	4	4	satellite	dish	3	18
dose	equivalent	1	1	satellite	phone	2	4
dose	rate	3	9	satellite	technology	1	9
draft	age	2	3	satellite	television	1	3
draft	night	3	1	satellite	tv	2	7
drag	queen	1	1	saturation	coverage	1	3
drawing	room	2	2	saturday	evening	1	12
dress	shirt	9	18	scale	model	1	5
drinking	water	1	11	school	bus	2	19
drip	drip	1	1	school	enrichment	1	1
drug	approval	1	3	school	prayer	1	14
drug	delivery	5	5	school	project	1	1
drug	education	1	2	school	violence	1	1
drug	money	6	3	school	yearbook	1	7
drug	policy	2	21	science	education	1	3
drug	possession	1	3	scrap	metal	1	4
drug	prevention	1	2	screen	name	4	6
drug	price	2	9	screen	time	1	5
drug	prohibition	1	1	sea	monster	1	1
drug	rehabilitation	1	5	search	result	1	4

Table 3c (cont.)

W1	W2	HYPHEN	OPEN	W1	W2	HYPHEN	OPEN
drug	smuggling	2	3	season	ticket	3	28
drug	testing	10	4	seat	pocket	1	1
drug	trade	1	12	securities	fraud	1	2
drug	treatment	1	19	securities	law	1	6
east	west	15	3	security	agency	6	15
efficiency	enhancing	15	1	seed	corn	2	2
election	day	1	27	self	assertion	2	1
election	law	1	18	self	control	23	1
election	night	1	9	self	destruct	3	1
election	spending	1	1	self	governance	4	1
election	time	2	6	self	help	4	6
electron	beam	9	16	self	ignition	1	1
electron	capture	1	1	self	image	16	1
electron	cyclotron	1	2	self	improvement	25	6
electron	donor	1	1	self	portrait	26	1
electron	energy	2	1	self	preservation	10	2
electron	probe	1	1	self	protection	9	5
elephant	dung	3	18	self	respect	9	1
elephant	man	1	4	service	industry	1	4
elevator	music	1	3	sewage	treatment	1	1
elite	university	1	4	sex	abuse	3	9
emerald	green	1	1	sex	change	3	1
employee	benefit	1	7	sex	education	3	15
employer	discrimination	1	3	sex	scandal	7	27
employment	cost	1	1	sex	talk	2	5
employment	discrimination	1	2	sex	toy	1	3
end	effect	1	1	shale	oil	1	2
end	use	4	1	share	price	2	24
end	user	7	17	share	shifting	1	1
end	zone	1	13	shell	model	2	2
energy	conservation	1	4	shipping	container	1	8
energy	flow	3	1	shirt	size	1	1
energy	saving	5	4	shock	rocker	1	2
enterprise	zone	1	2	shock	therapy	1	3
entertainment	industry	2	18	shoe	leather	2	2
entry	level	14	1	shop	floor	2	2
equity	market	1	5	shot	clock	1	1
estrogen	binding	3	1	shotgun	marriage	3	1
ethanol	subsidy	1	13	siamese	twin	1	4
ethics	commission	5	2	side	aspect	1	1
exchange	rate	2	32	side	dish	1	6
excision	repair	1	1	silk	road	1	1
executive	branch	1	24	silver	screen	1	4
executive	compensation	1	7	site	characterization	1	1
executive	producer	1	9	site	selection	1	6
exercise	machine	1	3	ski	lift	2	12
expense	account	3	9	ski	vacation	1	1
experience	life	1	1	skirt	chaser	1	1
explosive	detection	3	1	slag	heap	1	3
eye	candy	1	1	slave	labor	3	11
eye	level	1	1	slave	revolt	1	4
fabric	softener	1	1	slave	ship	2	2
fact	check	1	3	slave	trade	1	15
fact	checker	6	6	slave	trader	1	1
fact	finder	2	1	sleep	deprivation	1	2
factory	floor	2	8	snail	mail	6	8
fall	line	1	5	snake	oil	4	3
family	assistance	1	1	snuff	film	1	2
family	man	1	10	soap	star	1	1
family	planning	13	1	soccer	ball	1	3
family	reunification	1	2	soccer	mom	1	14
fantasy	play	2	11	software	development	1	23
fantasy	theme	1	1	sound	wave	1	3

Table 3c (cont.)

W1	W2	HYPHEN	OPEN	W1	W2	HYPHEN	OPEN
fantasy	type	2	1	source	receptor	1	1
farm	assistance	1	1	source	rock	1	1
farm	state	3	2	sovereign	immunity	1	1
fashion	magazine	3	11	soviet	era	3	1
fashion	plate	1	1	space	age	4	4
father	son	8	1	space	agency	1	5
fbi	file	1	14	space	charge	2	1
fbi	lab	1	3	space	heating	1	2
feature	film	1	11	space	probe	1	1
ferrite	garnet	1	1	spear	gun	1	1
fever	dream	1	2	speech	recognition	2	11
field	effect	2	1	speed	cable	1	2
field	goal	1	10	sperm	bank	3	2
field	line	1	2	spin	glass	2	1
fig	leaf	1	9	spirit	world	1	1
fighter	bomber	4	1	sport	sedan	1	3
fighter	pilot	1	10	sport	utility	3	7
figure	skating	2	11	sports	event	1	6
film	business	1	4	sports	team	1	10
film	collector	1	1	sports	viewership	1	1
film	music	1	2	spot	market	2	2
film	student	1	3	spot	price	1	3
film	study	1	1	spray	paint	1	4
film	style	1	1	spy	museum	1	1
finance	industry	1	3	spy	scandal	1	2
finger	food	1	3	square	foot	1	7
finger	work	1	1	stage	father	1	1
fire	detection	1	1	standard	issue	4	2
fire	engine	1	5	standard	size	2	1
fire	pump	1	1	star	director	1	1
fire	rescue	3	1	star	game	1	6
fire	retardant	1	1	state	action	1	26
fire	safety	1	2	state	agency	1	15
fire	sale	3	6	state	assembly	1	3
fire	suppression	2	1	state	college	1	14
firm	size	1	3	state	election	1	4
fish	market	1	1	state	security	1	11
fission	product	2	7	steam	engine	1	4
flag	burner	1	1	steel	band	1	1
flag	cover	1	1	steel	jaw	1	1
flag	waver	1	2	steel	mill	1	5
flea	control	2	4	steel	pipe	1	1
flea	market	1	11	stimulus	response	1	1
flight	operation	1	2	stock	car	2	10
flood	relief	1	1	stock	exchange	1	18
flood	victim	1	2	stock	index	1	26
floor	window	1	1	stock	picker	1	1
floppy	disk	1	8	store	window	1	3
flu	vaccine	2	1	storm	trooper	2	2
flue	gas	3	4	story	note	1	10
fluid	bed	2	1	strain	rate	1	4
fly	ash	1	2	straw	poll	3	26
focus	group	3	27	street	corner	4	17
folk	music	1	11	street	crime	1	11
food	distribution	3	4	street	level	9	1
food	fight	1	3	street	name	1	6
food	inspection	4	2	street	savvy	1	2
food	safety	1	11	stress	corrosion	1	1
food	shop	1	3	stress	relief	4	2
football	field	1	7	strip	club	2	6
fossil	fuel	5	10	strip	mall	1	2
frat	boy	8	7	strip	mining	2	1
frat	house	2	1	stroke	victim	1	1

Table 3c (cont.)

W1	W2	HYPHEN	OPEN	W1	W2	HYPHEN	OPEN
freeze	frame	1	8	student	aid	1	1
frequency	dependence	1	1	student	faculty	3	2
front	end	19	3	student	loan	2	31
fuel	injection	1	6	student	teacher	2	9
fuel	line	1	5	studio	musician	1	1
fuel	oil	1	4	stun	fencing	1	4
fuel	pipe	1	1	stun	gun	1	5
fuel	tank	2	16	style	sheet	2	3
fund	drive	1	3	submarine	crew	1	1
funeral	home	1	4	subscriber	acquisition	1	1
fur	industry	1	1	substance	abuse	3	15
fusion	energy	1	1	subway	turnstile	1	1
gale	force	1	2	sugar	beet	4	3
gambling	addiction	1	2	sui	generis	1	2
game	day	1	2	suicide	bomb	1	6
game	winner	2	3	sulfur	dioxide	1	11
gamma	irradiation	3	2	summer	blockbuster	2	9
gamma	radiation	3	6	summer	camp	1	22
gang	rape	1	1	summer	development	1	2
gangsta	rap	1	14	summer	league	5	17
garage	door	2	6	summer	movie	1	8
garden	shop	1	1	summer	work	2	2
garden	variety	12	5	superfund	site	1	3
gas	emission	1	8	supermarket	tabloid	1	10
gas	engine	1	3	supply	chain	2	13
gas	guzzler	4	5	supply	sider	17	1
gas	guzzling	7	1	support	group	1	15
gas	molecule	1	4	suv	owner	1	3
gas	oven	1	1	sweeps	month	1	1
gas	phase	4	6	switch	hitter	1	1
gas	turbine	1	3	system	software	1	5
gas	water	1	1	systems	integration	1	3
gel	filtration	1	1	tag	team	4	2
gen	xer	13	5	tail	end	1	5
gender	equality	3	9	talent	search	1	1
gender	equity	1	2	tap	tap	2	1
gender	gap	1	13	tape	cassette	1	1
gender	integration	1	1	tape	record	2	2
gender	role	1	5	target	cell	1	3
gender	study	1	1	tariff	reduction	1	1
gender	war	1	1	tax	collection	2	6
generator	coordinate	2	1	tax	cutter	2	2
gentleman	scholar	1	3	tax	dodge	1	1
germ	warfare	1	6	tax	evasion	1	15
gift	giver	1	2	tax	fraud	1	10
gift	tax	1	8	tax	hike	2	18
gift	wrap	1	1	tax	law	1	22
girl	power	1	3	tax	loss	1	4
glucose	tolerance	1	1	tax	reduction	2	10
goal	line	1	2	tax	relief	1	15
gold	medal	3	18	tax	savings	1	2
gold	rush	1	9	tea	leaf	1	1
golf	ball	1	30	tea	party	1	2
golf	cart	1	12	team	record	1	2
gourmet	food	2	3	tear	gas	2	28
government	business	2	2	tech	gadget	2	1
government	issue	1	3	tech	industry	1	32
government	newspaper	1	4	tech	sector	1	7
government	program	1	22	tech	stock	2	24
government	subsidy	1	18	technology	shock	1	1
grab	bag	1	8	technology	transfer	3	13
grad	school	1	3	telephone	answering	2	1
grad	student	2	15	television	ambush	4	1

Table 3c (cont.)

W1	W2	HYPHEN	OPEN	W1	W2	HYPHEN	OPEN
grade	point	2	8	television	crisis	4	1
grade	school	5	10	television	murder	4	1
grade	schooler	1	1	television	news	1	13
graduate	level	1	1	television	oz	4	1
graduation	speech	1	1	television	production	1	1
grain	growth	1	1	television	roar	4	1
grammar	school	2	6	television	sex	4	1
grass	court	2	3	television	sister	4	1
grass	grass	2	1	television	sitcom	4	2
gray	metal	1	2	television	station	1	18
ground	floor	1	11	television	winter	3	1
ground	level	3	8	temper	tantrum	1	6
ground	rule	3	6	tennis	court	1	21
ground	state	4	7	tenure	track	1	1
ground	war	1	23	term	paper	1	4
group	name	1	1	test	facility	1	1
group	sex	1	1	test	result	1	15
group	therapy	1	1	test	score	4	29
grudge	match	1	3	test	taker	5	1
guerrilla	leader	2	2	test	track	1	1
guinea	pig	2	11	test	tube	4	8
gun	nut	1	6	textile	mill	1	5
gun	ownership	1	10	thatcher	era	1	1
hair	care	3	5	theater	side	4	1
hair	loss	1	3	theater	triumph	4	1
half	billion	1	1	theme	park	7	14
half	century	1	8	theme	song	1	4
half	cup	1	1	thought	experiment	2	3
half	cycle	1	1	thread	count	1	1
half	dozen	1	1	three	year	3	1
half	live	4	1	thrift	shop	2	1
half	period	1	1	thrift	store	2	2
half	sister	1	1	thrust	fault	1	1
hand	eye	1	1	ticket	taker	1	2
hand	gesture	1	5	tie	breaker	1	1
harm	reduction	1	1	tie	game	1	4
hatchet	man	1	4	timber	industry	1	1
hate	crime	1	24	time	bomb	1	8
hay	bale	1	1	time	capsule	2	1
health	food	6	20	time	reversal	1	1
health	maintenance	2	6	time	sery	1	1
health	policy	2	6	time	shift	1	1
health	science	1	2	time	synchronization	1	1
health	study	1	10	time	trial	4	24
hearing	aid	1	2	time	trialer	1	1
heart	bypass	1	1	time	zone	3	9
heart	lung	1	1	tin	ear	1	8
heart	valve	8	11	tongue	twister	1	1
heat	energy	1	1	topsy	turvy	15	1
heat	exchanger	1	10	tote	bag	1	2
heat	pump	3	21	touch	screen	2	2
heat	transfer	4	17	town	hall	3	21
heat	wave	2	20	toy	industry	1	1
hip	replacement	1	4	toy	store	2	5
hiv	aid	2	2	tract	house	1	1
hiv	transmission	1	10	trade	liberalization	2	7
holiday	inn	4	7	trade	show	2	7
holiday	movie	1	3	trade	union	1	15
holiday	shopping	1	1	trailer	park	3	2
holiday	spirit	1	2	trailer	trash	2	2
hollywood	actor	1	1	training	mission	3	7
holocaust	memorial	1	1	transit	agency	1	1
holocaust	survivor	1	12	transition	class	1	4

Table 3c (cont.)

W1	W2	HYPHEN	OPEN	W1	W2	HYPHEN	OPEN
home	automation	2	1	transition	metal	1	6
home	brew	1	1	transmission	safety	1	1
home	delivery	1	6	travel	book	1	1
home	design	1	1	travel	industry	1	8
home	economics	1	5	travel	office	8	2
home	entertainment	1	3	trend	line	1	2
home	equity	2	2	trend	spotter	2	1
home	field	2	1	trial	balloon	1	2
home	furnishing	1	5	trial	court	1	8
home	health	3	6	trial	lawyer	1	28
home	improvement	3	13	trick	pony	1	3
home	mortgage	2	7	trillion	dollar	5	14
home	office	1	9	truck	bomb	1	5
home	pc	1	1	trumpet	player	1	4
home	plate	2	17	tv	campaign	1	1
home	rule	1	4	tv	column	3	3
home	security	1	2	tv	folk	1	1
home	study	1	1	tv	movie	2	19
home	team	1	6	tv	network	1	26
home	theater	1	1	tv	right	1	1
homogeneity	region	1	1	tv	sunday	1	1
horror	film	1	11	tv	talk	1	11
horror	movie	2	12	tv	watcher	2	4
horse	breeder	1	1	twin	engine	1	1
host	family	2	1	union	election	1	5
hostage	taker	1	1	union	leader	2	24
hotel	casino	1	1	unit	cell	7	4
house	arrest	1	20	urinal	drain	1	1
household	name	1	14	user	group	1	1
housing	discrimination	1	1	uv	irradiation	3	1
human	resource	1	1	valve	replacement	1	3
husband	girlfriend	1	1	vanity	plate	1	1
hydrogen	form	1	1	vapor	phase	1	4
hydrogen	ion	1	2	vehicle	price	1	1
hydrogen	sulfide	1	3	velocity	distribution	1	2
ibm	pc	2	13	venture	capital	5	14
ice	skating	2	4	vice	president	15	12
identity	theft	1	24	vice	squad	1	1
immigration	reform	1	3	video	editing	1	1
income	earner	2	2	video	poker	1	24
income	housing	1	2	video	store	1	15
income	redistribution	1	5	vietnam	era	4	7
index	fund	5	27	virus	type	1	1
industry	standard	1	21	voice	dictation	2	1
infant	mortality	2	13	voice	recognition	5	5
inflation	adjustment	1	2	volume	loss	1	1
influence	peddler	1	2	voter	registration	2	12
influence	peddling	7	2	wack	job	1	1
information	age	2	32	wage	slave	1	2
information	collection	1	8	wall	hanging	1	2
information	gathering	2	2	wall	painting	1	1
information	integration	3	4	war	book	1	2
information	sharing	5	27	war	era	10	6
instant	replay	1	6	war	reparation	1	2
insurance	business	1	2	warp	speed	2	1
insurance	fraud	1	1	wartime	atrocity	1	1
internet	boom	1	3	washington	insider	1	8
internet	business	1	8	washington	politician	1	5
internet	porn	1	7	waste	disposal	3	15
interstate	highway	1	10	waste	incineration	1	1
investment	banking	4	15	water	bottle	1	5
investment	grade	5	5	water	level	1	7
investor	relation	1	2	water	power	1	1

Table 3c (cont.)

W1	W2	HYPHEN	OPEN	W1	W2	HYPHEN	OPEN
iodine	containing	1	1	water	pressure	1	5
ion	exchange	8	3	water	purification	1	1
ion	implantation	2	4	water	quality	5	3
ion	pair	2	1	water	use	1	2
ion	plating	1	2	wax	figure	1	1
iron	ore	1	2	wealth	creation	1	4
issue	advocacy	2	9	weather	modification	2	10
ivory	tower	4	13	web	business	1	1
japanimation	style	1	1	web	design	1	1
jeans	maker	1	1	web	savvy	2	1
jersey	wearer	1	1	web	surfing	1	3
jet	lag	1	7	web	web	1	1
job	approval	4	7	wedding	chapel	3	2
job	creation	1	11	wedding	night	3	3
job	discrimination	1	11	welfare	commission	2	1
job	loss	3	6	welfare	entitlement	1	1
job	program	1	1	welfare	office	1	4
job	search	1	11	welfare	policy	1	5
job	security	1	27	welfare	tanf	1	1
job	training	4	16	west	bank	1	4
july	august	2	1	wife	beater	1	1
jumbo	jet	1	1	window	dresser	1	1
junk	bond	9	12	window	treatment	1	2
junk	food	2	18	wine	shop	1	1
jury	leak	1	1	wire	service	1	10
karaoke	bar	1	2	witch	hunt	2	14
kennedy	era	2	2	wood	frame	2	3
kid	culture	1	1	word	choice	1	4
kid	glove	5	2	word	processing	4	1
kiddie	cash	8	3	work	ethic	1	17
kidney	dialysis	1	2	work	life	5	3
kidney	failure	1	4	work	rule	1	6
king	size	3	1	work	sharing	1	1
labor	law	1	10	work	task	2	11
labor	lawyer	1	5	work	time	1	3
labor	management	5	1	worker	hour	1	1
labor	market	2	25	working	class	1	2
labor	negotiation	1	7	workplace	harassment	1	2
labor	relation	2	16	world	aid	1	1
laissez	faire	10	1	world	champion	1	13
lance	writer	1	2	world	end	1	1
land	management	1	3	world	government	1	12
land	use	2	3	world	record	2	14
language	column	1	1	world	story	1	2
law	partner	1	8	world	trade	1	8
law	review	1	3	world	weariness	2	2
lawn	care	2	2	wright	patterson	2	1
lawyer	letter	1	1	writer	director	22	1
leading	edge	1	2	yadda	yadda	1	3
left	wing	4	1	yard	line	1	4
lemon	lime	2	2	yard	sale	1	2
leopard	print	3	1	year	end	29	3
leopard	skin	2	2	yin	yang	1	1
letter	writer	4	15	youth	culture	1	6
life	form	2	5	yttrium	aluminium	1	1

Table 3d - closed/open variants

W1	W2	CLOSED	OPEN	W1	W2	CLOSED	OPEN
air	crew	3	5	motor	boat	8	4
air	show	1	3	motor	coach	1	5
air	stream	3	1	motor	home	2	25
air	strip	10	2	motor	sport	1	1
altar	piece	12	1	movie	talk	1	1
apple	sauce	4	1	mud	flat	1	1
arm	band	2	1	mug	shot	1	4
arm	pit	16	1	muscle	man	1	1
art	forum	1	1	name	tag	1	1
art	news	3	1	nap	time	1	5
ash	tray	9	2	needle	point	4	3
attic	wall	1	1	net	system	1	1
audio	book	5	1	news	box	2	19
audio	tape	11	2	news	break	1	1
auto	worker	1	7	news	brief	1	1
baby	food	1	9	news	cast	26	3
back	door	13	9	news	gathering	1	1
back	room	10	8	news	page	1	7
back	seat	28	1	news	person	4	1
back	side	14	3	news	play	2	2
back	street	13	1	news	print	27	5
ball	club	4	5	news	rack	1	1
ball	field	1	3	news	reader	2	1
ball	gown	1	3	news	service	1	31
band	leader	1	1	news	team	1	1
band	mate	1	1	news	wire	4	6
bank	card	5	1	news	writer	1	1
bank	head	1	2	newspaper	man	4	1
bank	note	1	4	night	spot	3	1
bar	code	1	33	night	stand	3	8
barber	shop	5	2	night	time	28	5
base	pair	2	1	noon	time	2	1
base	runner	6	1	north	side	3	11
bath	tub	26	2	nose	dive	4	5
bay	leaf	1	2	note	pad	5	1
bay	side	2	1	nursery	man	1	1
beach	side	10	3	nut	cake	1	1
bean	bag	2	1	nut	case	4	3
bean	field	1	6	nut	job	2	2
bed	bug	2	1	ocean	side	5	2
bed	roll	2	1	oil	fire	1	2
beer	hall	1	1	oil	man	5	1
belly	button	2	3	paint	brush	7	2
belt	line	3	5	paint	can	1	3
bill	collector	1	1	palm	pilot	10	6
bird	cage	3	2	pan	handle	21	1
bird	song	3	1	paper	clip	2	6
birth	date	3	10	paper	towel	1	16
birth	rate	5	12	park	land	24	2
black	man	7	2	particle	board	5	1
blood	bath	14	9	passage	way	27	1
blood	line	5	1	pasture	land	1	2
blood	stain	2	3	pawn	shop	6	2
blood	stream	11	3	pay	day	21	1
blow	dryer	2	1	pea	coat	1	2
blue	jean	1	1	peach	tree	15	6
blue	water	4	1	pet	food	1	2
blues	man	2	1	photo	journalism	3	1
board	room	23	2	pickle	jar	1	1
boarding	house	2	1	picture	book	1	3
boat	house	1	2	pigeon	hole	4	1
body	builder	8	1	pin	hole	1	1

Table 3d (cont.)

W1	W2	CLOSED	OPEN	W1	W2	CLOSED	OPEN
bond	strength	1	1	pin	prick	3	2
boogie	man	1	1	pin	stripe	5	1
book	selling	7	1	place	mat	1	3
book	shelf	22	1	plane	load	4	1
book	shop	29	1	plaster	board	2	1
boom	box	2	4	play	date	2	2
boom	town	4	1	play	room	22	1
border	line	32	1	play	space	2	5
bottle	brush	1	1	plot	line	4	12
box	car	8	1	pop	eye	5	1
bread	line	3	1	poster	boy	2	8
bread	winner	4	1	pot	belly	3	1
brick	layer	2	1	pot	luck	7	1
brush	stroke	4	1	power	boat	3	3
bull	whip	5	1	power	book	4	1
bully	boy	1	3	power	broker	2	8
bus	load	7	1	power	train	1	1
business	woman	5	5	press	box	1	2
button	hole	2	2	press	time	2	5
buzz	word	22	2	punch	bowl	2	2
cab	driver	5	7	quarter	horse	1	2
camera	man	22	1	racquet	ball	30	2
camera	work	3	2	radiation	dose	1	8
camp	fire	27	6	radio	shack	5	5
car	line	1	1	radio	signal	1	4
car	maker	17	3	radio	therapist	1	2
car	port	8	2	rail	car	5	2
card	holder	2	3	rain	drop	7	2
care	giver	31	3	rain	man	2	5
case	load	14	1	rain	storm	5	5
cat	fight	1	3	rat	hole	1	5
catch	phrase	5	1	rat	lung	1	1
cave	man	2	1	rat	pack	1	11
cell	mate	2	1	record	keeping	1	1
center	fielder	1	2	red	brick	1	1
chain	saw	24	2	repair	man	9	3
chalk	board	3	1	rest	room	27	2
checker	board	6	1	rib	cage	1	1
cheek	bone	8	1	rice	field	1	1
cheese	cake	29	4	river	bank	9	3
chick	pea	7	1	river	dale	1	1
chimney	hill	2	2	road	block	21	1
choir	boy	4	1	road	map	5	4
city	folk	3	1	road	runner	2	2
city	life	2	3	road	show	4	2
class	work	1	2	road	test	1	1
cliff	top	2	1	roller	blade	2	3
clock	tower	1	1	roof	line	1	1
clothes	line	5	1	row	boat	11	1
coal	field	3	1	row	house	1	2
cod	piece	6	1	rule	book	1	1
coffee	maker	2	6	rule	making	19	5
computer	world	14	4	sales	clerk	1	2
congress	person	2	1	sales	person	21	3
cop	car	1	2	salt	box	2	1
copper	mine	2	2	sand	bar	3	2
copy	writer	6	1	sand	castle	2	2
corn	bread	2	8	sand	dollar	1	2
corn	field	23	1	sand	pile	9	4
corn	flake	1	2	sauce	pan	17	1
corn	meal	2	4	school	book	1	2
corn	starch	16	2	school	boy	22	3

Table 3d (cont.)

W1	W2	CLOSED	OPEN	W1	W2	CLOSED	OPEN
corner	shop	1	1	school	child	27	1
council	member	2	12	school	day	1	10
council	woman	9	1	school	girl	10	6
counter	top	2	1	school	house	15	2
course	work	4	3	school	kid	3	24
cover	line	1	3	school	work	8	1
crab	grass	8	1	school	yard	17	3
crawl	space	1	2	science	center	1	2
crew	member	4	14	score	card	15	1
crop	duster	4	1	scout	master	11	2
crow	bar	3	1	scrap	book	6	1
cuff	link	2	1	sea	bed	3	1
curve	ball	19	4	sea	bird	4	3
dash	board	27	1	sea	coast	5	3
data	bank	2	1	sea	floor	1	1
data	point	1	7	sea	hawk	21	1
data	set	1	1	sea	life	2	1
day	lily	2	4	sea	port	12	1
day	trip	4	19	sea	shell	6	4
dinner	time	6	5	sea	shore	11	1
dirt	ball	1	1	see	page	2	1
dog	fight	2	1	seed	bed	2	1
dog	house	4	1	service	man	19	2
doll	house	7	2	service	member	1	16
door	bell	7	1	service	person	1	8
door	knob	8	2	sex	orgy	1	2
dream	boat	4	1	share	space	1	1
dream	life	1	7	sheep	dog	3	2
dream	world	1	5	sheet	rock	4	6
drill	hole	4	1	shirt	sleeve	3	2
drive	shaft	1	2	shoe	box	1	1
drive	train	3	3	shop	keeper	22	1
drug	lord	1	19	shoulder	blade	1	1
drug	maker	4	14	show	biz	13	2
dump	site	1	2	show	girl	17	1
dust	bowl	1	2	show	piece	4	1
dust	buster	1	1	show	tune	1	3
ear	muff	3	1	side	arm	1	4
egg	nog	2	1	side	car	1	2
egg	shell	7	2	side	light	5	1
entrance	way	3	2	side	show	18	4
entry	point	2	2	side	trip	1	2
entry	way	10	3	side	wall	6	1
fact	sheet	1	7	sight	line	2	1
falling	water	1	1	silica	gel	1	3
fan	base	1	6	sing	song	1	1
fantasy	land	2	1	skin	flint	2	1
fashion	sense	4	2	sky	box	5	1
feast	day	2	4	sky	diving	4	2
fig	tree	1	3	sleaze	bag	2	1
film	center	1	1	smoke	screen	2	2
film	star	1	8	smoke	stack	12	4
fire	ant	1	5	snake	pit	2	1
fire	cracker	11	1	snap	dragon	1	1
fire	fighting	1	1	snow	ball	10	2
fire	hose	1	1	snow	bird	3	1
fire	house	13	1	snow	man	6	1
fire	recipe	1	1	snow	melt	4	1
fire	storm	26	1	snow	plow	1	3
fire	truck	2	4	snow	storm	17	3
fire	wheel	8	1	soap	box	8	5
fish	net	1	1	soap	dish	4	1

Table 3d (cont.)

W1	W2	CLOSED	OPEN	W1	W2	CLOSED	OPEN
fish	pond	1	1	song	writer	24	4
flag	pole	3	1	sound	bite	6	6
flash	card	2	1	sound	card	1	1
flash	point	2	4	sound	stage	3	1
flood	water	2	7	source	book	1	5
floor	show	2	2	south	side	2	12
flow	chart	2	2	space	craft	31	1
flower	bed	5	29	space	suit	4	1
folk	song	2	3	spark	plug	1	10
food	chain	1	20	speaker	phone	1	4
food	stuff	11	1	speech	writer	16	3
form	factor	2	2	speed	boat	5	1
game	board	1	1	spot	size	1	1
gang	banger	1	1	spring	creek	2	5
gang	gang	1	1	spring	time	16	8
ghost	writer	7	1	stair	master	8	4
gift	shop	1	14	star	man	1	1
glove	box	3	3	star	war	4	7
goal	post	4	2	state	house	3	2
god	sake	1	1	station	house	3	3
gold	mark	2	1	steak	house	13	7
gold	mine	6	5	steel	head	2	1
grape	vine	25	1	stick	figure	2	2
grave	site	2	1	stock	broker	20	2
green	field	24	1	stomach	ache	2	1
green	glass	1	2	stone	wall	15	9
green	light	4	2	store	house	6	2
green	man	1	3	story	time	1	2
greyhound	bus	1	1	straw	man	2	4
grill	work	5	1	street	sweeper	1	1
ground	stroke	2	2	string	field	1	3
ground	swell	13	1	stunt	man	1	1
ground	work	27	1	style	book	2	1
grounds	keeper	5	1	sugar	man	1	1
guest	room	1	3	summer	house	1	2
gulf	stream	6	1	sun	bathe	1	1
gun	battle	8	7	sun	belt	3	2
gun	show	1	1	sun	burn	6	1
hack	work	6	1	sun	roof	5	13
hail	storm	2	2	sun	room	2	5
hair	dryer	1	6	sun	screen	4	1
hair	spray	7	4	surf	board	9	1
hair	stylist	2	2	sweat	pant	8	4
half	way	12	4	sweat	shirt	22	12
hand	ball	4	1	table	top	6	2
hand	basket	3	2	tag	line	2	14
hand	print	3	1	tail	pipe	4	5
handle	bar	10	2	tail	wind	3	1
harbor	side	3	1	tank	wagon	1	3
hay	loft	1	1	tar	heel	2	3
head	count	1	1	tax	man	1	2
head	light	17	1	taxi	cab	2	4
head	phone	12	1	tea	cup	7	1
health	system	1	11	thumb	print	2	3
heart	beat	20	2	thunder	clap	2	1
heart	breaker	2	3	thunder	storm	27	1
heart	string	1	1	time	scale	23	2
heart	throb	14	1	time	share	2	5
help	desk	9	1	tin	man	1	1
hen	house	2	1	tooth	brush	9	1
hog	wash	7	1	touch	tone	3	2
home	boy	4	2	town	car	4	4

Table 3d (cont.)

W1	W2	CLOSED	OPEN	W1	W2	CLOSED	OPEN
home	builder	3	3	town	home	1	9
home	buyer	5	4	toy	box	1	1
home	maker	27	1	train	load	1	1
home	ownership	3	12	trap	door	2	1
home	style	4	2	trash	can	1	16
horse	meat	1	1	tree	house	1	6
horse	power	11	2	trouble	maker	20	4
horse	shoe	26	1	truck	stop	2	2
house	cat	1	2	tube	plate	1	1
house	dress	1	1	user	name	2	1
house	guest	7	1	video	camera	1	25
house	mate	7	2	video	cassette	4	1
house	plant	5	2	video	disc	4	2
iron	man	6	2	video	taping	5	1
jazz	man	1	1	waist	line	9	1
jig	saw	9	1	wait	staff	1	3
job	holder	1	1	wake	field	20	1
john	doe	1	10	war	game	1	4
joy	stick	2	1	war	plane	15	3
junk	yard	11	1	wash	tub	1	1
key	stroke	3	1	waste	form	11	2
key	word	14	3	water	bed	1	3
king	sport	2	1	water	color	28	3
kitchen	ware	5	2	water	content	1	3
knee	cap	7	2	water	course	10	2
knife	point	1	2	water	lily	1	4
lady	friend	2	4	water	line	1	3
lady	love	1	2	water	mark	2	6
lake	shore	4	7	water	park	4	4
lake	side	25	1	water	shed	28	2
lamb	chop	1	5	water	side	8	1
lamp	shade	2	1	water	slide	3	1
land	mass	5	2	water	sport	15	15
lap	dog	9	5	water	tower	1	2
law	breaker	4	1	wave	form	11	2
law	court	1	1	wave	function	3	13
law	making	3	1	weather	boy	1	1
law	man	2	1	weather	folk	1	1
leader	board	4	3	weather	man	6	8
leather	good	1	2	weather	vane	2	1
left	field	1	3	web	link	1	4
leg	work	8	2	web	surfer	2	13
life	boat	5	1	web	world	1	1
life	care	3	1	week	night	9	1
life	cycle	1	8	west	east	1	2
life	jacket	3	2	west	gate	3	1
life	line	14	1	west	lake	6	1
life	work	2	1	west	point	1	6
light	house	30	1	wheat	field	1	1
line	width	6	1	wheel	barrow	1	1
litter	box	1	12	white	house	4	3
loin	cloth	3	1	white	man	4	2
lover	man	1	1	whore	house	2	1
lunch	box	3	3	wind	storm	2	2
lunch	meat	2	3	window	sill	4	4
market	watch	4	1	wine	glass	1	3
marsh	land	6	1	wing	span	4	1
may	day	4	1	wolf	pack	3	1
meal	time	6	1	wood	bridge	2	1
meat	ball	16	2	wood	carving	4	1
meat	loaf	7	2	wood	chip	1	3
meat	packer	1	1	wood	worker	6	1

Table 3d (cont.)

W1	W2	CLOSED	OPEN	W1	W2	CLOSED	OPEN
meat	packing	1	5	word	play	10	7
middle	ground	1	2	work	bench	1	2
milk	truck	1	2	work	day	11	4
minute	man	2	4	work	group	1	2
money	bag	5	1	work	load	29	3
moon	walk	2	1	work	mate	1	2
morning	star	18	1	work	space	11	4
mother	board	14	3	work	world	1	1
mother	lode	4	1	working	man	1	2
mother	ship	2	1	worry	wart	3	1

Appendix B-2: English Regression Analyses

Tables 12 - 14 Phonological Feature Models

ANOVA						
Model		**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**
	Regression	65.611	4	16.403	116.151	.000
	Residual	99.136	702	.141		
	Total	164.746	706			
Predictors: (Constant), SYLL, STRESS, BSYLL, DOUBLEC						
Dependent Variable: PCLOSED						

Table 12a

Coefficients					
	Unstandardized Coefficients		**Standardized Coefficients**	**t**	**Sig.**
Model	**B**	**Std. Error**	**Beta**		
(Constant)	1.142	.072		15.888	.000
SYLL	-.306	.021	-.669	-14.720	.000
STRESS	.230	.039	.181	5.891	.000
BSYLL	.271	.064	.189	4.259	.000
DOUBLEC	-.199	.075	-.078	-2.636	.009
Dependent Variable: PCLOSED					

Table 12b

ANOVA						
Model		**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**
	Regression	2.745	2	1.372	38.153	.000
	Residual	25.323	704	.036		
	Total	28.068	706			
Predictors: (Constant), STRESS, BSYLL						
Dependent Variable: PHYPHEN						

Table 13a

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model	B	Std. Error	Beta		
(Constant)	.208	.018		11.432	.000
STRESS	-.168	.019	-.321	-8.735	.000
BSYLL	-4.523E-02	.022	-.076	-2.076	.038
Dependent Variable: PHYPHEN					

Table 13b

ANOVA					
Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	53.760	3	17.920	123.921	.000
Residual	101.661	703	.145		
Total	155.421	706			
Predictors: (Constant), SYLL, DOUBLEC, BSYLL					
Dependent Variable: POPEN					

Table 14a

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model	B	Std. Error	Beta		
(Constant)	-.377	.057		-6.661	.000
SYLL	.296	.021	.665	14.370	.000
DOUBLEC	.245	.076	.098	3.209	.001
BSYLL	-.185	.064	-.133	-2.881	.004
Dependent Variable: POPEN					

Table 14b

Tables 17 – 19 Distributional (Lexical) Feature Models

ANOVA						
Model		**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**
	Regression	110.285	10	11.029	140.941	.000
	Residual	54.461	696	.078		
	Total	164.746	706			
Predictors: (Constant), W2CTOKLF, W1CTYPLF, W1OTOKLF, W2OTOKLF, W2CTYPLF, W1HTOKLF, W2TOKFR, W2HTOKFR, TOTAL, W2TYPLF						
Dependent Variable: PCLOSED						

Table 17a

Coefficients					
	Unstandardized Coefficients		**Standardized Coefficients**	**t**	**Sig.**
Model	**B**	**Std. Error**	**Beta**		
(Constant)	.346	.045		7.654	.000
W2CTOKLF	5.115E-02	.011	.294	4.500	.000
W1CTYPLF	.153	.011	.382	14.509	.000
W1OTOKLF	-4.175E-02	.008	-.164	-4.935	.000
W2OTOKLF	-7.537E-02	.018	-.270	-4.172	.000
W2CTYPLF	8.512E-02	.024	.272	3.606	.000
W1HTOKLF	-3.613E-02	.011	-.108	-3.317	.001
W2TOKFR	-5.058E-05	.000	-.099	-2.975	.003
W2HTOKFR	-5.394E-04	.000	-.061	-2.615	.009
TOTAL	1.245E-04	.000	.051	2.286	.023
W2TYPLF	5.350E-02	.027	.142	1.979	.048
Dependent Variable: PCLOSED					

Table 17b

ANOVA						
Model		**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**
	Regression	11.589	9	1.288	54.462	.000
	Residual	16.479	697	.024		
	Total	28.068	706			
Predictors: (Constant), W1HTOKFR, W2TOKLF, W2HTOKLF, W2TYPFR, W2TYPLF, W1CTYPLF, W1HTYPLF, W1OTOKLF, W2CORPFR						
Dependent Variable: PHYPHEN						

Table 18a

Coefficients						
		Unstandardized Coefficients		**Standardized Coefficients**	**t**	**Sig.**
Model		**B**	**Std. Error**	**Beta**		
	(Constant)	.467	.027		17.405	.000
	W1HTOKFR	5.543E-04	.000	.302	8.761	.000
	W2TOKLF	-3.656E-02	.009	-.325	-4.138	.000
	W2HTOKLF	3.063E-02	.006	.207	5.138	.000
	W2TYPFR	4.921E-04	.000	.240	4.444	.000
	W2TYPLF	-5.896E-02	.014	-.379	-4.252	.000
	W1CTYPLF	-3.284E-02	.005	-.199	-5.985	.000
	W1HTYPLF	3.728E-02	.010	.204	3.915	.000
	W1OTOKLF	-2.692E-02	.005	-.257	-5.735	.000
	W2CORPFR	3.629E-06	.000	.096	2.422	.016
Dependent Variable: PHYPHEN						

Table 18b

ANOVA					
Model	**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**
Regression	105.971	10	10.597	149.151	.000
Residual	49.450	696	.071		
Total	155.421	706			
Predictors: (Constant), W2CTYPLF, W2OTOKLF, W1CTYPLF, W1OTOKLF, W1HTOKFR, W2CTOKLF, TOTAL, W2HTYPLF, W2TOKLF, W2HTOKFR					
Dependent Variable: POPEN					

Table 19a

Coefficients					
	Unstandardized Coefficients		**Standardized Coefficients**	**t**	**Sig.**
Model	**B**	**Std. Error**	**Beta**		
(Constant)	.203	.042		4.832	.000
W2CTYPLF	-8.447E-02	.020	-.278	-4.273	.000
W2OTOKLF	6.838E-02	.014	.252	4.770	.000
W1CTYPLF	-.121	.009	-.310	-12.859	.000
W1OTOKLF	7.383E-02	.006	.299	11.629	.000
W1HTOKFR	-3.698E-04	.000	-.086	-3.935	.000
W2CTOKLF	-4.339E-02	.011	-.257	-3.779	.000
TOTAL	-1.531E-04	.000	-.065	-2.953	.003
W2HTYPLF	-5.609E-02	.016	-.114	-3.459	.001
W2TOKLF	3.666E-02	.017	.138	2.192	.029
W2HTOKFR	4.048E-04	.000	.047	2.010	.045
Dependent Variable: POPEN					

Table 19b

Tables 20 - 22 All-Feature Models

ANOVA					
Model	**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**
Regression	112.386	11	10.217	135.614	.000
Residual	52.360	695	.075		
Total	164.746	706			

Predictors: (Constant), W2CTOKLF, W1CTYPLF, W1OTOKLF, W2OTOKLF, STRESS, W2CTYPLF, W2HTOKFR, SYLL, BSYLL, W1HTOKLF, TOTAL

Dependent Variable: PCLOSED

Table 20a

Coefficients					
	Unstandardized Coefficients		**Standardized Coefficients**	**t**	**Sig.**
Model	**B**	**Std. Error**	**Beta**		
(Constant)	.610	.070		8.720	.000
W2CTOKLF	3.878E-02	.011	.223	3.457	.001
W1CTYPLF	.128	.011	.321	11.275	.000
W1OTOKLF	-3.569E-02	.008	-.141	-4.207	.000
W2OTOKLF	-5.168E-02	.007	-.185	-7.308	.000
STRESS	.116	.030	.092	3.873	.000
W2CTYPLF	8.104E-02	.020	.259	4.080	.000
W2HTOKFR	-6.465E-04	.000	-.073	-3.223	.001
SYLL	-8.474E-02	.018	-.185	-4.672	.000
BSYLL	.146	.047	.102	3.114	.002
W1HTOKLF	-3.410E-02	.011	-.102	-3.152	.002
TOTAL	1.263E-04	.000	.052	2.365	.018

Dependent Variable: PCLOSED

Table 20b

ANOVA						
Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	12.032	11	1.094	47.404	.000
	Residual	16.036	695	.023		
	Total	28.068	706			

Predictors: (Constant), W1HTOKFR, W2TOKLF, W2HTOKLF, STRESS, W2TYPFR, W2TYPLF, W1HTOKLF, W2OTYPFR, W1OTOKLF, W1CTYPLF, W2CORPFR

Dependent Variable: PHYPHEN

Table 21a

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model	B	Std. Error	Beta		
(Constant)	.507	.028		18.197	.000
W1HTOKFR	4.926E-04	.000	.268	7.194	.000
W2TOKLF	-3.201E-02	.009	-.284	-3.513	.000
W2HTOKLF	3.080E-02	.006	.208	5.171	.000
STRESS	-6.814E-02	.017	-.130	-4.102	.000
W2TYPFR	3.057E-04	.000	.149	1.985	.048
W2TYPLF	-6.028E-02	.015	-.388	-4.109	.000
W1HTOKLF	2.796E-02	.008	.203	3.705	.000
W2OTYPFR	3.316E-04	.000	.122	1.757	.079
W1OTOKLF	-2.979E-02	.005	-.284	-6.236	.000
W1CTYPLF	-2.502E-02	.006	-.151	-4.377	.000
W2CORPFR	3.000E-06	.000	.079	2.007	.045

Dependent Variable: PHYPHEN

Table 21b

ANOVA						
Model		**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**
	Regression	107.329	12	8.944	129.069	.000
	Residual	48.092	694	.069		
	Total	155.421	706			
l Predictors: (Constant), W2CTYPLF, W2OTOKLF, W1CTYPLF, W1OTOKLF, SYLL, W1HTOKFR, TOTAL, W2HTYPFR, W2HTOKFR, W2CTOKLF, DOUBLEC, BSYLL						
Dependent Variable: POPEN						

Table 22a

Coefficients					
	Unstandardized Coefficients		**Standardized Coefficients**	**t**	**Sig.**
Model	**B**	**Std. Error**	**Beta**		
(Constant)	1.347E-02	.064		.210	.834
W2CTYPLF	-7.882E-02	.019	-.259	-4.125	.000
W2OTOKLF	8.050E-02	.007	.296	11.586	.000
W1CTYPLF	-.102	.010	-.261	-9.873	.000
W1OTOKLF	6.957E-02	.006	.282	10.878	.000
SYLL	7.150E-02	.017	.161	4.124	.000
W1HTOKFR	-3.784E-04	.000	-.088	-4.083	.000
TOTAL	-1.486E-04	.000	-.063	-2.900	.004
W2HTYPFR	-4.491E-03	.001	-.090	-3.508	.000
W2HTOKFR	5.221E-04	.000	.061	2.541	.011
W2CTOKLF	-2.761E-02	.011	-.164	-2.561	.011
DOUBLEC	.121	.053	.049	2.285	.023
BSYLL	-8.874E-02	.045	-.064	-1.974	.049
Dependent Variable: POPEN					

Table 22b

BIBLIOGRAPHY

Abercrombie, D. (1967) Elements of General Phonetics. Chicago: Aldine Publishing Company.

Anderson, Stephen (1992) A-morphous Morphology. Cambridge: Cambridge University Press.

Anttila, A. T. (1997) Variation in Finnish phonology and morphology, doctoral dissertation, Stanford University.

Aronoff, M. (1994) Morphology by Itself. Cambridge, MA: MIT Press.

Aronoff, M. (1976) Word Formation in Generative Grammar. Cambridge,MA: MIT Press.

Balota, D., Pilotti, M., and Cortese, M. (2001) "Subjective frequency estimates for 2,938 monosyllabic words" in Memory and Cognition, 29 (4), pp.639-647.

Bauer, L. (2001) Morphological Productivity. Cambridge: Cambridge University Press.

Bauer, L. (1998) When is a sequence of two nouns a compound in English?" in English Language and Linguistics, 2 (1), pp.65-86.

Bauer, L. (1983) English Word-formation. Cambridge: Cambridge University Press.

Bertran, A.P. (1999) "Prosodic Typology: On the Dichotomy between Stress-Timed and Syllable-Timed Languages" Language Design, 2, pp.103-130.

Biber, D., Conrad, S., Reppen, R. (1998) Corpus Linguistics: Investigating language structure and use. Cambridge: Cambridge University Press.

Bloomfield, L. (1933) Language. London: Allen and Unwin.

Bock, J.K. (1986) "Syntactic Persistence in Language Production" in Cognitive Psychology, 18, pp.355-387.

Booij, G. (1985) "Coordination reduction in complex words: a case of prosodic phonology" in H. van der Hulst and NSH Smith (eds.), Advances in non-linear phonology. Dordrecht: Foris, 143-160.

Booij, G. (2004) "Compounding and derivation: evidence for Constructional Morphology" in W. U. Dressler, P. Kastovsky, & F. Rainer (eds.) Morphologia 2004. Amsterdam: Benjamins, pp.109-132.

Burstein, J. (1992) The Stress and Syntax of Compound Nominals. CUNY Dissertation.

Catford, J.C. (1988) A Practical Introduction to Phonetics. New York: Oxford University Press.

Chambers, J.K. (2000) "Region and Language Variation" in English World-Wide 21, pp.1-31.

Chomsky, N. and Halle, M. (1968) The Sound Pattern of English. New York: Harper & Row.

Clark, Eve (1993) The Lexicon in Acquisition. Cambridge: Cambridge University Press.

Clark, J. and Yallop, C. (1995) An Introduction to Phonetics and Phonology,2nd ed. Oxford: Blackwell.

Collinder, B. (1969) Survey of the Uralic Languages. Stockholm: Almqvist and Wiksell Förlag.

de Jong, N.H., Feldman, L., Schneider, R., Pastizzo, M., Baayen, H. (2002) "The Processing and Representation of Dutch and English Compounds: Peripheral Morphological and Central Orthographic Effects" in Brain and Language 81, pp.555-567.

DiSciullo, A.M. and Williams, E. (1987) On the Definition of Word. Cambridge: MIT Press.

Eek, A. and Meister, E. (1997) "Some perception Experiments on Estonian Word Prosody: Foot Structure vs. Segmental Quantity" in Lehiste and Roos (eds.), Proceedings of the International Symposium on Estonian Prosody, Tallinn, Estonia, Oct.1996, pp.71-99.

Erelt, M., R. Kasik, H. Metslang, H. Rajandi, K. Ross, H. Saari, K. Tael & S. Vare. (1995). Eesti keele grammatika: morfoloogia. Tallinn: Eesti Teaduste Akadeemia Eesti Keele Instituut.

Fabb, N. (1998) "Compounding" in Spencer and Zwicky, The Handbook of Morphology. Oxford: Blackwell Publications, pp.66-83.

Fromkin, V., Rodman, R., and Hyams, Nina (2003) An Introduction to Language, 7th ed. Boston: Thompson-Heinle.

Gordon, Matthew (1997) "The Phonetic Correlates of Stress and the Prosodic Hierarchy in Estonian" in Lehiste and Roos (eds.), Proceedings of the International Symposium on Estonian Prosody, Tallinn, Estonia, Oct.1996, pp.100-123.

Hayes, Bruce (1995) Metrical Stress Theory: Principles and Case Studies. Chicago: University of Chicago Press.

Hedlund, T. (2002) "Compounds in dictionary-based cross-language information retrieval" Information Research, **7**(2) (Available at http://InformationR.net/ir/7-2/paper128.html).

Hint, M. (1997) "The Estonian Quantity Degrees in Prosody and Morphophonology" in Lehiste and Roos (eds.), Proceedings of the International Symposium on Estonian Prosody, Tallinn, Estonia, Oct.1996, pp.125-134.

Hint, M. (1998) Häälikutest Sõnadeni. Tallinn: Eesti Keele Sihtasutus.

Hume, E. and Johnson, K. (2001) "A Model of the Interplay of Speech Perception and Phonology" in E. Hume and K. Johnson (eds.) The Role of Speech Perception in Phonology. San Diego, CA: Academic Press, pp.3-26.

Ide, N., Reppen, R. and Suderman, K. (2003) ANC First Release. Philadelphia: Linguistic Data Consortium.

Jones, Daniel, edited by P. Roach, J. Hartman, and J. Setter (2003) (Cambridge Pronouncing Dictionary, 16th edition. Cambridge: Cambridge University Press.

Kaalep, H. and Muischnek,K. (2002) Eesti Kirjakeele Sagedussõnastik. Tartu: TÜ kirjastus.

Kaalep, H. (1997) "An Estonian Morphological Analyser and the Impact of a Corpus on its Development" Computers and the Humanities, 31, pp.115-133.

Kager, R. (2001) Optimality Theory. Cambridge University Press.

Kiparsky, P. (2005) "Grammaticalization as optimization" (Draft), available at www.stanford.edu/~kiparsky/.

Krull, D. (1999) "Foot Isochrony in Estonian" in Proceedings of the XIVth ICPhS, August 1-7, San Francisco, CA.

Kučera, H. and Francis, W.N. (1967) Computational Analysis of Present-Day American English. Providence: Brown University Press.

Lapata, M. and Lascarides, A. (2003) Detecting Novel Compounds: The Role of Distributional Evidence" in Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, pp.235-242.

Lehiste, I. (1977) "Isochrony Revisited" Journal of Phonetics v.5, pp.253-263.

Lehiste, I. (1997a) "The Search for Phonetic Correlates in Estonian Prosody" in Lehiste and Roos (eds.), Proceedings of the International Symposium on Estonian Prosody, Tallinn, Estonia, Oct.1996, pp.11-35.

Lehiste, I. (1997b) "The Structure of Trisyllabic Words" in Lehiste and Roos (eds.), Proceedings of the International Symposium on Estonian Prosody, Tallinn, Estonia, Oct.1996, pp.149-164.

Lehiste, I. (2004) "Prosody in Speech and Singing", a paper presented at the 2004 Speech Prosody Conference in Nara, Japan.

Lieber, R. (1983) "Argument Linking and Compounding in English" Linguistic Inquiry, 14, pp.251-286.

Manning, C. and Schütze, H. (2000) "Collocations", ch.5 in Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press, pp.151-190.

Marchand, H. (1969) The Categories and Types of Present-day English Word Formation. München: Beck.

McCarthy, J. (2001) A Thematic Guide to Optimality Theory. Cambridge University Press.

McCarthy, J. and Prince, A. (1995) "Prosodic Morphology" in Goldsmith, J. (ed.) The Handbook of Phonological Theory. Cambridge: Blackwell Publishers, pp.318-366.

Mellenius, I. (1997) The Acquisition of Nominal Compounding in Swedish. Lund, Sweden: Lund University Press.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K.J. (1990) "Introduction to WordNet: an on-line lexical database" in International Journal of Lexicography 3 (4), pp. 235 - 244.

Mürk, H. (1997) A Handbook of Estonian: Nouns, Adjectives and Verbs. Bloomington:Indiana University Research Institute for Inner Asian Studies.

Neef, M. (2002), The Reader's View: Sharpening in German" in Neef, M. Neijt, A., and Sproat, R. (eds.) The Relation of Writing to Spoken Language. Tubingen: Max Niemeyer Verlag, pp.169-192.

Neef, M., Neijt, A., and Sproat, R. (2002), "Introduction" to Neef, M. Neijt, A., and Sproat, R. (eds.) The Relation of Writing to Spoken language. Tubingen: Max Niemeyer Verlag, pp.1-10.

Neijt, A. (2002) "The Interfaces of Writing and Grammar" in Neef, M. Neijt, A., and Sproat, R. (eds.) The Relation of Writing to Spoken language. Tubingen: Max Niemeyer Verlag, pp. 11-34.

Nesselhauf, N. (2004) "What are collocations?" in Phraseological Units: basic concepts and their application. Basel: Schwabe, pp.1-21.

Nishimoto, E. (2004) A Corpus-Based Delimitation of New Words: Cross-segment comparison and morphological productivity, CUNY Dissertation.

Noack, C. (2002) Regularities in German Orthography: A Computer-Based Comparison" in Neef,M., Neijt,A., and Sproat,R. (eds.) The Relation of Writing to Spoken Language. Tubingen: Max Niemeyer Verlag, pp.149-168.

Odden, D. (1997) "Some Theoretical Issues in Estonian Prosody" in Lehiste and Roos (eds.), Proceedings of the International Symposium on Estonian Prosody, Tallinn, Estonia, Oct. 1996, pp.165-194.

Oinas, F. (1975) Basic Course in Estonian, 4th ed. Bloomington: Indiana University.

The Oxford American Dictionary and Language Guide (1999). New York: Oxford University Press.

Pickering, M. and Branigan, H. (1999) "Syntactic Priming in Language Production" in Trends in Cognitive Sciences, vol.3, no.4, pp.136-141.

Pinker, S. (2000) Words and Rules: The Ingredients of Language. New York: Perennial Books.

Prince, Alan (1980) "A Metrical Theory for Estonian Quantity," Linguistic Inquiry vol.11, no.3, pp.511-562.

Roach, P. (1982) "On the distinction between 'stress-timed' and 'syllable-timed languages'" in D. Crystal (ed.), Linguistic Controversies. London: Edward Arnold, pp.73-79.

Ryder, M. E. (1994) Ordered Chaos: The Interpretation of English Noun-Noun Compounds. Berkeley and Los Angeles, CA: University of California Press.

Saagpakk, P. (1955) "A Grammatical Survey of the Estonian Language", preface to an Estonian-English Dictionary. New York: Nordic Press, pp.XLV-LXI.

Sadock, J. (1998) "On the Autonomy of Compounding Morphology" in Lapointe, S., Brentari, D., and Farrell, P.M. (1998) Morphology and its Relation to Phonology and Syntax. Stanford, CA: CSLI Publications, pp. 161-187.

Silvet, J. (ed.) (1964) Eesti Inglise Sõnaraamat Toronto: ORTO.

Sproat, R. (1992) Morphology and Computation. Cambridge, MA: MIT Press.

Sproat, R. (2000) A Computational Theory of Writing Systems. AT&T Labs Research.

Tauli, V. (1973) Standard Estonian Grammar, Part I: Phonology, Morphology, Word-Formation. Uppsala University.

Tatham, M. and Morton, K. (2001) "Intrinsic and Adjusted Unit Length in English Rhythm Synthesis" in Proceedings of the Institute of Acoustics – WISP 2001, St Albans: Institute of Acoustics, pp.189-200.

ten Hacken, P. (2004) "What are Compounds?" in Phraseological Units: basic concepts and their application. Basel: Schwabe, pp.53-66.

ten Hacken, P. (1999) "Motivated Tests for Compounding" in Acta Linguistica Hafniensia, v.31, pp.27-58.

ten Hacken, P. (1994) Defining Morphology: A Principled Approach to Determining the Boundaries of Compounding, Derivation, and Inflection. Hildesheim: Olms, pp.1-143.

U.S. Department of Energy (1991) "Department of Energy abstracts", files 1-7, part of the ACL Data Initiative. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Vago, R. (1980) The Sound Pattern of Hungarian. Washington, D.C.: Georgetown University Press.

Viks, Ülle (1994a) Eestikeele Klassifikatoorne Morfoloogia. Tartu: Tartu Ülikooli Kirjastuse Trükikoda, pp.23-31.

Viks, Ülle (1994b) "A Morphological Analyzer for the Estonian Language: The Possibilities and Impossibilities of Automatic Analysis" in Viks, Ü. (ed.) Automatic Morphology of Estonian I: Research Report. Tallinn: Institute of the Estonian Language.

Volk, M. (2002) "Using the Web as Corpus for Linguistic Research" in Renate Pajusalu and Tiit Hennoste (eds.): Tähendusepüüdja. *Catcher of the Meaning:. A Festschrift for Professor Haldur Õim.* Publications of the Department of General Linguistics 3. University of Tartu) volk@ifi.unizh.ch

Von Schneidemesser, L. (1996) "Soda or Pop?" in Journal of English Linguistics, vol.24, no.4, pp.270-287.

Zipf, G.K. (1935) The Psycho-Biology of Language. Cambridge: MIT Press.

Websources:

Google (www.google.com).

Kerstens, J., Ruys, E., Zwarts, J. (1996-2001) Lexicon of Linguistics. Utrecht, The Netherlands: Utrecht Institute of Linguistics (http://www2.let.uu.nl/UiL)

McFedries, P. (1995-2005) The Word Spy (www.wordspy.com)

Miller, J., (ed.) (2005) "A Collection of Word Oddities and Trivia", a Yahoo! group (http://members.aol.com /gulfhigh2/words11.htm).

Morris, E. (1995-2005) The Word Detective (www.word-detective.com).

MS Encarta Dictionary (http://encarta.msn.com/encnet/features/dictionary).

Wikipedia, The Free Encyclopedia. (http://en.wikiquote.org/wiki/Floccinaucinihilipilification)

Wikipedia, The Free Encyclopedia. (http://en.wikipedia.org/wiki/Brown_Corpus)

WordNet 2.1 (2005) Princeton: Princeton University Cognitive Science Laboratory. (http://wordnet.princeton.edu/perl/webwn)