A CORPUS-BASED DELIMITATION OF NEW WORDS: CROSS-SEGMENT COMPARISON AND MORPHOLOGICAL PRODUCTIVITY

by

EIJI NISHIMOTO

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

2004

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

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

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

April 20, 2004
Date

Dr. Dianne Bradley
Chair of Examining Committee

April 20, 2004
Date

Dr. Gita Martohardjono
Executive Officer

Dr. Martin Chodorow
Dr. Virginia Teller
Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

A CORPUS-BASED DELIMITATION OF NEW WORDS:
CROSS-SEGMENT COMPARISON AND
MORPHOLOGICAL PRODUCTIVITY

by

Eiji Nishimoto

Adviser: Dr. Dianne Bradley

The dissertation explores methods of identifying new words in a large corpus of texts, the British National Corpus (BNC) of 100 million English words, and of assessing productivity in derivational affixation. Adopting a smoothing technique, deleted estimation, from the Language Technology literature, we show that new words can be detected when segments of a corpus are cross-compared to find which word types are shared (or unshared). When each corpus segment is created so as to reflect a set of words used by a group of randomly sampled speakers, through a randomization respecting document boundaries, the cross-comparison of corpus segments can be interpreted as revealing the usage distribution of words across groups of speakers. A word shared by fewer corpus segments is more limited in its usage commonality and thus a more likely candidate for a new word. Morphological productivity, the potential of a word formation process involving an affix to form a new word, is assessed for 12 English derivational

suffixes (nominal *-ness*, *-ity*, *-er*, *-ee*, *-ion*, *-ment*, and *-th*; verbal *-ize* and *-ify*; adjectival *-ish* and *-ous*; adverbial *-ly*), based on new words identified in the BNC via deleted estimation. Quantifying the usage distribution of new word types across corpus segments opens many possibilities for assessing the productivity of affixes. Cross-comparing as few as two corpus segments offers a crude yet computationally simple method of separating new words (unshared) from non-new words (shared), to yield a productivity index for a given affix. Cross-comparing as many as six corpus segments supports a graded definition of a word's newness (words shared by fewer corpus segments being more likely new) and thereby a more detailed characterization of the productivity of affixes. The proposed methods of identifying new words and assessing productivity are shown to offer valuable insights into the issue of productivity in word formation.

Acknowledgments

I would like to thank my dissertation committee members for their continuous support for the present research. My adviser Dianne Bradley has always shared excitement in this research with me; without her insightful comments and help with analyzing data, this research would not have gained its current shape. Martin Chodorow has provided me with the backbone of this research; without his introduction to corpus studies, Perl programming, and statistical techniques, the data of the present research could not have been collected. Virginia Teller has encouraged me throughout my doctoral study; without her introduction to computational linguistics, this research would not have started.

I would also like to thank Harald Baayen and Mark Aronoff for their insightful comments on early developments of the present research. This research has been inspired by a series of studies done by Harald Baayen and his colleagues.

I have also benefited from my involvement in the learnability project at CUNY. William Sakas taught me many important principles of programming, and Janet Dean Fodor gave me many intellectual challenges that polished my academic skills.

I would also like to thank Woan Jen Liing for her help with Chinese data. My thanks also go to my long-time friends, Carrie Crowther, Alison Gabriele, Tomoyuki Yabe, Mary Sepp, Hope Cotton, and many others.

Finally, I would like to thank my wife Xiaotong Dong for her love and support.

The present research was funded by the Mario Capelloni Dissertation Fellowship granted by the Graduate Center of the City University of New York.

Table of Contents

List of Tables

List of Figures

List of Notations

V	Total number of word types with an affix
V_N	Number of new word types with an affix
V_{NN}	Number of non-new word types with an affix
P_{DE}	Productivity based on the deleted estimation method, the ratio of V_N to V
UC	Usage Commonality; the usage distribution of words across corpus segments
Sr	Number of word types with an affix that appear in r corpus segments ($1 \leq r \leq n$, given n corpus segments)
Ssum	Sum of all Sr data
%Sr	Percentage of Sr in Ssum
BNC	British National Corpus
WBP	The "Written Books and Periodicals" component of the BNC
OED	Oxford English Dictionary
WD	Webster's Third New International Dictionary

Chapter 1 Productivity in Word Formation

Although the notion of *morphological productivity* has been central to the study of word formation (Aronoff, 1976: 35; Bauer, 1988: 57; Plag, 1999: 6), productivity continues to be a contested issue (Bauer, 1983: 62; 2001: 1) that defies a solid, uniform description. In recent years, the morphology literature has seen a wealth of research, such as that of Bauer (2001) and Plag (1999), that are devoted specifically to investigating the issue of productivity. Bauer (2001), in particular, gives a comprehensive overview of the subject matter. In this introductory chapter, we will discuss what phenomenon researchers are interested in describing and what the aim of the present research is.

1.1 Target Phenomenon

Morphological productivity concerns word formation processes involving affixes that lend themselves to the coinage of new words. Although no uniform definition of productivity has been obtained in the literature (see Bauer, 2001; Plag, 1999), the consensus among researchers seems to be that "new" words are central to the discussion of productivity. The term *new word* will occur throughout the present research.

In any language, speakers use not only words that they are familiar with but also words that are newly coined. Most new words are understood by speakers without great difficulty. Thus, not only existing words but also new words are a target of morphological studies, and morphological theory must account for the rule-governed ways in which new words are formed.

Productive coinages of words are common in our daily use of language: we may find them in the public media (e.g., TV, radio, newspapers, and the *Web*[1]) and in our ordinary conversations. A person who is being gossiped about may be referred to as a *gossipee*, or a used book may be *cleanish*. If a play on stage is *wowable*, we may discuss the *wowability* of that play. Not all of these productively coined words will be "new" to all speakers (i.e., some may have been heard or used before), but these words are potentially new to the majority of English speakers. In the present study, we will examine the coinage of new words in English. Unless otherwise stated, all examples are drawn from English.

Discussion of productivity centers primarily on derivational affixation (Aronoff, 1976; Bauer, 1988), where an affix (mainly a prefix or suffix, in English) attaches to a *base* to form a word. Some examples are[2]:

(1) a. *chic* (adjective) + *-ness* (nominal suffix) → *chicness* (noun)

b. *fiction* (noun) + *-ish* (adjectival suffix) → *fictionish* (adjective)

c. *awful* (adjective) + *-ize* (verbal suffix) → *awfulize* (verb)

d. *beige* (adjective) + *-ly* (adverbial suffix) → *beigely* (adverb)

Word formation processes not involving affixation are commonly excluded from the discussion of productivity (Aronoff, 1976: 20; Bauer, 1988: 33, 39). Those that are

---

[1] We will use the term *Web* to refer to the *World Wide Web* (i.e., the *Internet*).

[2] Although these examples involve a change in the lexical category (noun, adjective, verb) of the base, there are also cases where the lexical category of the base does not change: for example, *full* (adjective) + *-ish* (adjectival suffix) → *fullish* (adjective).

excluded are *blends* (e.g., *smoke + fog → smog*), *acronyms* (e.g., *North Atlantic Treaty Organization → NATO*), and *clippings* (e.g., *delicatessen → deli*). Unlike derivational affixation, these processes lack regularity, and it is difficult to predict how new words will be formed. It is even questionable whether the processes underlying blends, acronyms, and clippings properly belong to the domain of morphology (Bauer, 1988: 33, 39; Haspelmath, 2002: 25). *Compounding* (e.g., *greenhouse*) is a productive process, but the discussion of its productivity enters only peripherally into a discussion of productivity in derivational affixation.

Accounting for the phenomenon of productivity is complicated by the many factors involved. One factor, for example, that contributes to the difficulty of obtaining a uniform definition of productivity is that there is a disagreement among researchers as to just what is being productive: it could be a process, an affix, or a rule that underlies word formation (Bauer, 2001: 12–15). Depending on the view of what is involved in productive coinages of words, a different definition of productivity may be obtained.

The aspect of productivity that most intrigues researchers is the fact that some word formation processes more easily lead to the coinage of words than others; that is, affixes differ in their *degree of productivity* (Aronoff, 1976: 35–45; Bauer, 2001: 125–162). It is not the case that the word formation process for a given affix can be used freely to form any word. For example, although the suffix *-ment* attaches to many verbs to form a noun, it is not the case that *-ment* attaches to any verb:

(2) a. *settle + -ment → settlement*

b. *accomplish + -ment → accomplishment*

c. *provide* + *-ment* → **providement*

d. *express* + *-ment* → **expressment*

There are words that are possible and not possible, and some affixes can be used to form more possible words than others (Aronoff, 1976). The degree of productivity will be discussed in the next section.

Although views may differ as to what is involved in productive coinages of words, the most fundamental element underlying the study of productivity seems to be the potential for the coinage of new words. Bauer (2001: 41), for example, notes: "Productivity is all about potential. A process is productive if it has the potential to lead to new coinages, or to the extent to which it does lead to new coinages." What we will pursue in the present study is what exactly that potential is, how it relates to the coinage of new words, and how it could be described. In particular, the present study investigates the quantification of productivity based on a large corpus of texts. The corpus-based study of morphological productivity has recently been made prominent, notably by Baayen (1989, 1992, 2001) and his colleagues (e.g., Baayen & Lieber, 1991; Baayen & Renouf, 1996).

1.2 Concepts and Terms

In this section, we will review some concepts and terms involved in discussion of productivity that are relevant to the present study.

1.2.1 Degrees of Productivity

How productivity varies among affixes is well illustrated by rival suffixes *-ness* and *-ity* in English. Aronoff (1976: 35–45) discusses varying degrees of productivity between *-ness* and *-ity* as they attach to a base that ends in *-ous*. Given *-ous* bases, words formed with *-ness* are generally preferred over words formed with *-ity* (Aronoff, 1976: 37):

(3) a. *fabulousness* > *fabulosity*

b. *dubiousness* > *dubiety*/*dubiosity*

Here, the symbol > indicates which word is apparently more preferred. There are also word formations with *-ity* that are not possible with *-ous* bases (Aronoff, 1976: 38):

(4) a. *acrimonious* → *acrimoniousness*, **acrimoniosity*

b. *euphonious* → *euphoniousness*, **euphoniosity*

c. *famous* → *famousness*, **famosity*

One of the accounts that Aronoff (1976: 38–39) offers as to why differences in the degree of productivity arise is *semantic coherence*. The meaning of a derived word with *-ness* is predictable from the word formation, and is semantically coherent with the meaning of the base. By contrast, the meaning of a derived word with *-ity* is often idiosyncratic and cannot be easily predicted. Another account of the difference in the degree of productivity, *blocking*, will be discussed in the next section.

If we consider degrees of productivity to constitute a continuum on a scale of productivity, *-ness* is expected to be on the productive end of the scale, while the suffix *-th* is expected to be on the unproductive end. The suffix *-th* has long been unsuccessful in coining a new word that survives, despite attempts at terms like *coolth* (Aronoff & Anshen, 1998: 243).

The use of the term *productive* varies between two senses: in the discussion of *-ness* and *-ity*, for example, one affix is said to be more productive than the other. In the discussion of *-ness* and *-th*, on the other hand, affixes are said to be productive or not productive at all.

Bauer (2001: 205–211) proposes a division of the ambiguous notion of *productivity* into two senses: *availability* and *profitability*. *Availability* refers to the question whether a word formation process is available at all for the coinage of a new word. The word formation process for *-th* is apparently no longer available, because it does not coin any new word, while the word formation processes for *-ness* and *-ity* are available because they both coin new words. *Profitability* refers to the extent with which a given word formation process lends itself to the coinage of new words. The comparison of *-ness* and *-ity* is based on the profitability of these suffixes. The main focus of the present study can be said to be the profitability of affixes, in Bauer's terms.

As compared with inflectional processes, derivational processes are limited in that a given affix attaches to a range of bases that is restricted in some way, within a lexical category. Some researchers (e.g., Matthews, 1991: 69) regard derivational processes as being *semi-productive*.

The productive/semi-productive dichotomy between inflectional and derivational processes relates to views of the *Lexicalist Hypothesis* which, following Chomsky's (1970) proposal to separate morphology from syntax, consider inflection and derivation to take place in syntax and the lexicon, respectively (Anderson, 1982; Aronoff, 1994). Although the extent to which derivation is confined to the lexicon is a matter of debate (see Scalise, 1984), a distinction is commonly drawn in English between inflection and derivation (Aronoff, 1976, 1994). We will adhere to that distinction in what follows, and will discuss only clear cases of English derivation in the present study.

1.2.2 Restrictions/Constraints

Derivational processes are limited by various conditions. The term *restriction* has been traditionally used in the literature to describe a limiting condition on word formation, but the term *constraint* has recently gained more usage, with an advantage that *constraint* does not imply that a limiting condition is absolute (Bauer, 2001: 126). The use of the term constraint may or may not be associated with the views of *Optimality Theory* (*OT*, see Archangeli & Langendoen, 1997). Plag (1999), for example, proposes an output-oriented view of morphology and discusses constraints in the framework of OT. Bauer (2001), on the other hand, discusses constraints, independent of the OT framework. In general, the two terms *restriction* and *constraint* appear to be used interchangeably especially when no theoretical implication is intended (e.g., Plag, 2003: 60). We will consistently use the term *restriction*, without any intended nuance.

Restrictions on word formation are mainly classified into morphological, phonological, syntactic, semantic, pragmatic types (e.g., Bauer, 2001: 128–139; Plag,

2003: 61). It is clear from this classification that such restrictions are present at every level of linguistic representation. Since detailed analyses of various restrictions are available elsewhere (e.g., Aronoff, 1976; Bauer, 2001; Plag, 1999), only well-known examples of restrictions will be reviewed below. Additional restrictions will be discussed in later chapters as they become relevant in the discussion of particular affixes.

At the morphological level, the *Latinate Restriction* (Aronoff, 1976: 51–52; Plag, 1999: 57–60) requires the base of a derived word to have the [+Latinate] feature. The suffix *-ity* is subject to this restriction and attaches only to bases with the [+Latinate] feature,[3] whereas *-ness*, which is free of the restriction, attaches to a wide variety of both Latinate [+Latinate] and Germanic [−Latinate] bases. This point is exemplified in:

(5) a. *pure* [+Latinate] + *-ness* → *pureness*

b. *pure* [+Latinate] + *-ity* → *purity*

c. *clean* [−Latinate] + *-ness* → *cleanness*

d. *clean* [−Latinate] + *-ity* → **cleanity*

In cases such as *readability* and *knowability*, *read* and *know* have the [−Latinate] feature, but since *-able* is [+Latinate], the requirement that *-ity* only attaches to a [+Latinate] base is maintained (Aronoff, 1976: 52). These cases show that this restriction is pertinent at the morphological level.

The word formation process for *-ity* is also limited by a phenomenon called *blocking* (Aronoff, 1976: 43–45). The formation of a given word is blocked if an

[3] An exception that Aronoff (1976: 51) points out is *oddity*, where *odd* is [−Latinate].

equivalent word with an identical meaning already exists. Aronoff (1976: 44) shows that given *-ous* bases, many *-ity* words (but not *-ness* words) are blocked:

(6) a. *glorious* + *-ness* → *gloriousness*

b. *glorious* + *-ity* → **gloriosity* (cf. *glory*)

c. *furious* + *-ness* → *furiousness*

d. *furious* + *-ity* → **furiosity* (cf. *fury*)

According to Aronoff (1976: 45), no *-ness* word with an *-ous* base is blocked because *-ness* words are not stored in the lexicon and are formed as needed. On the other hand, *-ity* words are subject to storage in the lexicon, and the formation of an *-ity* word will be blocked if the slot in the lexicon for listing that word is already filled with an existing word (assuming that only one slot is available for each canonical meaning).

Blocking is known to be not absolute (Bauer, 2001: 137; Matthews, 1991: 77), and whether it occurs depends on semantic needs (Plag, 1999: 51); for instance, the use of *stealer* in *ten stealers* by Shakespeare in the sense of "ten fingers" is not blocked by the existing word *thief* (Bauer, 1988: 67). Blocking may also be overridden due to a temporary loss of memory of a particular word, as commonly seen in children (Aronoff & Anshen, 1998: 240); for example, *famousness* may be used when *fame* is temporarily forgotten.

1.2.3 Words, Lexemes and Word Forms

In discussing the coinage of a word, the term *word* is ambiguous among three senses (Bauer, 1988: 7–9; Matthews, 1991: 24–31): *word form*, *lexeme*, and *grammatical word*. Consider the word *recognize* in the following sentences:

(7) a. I *recognized* him immediately.

b. The film was *recognized* as the best of the year.

c. He *recognizes* me, too.

Orthographically (and phonologically), there are two distinct word forms in (7): one is *recognized* in (7a) and (7b), and the other is *recognizes* in (7c). These two word forms, although distinct on the surface, belong to the same lexeme RECOGNIZE.[4] A lexeme is an abstract basic entity of the lexicon and corresponds roughly to a headword of a dictionary; for example, it is RECOGNIZE, not *recognized* or *recognizes*, by which we search a dictionary for a word definition. The word form *recognized* in (7a) and (7b) can be separated into two grammatical words: *recognized* is "RECOGNIZE + past tense" in (7a), versus *recognized* is "*RECOGNIZE* + past participle" in (7b). Since *recognizes* in (7c) is "RECOGNIZE + third person singular present tense," we have three grammatical words in (7). In summary, there are two word forms and three grammatical words in (7), all belonging to one lexeme RECOGNIZE.

Now let us consider another example:

---

[4] For the purpose of disambiguating the term *word*, we will follow the convention of writing a lexeme in small capitalized letters in this section.

(8) The *recognition* of his work is widespread.

Here, *recognition* belongs to a lexeme RECOGNITION that is distinct from RECOGNIZE: the suffixation of *-ion* to lexeme RECOGNIZE forms another lexeme RECOGNITION. It is lexeme that we will associate with the term *word* in discussing new words in the present study. Any use of the term *word* in isolation can be replaced with *lexeme*. Thus, the coinage of new words that we seek can be rephrased as the coinage of new lexemes. The term *word form* will be used in referring to orthographical form variations (as we will examine words in texts).

Although the matter is not pursued in the present study, it must be noted that productivity in inflectional processes involves word forms. For example, in (9) there are three word forms using the English plural suffix *-s*.

(9) The *books* are given to both *teachers* and *students*.

We could discuss the productivity of *-s* that leads to three distinct word forms in (9). Pluralization by *-s* has high regularity, being applicable to the majority of nouns.[5] Inflectional processes in English are generally considered to be fully productive (Plag, 2003: 16; Scalise, 1984: 114), and they are often cited as parade examples of high productivity.

[5] Some exceptions are: *child* → *children*, *fish* → *fish*, and *goose* → *geese*.

1.2.4 Processes, Rules and Patterns

In statements of the form "the word formation process is productive for *-ness*," the use of the term *process* offers the advantage that it does not necessarily imply that an affix itself is the cause of productivity (Bauer, 2001: 12–13). However, in comparing affixes with respect to their productivity, an affix is commonly said to be productive (or unproductive), as in "*-ness* is productive while *-th* is not." Such an expression should be interpreted as a shorthand for "the word formation process is productive for *-ness*, while the word formation process is not productive for *-th*." In the research reported in this dissertation, our focus is on characterizing observed differences in productivity among affixes, rather than on what causes productivity (or unproductivity) in any instance.

The term *process* is also neutral with respect to the involvement of a rule in word formation. In the generative framework (e.g., Aronoff, 1976), regularity in word formation is traditionally captured in a *Word Formation Rule* (*WFR*). Under the rule-based view of word formation, productivity is discussed for the WFR of an affix. Bauer (2001: 13) sees little difference between discussions of processes and discussions of rules:

> The difference between saying that productivity is a feature of morphological processes and saying that it is a feature of rules seems to me to be largely a matter of the author's conception of the grammar. The rule is a precise statement of how the morphological process operates. Saying that the process is productive presupposes the reality of the process; saying that the rule is productive presupposes the reality of the grammar. It does not seem to me that there is any empirical difference between these two versions.

In his use of the term *process*, Bauer is simply taking a neutral position regarding the involvement of rules in word formation.

Another useful term in discussing word formation is *pattern*. In referring to, for example, words ending in *-ization* or *-ification*, the term *pattern* seems preferable to the terms *process* or *rule*; *pattern* neutrally accommodates the fact that multiple affixes are involved in *-ization* and *-ification*. Aronoff (1983: 166) sees word formation patterns as subcategories of the output of WFRs.

The term *pattern* is also associated with an interpretation in favor of an analogy-based view of morphology that denies the involvement of rules. Models invoking *connectionism* (or *Artificial Neural Networks*), following Rumelhart and McClelland (1986), have challenged the classic view of morphology which assumes the involvement of rules. There have been heated debates over whether rules are involved in the processes of past-tense inflection in English. Some researchers (e.g., Pinker & Prince, 1988) oppose the analogy-based view of connectionism by postulating that inflectional processes involve both rules (computation) for regular inflection and analogy (associative memory) for irregular inflection. On the other hand, researchers in favor of connectionism (e.g., MacWhinney & Leinbach, 1991) maintain that no rule is involved in inflectional processes.

Regarding the debate over rules versus analogy, an argument in favor of the rule-based approach is offered by Bauer (2001: 97):

> If we give up on rules immediately, and say that everything is analogy, we will never discover the hidden regularities. If we try to describe things in terms of rules, we always have analogy to fall back on. Rule-governed productivity may thus be a better research strategy than analogy, even if analogy, by allowing for greater variation in output, permits the researcher to present a more accurate picture of speaker behaviour.

Bauer suggests that the rule-based view is the conservative option because it does not abandon *a priori* a potentially useful means of depicting regularities in morphological processes.

Another use of the term *pattern* is seen in Bybee's (e.g., 1998) view of morphology, in which a word is derived based on a pattern arising from similarity among phonologically and semantically related words.

We will summarize our use of terminology regarding processes, rules, and patterns as follows. We will be discussing productivity in the word formation "process" (rather than "rule") for an affix. In comparing affixes, an affix will be said to be productive, with the intended meaning that the word formation process for that affix is productive. Productivity in a word formation "pattern" will be discussed for cases where what is examined involves multiple affixes, as in *-ization* and *-ification*. These notational preferences are not based on any special view regarding the involvement of rules or analogy. Plag (2003: 179) notes that "adopting a particular theory is often unnecessary for the solution of particular empirical problems." The empirical questions that we will address beginning next chapter are not bound by a particular view of what is involved in word formation.

1.3 Speakers' Perception/Production of Words

As a first step in examining how the degree of productivity could be investigated, we will review a series of experimental studies that investigate the relationship between the degree of productivity and speakers' perception/production of words. Aronoff (1976: 37) suggests that native speakers of a language have intuitions about which words are

acceptable words of their language, and that there are varying degrees with which words are deemed acceptable. For example, presented with *perceptiveness* and *perceptivity*, most English speakers would prefer *perceptiveness* (Aronoff, 1976: 37).

Aronoff and Schvaneveldt (1978) investigated native English speakers' graded acceptance of possible words formed with *-iveness* and *-ivity* patterns.[6] Possible words were those for which the *-ive* bases were attested but nominalizations in *-ness* and *-ity* were not (e.g., *recursiveness*, *recursivity*). When presented with possible words intermingled with attested words (e.g., *decisiveness*, *relativity*) and non-words (forms with non-occurring bases, e.g., *affertiveness*, *affertivity*), subjects judged more *-iveness* words as acceptable words of their language. These data confirmed Aronoff and Schvaneveldt's suggestion that *-iveness* is more productive than *-ivity*, and matched their prediction that words formed with a more productive pattern are accepted more readily by speakers. The data also converge with their observation that there are more word types listed in *Walker's Rhyming Dictionary* (Walker, 1936) for *-iveness* (140) than for *-ivity* (28).

It is not the case, however, that *-ness* is always preferred to *-ity*; rather, the outcome is pattern-dependent. In a follow-up study using *-ible* rather than *-ive* bases, Anshen and Aronoff (1981) found that possible words with *-ity* were accepted more often

---

[6] Aronoff (1980) suggests that testing speakers' graded acceptance of possible words has the advantage that it enables us to focus on productivity in a synchronic sense. Productivity in a synchronic sense concerns the productivity of a word formation process at a given point in time in the history of a language. On the other hand, a change in the productivity of a word formation process between two points in time (say, between the years 1800 and 2000) can be addressed in a diachronic sense. Since possible words are those that have not existed in the language, an experiment involving possible words tests productivity in a synchronic sense, and it satisfies the view (Aronoff, 1980) that only the synchronic aspect of productivity is of interest.

than those with *-ness*. The finding again coincides with their observation that there are more word types listed in Walker (1936) for *-ibility* (112) than for *-ibleness* (28).

Cutler (1980) attempted to account for speakers' preference for possible words in terms of their *phonological transparency*. A derived word is phonologically transparent if there is no phonological change in the base (e.g., *clear* + *-ness* → *clearness*), and it is phonologically opaque if there is a phonological change in the base (e.g., the vowel change observed in *clear* + *-ity* → *clarity*). Cutler found that more English speakers preferred words with a phonologically transparent base. Nevertheless, Anshen and Aronoff's (1981) findings discussed above constitute counterevidence to Cutler's (1980) claim: phonologically opaque *-ibility* words were more readily accepted by subjects (cf. Cutler, 1981).

There are three points to be noted about experiments investigating speakers' graded acceptance of possible words. First, since what is examined is the perception, rather than the production, of words by speakers, it is unclear whether word formation patterns that are preferred in these experiments are also those that lead to more coinages of words in word production. Second, the possible words presented to subjects were prepared by researchers, and since these words may or may not be actually used by speakers, it is unclear how the findings based on possible words relate to productivity in the actual use of language. Third, it is uncertain whether a higher acceptance rate of possible words following a word formation pattern is correlated with a higher degree of productivity of that pattern, or with the existence of a larger number of words following that pattern in dictionary listings.

To investigate productivity in the production of words, Anshen and Aronoff (1988) conducted an experiment in which subjects were asked to freely list words ending in particular word formation patterns. There were two points to note in their results. First, considering words present in Kučera and Francis (1967) as existing words, more non-existing words for *-ness* were listed by subjects. Second, there was greater overlap among the *-ity* words listed by the subjects. Anshen and Aronoff (1988) account for these findings in terms of the lexical storage of words. They suggest that *-ity* words are stored in the lexicon and are retrieved in full form, whereas *-ness* words are not stored in the lexicon and are formed as needed. Since the subjects retrieved *-ity* words from their lexicons, there was more overlap among the *-ity* words listed. On the other hand, the *-ness* words listed by the subjects were non-existing because the subjects actively constructed those words.

Another experiment on productivity in word production was conducted in Dutch by Baayen (1994). Baayen refined Anshen and Aronoff's (1988) methodology by the use of computers. Native speakers of Dutch were asked to type in words that ended with particular affixes. Prior to the experiment, Baayen used a corpus-based productivity measure (Baayen, 1989, 1992, 2001; see Section 2.4) to estimate differences in the degree of productivity between rival affixes. Among the words listed by the subjects, Baayen defined new words as those not present in his reference corpus of 42 million Dutch words, and found that more new words were listed by the subjects for affixes that the corpus-based productivity measure identified as more productive.

One thing to note about experiments investigating productivity in the production of words is that the environment in which the subjects are asked to list words does not resemble the environment where the language is actually used.[7]

The review of experiments investigating speakers' perception and production of words reveals that, to understand what the experimental data correlated with, it is important for us to have an independent method of assessing the degree of productivity of a word formation process, such as the one used in Baayen (1994). In the next chapter, we will examine what types of productivity measures are available in the literature.

1.4 Summary

This chapter introduced the issue of morphological productivity in word formation. Productivity is seen in a derivational word formation process involving an affix that lends itself to the coinage of a new lexeme. The word formation process for an affix may be subject to various restrictions, and affixes vary in their degree of productivity. The degree of productivity is the aspect of productivity that most intrigues researchers and it is that which is the focus of the present research. A review of experiments that investigated speakers' perception and production of words revealed that a clear method of assessing the degree of productivity of a word formation process is needed. It has become clear from the discussion in this chapter that what is crucially needed is a quantitative productivity measure.

---

[7] For example, since it is difficult to control an experiment involving the production, not the perception, of words, subjects may have adopted a systematic pattern in arriving at suitable words (as an extreme example, *redness* leading to *greenness* and *blueness*).

Chapter 2 Productivity Measures

It has become clear in the preceding chapter that our primary interest in the issue of productivity lies in the degree of productivity. To account for varying degrees of productivity among word formation processes, we need a method of quantifying productivity. Various proposals have already emerged in the literature as to how productivity might be quantified. Those proposals will be reviewed in this chapter.

2.1 Restriction-Based Approach

Some researchers (Schultink, 1961; Booij, 1977; Van Marle, 1985) have regarded the degree of productivity of a word formation process as inversely proportional to the number of restrictions on that process. Given that word formation processes are subject to various restrictions, it is reasonable to expect the degree of productivity of an affix to be affected by such restrictions. However, an attempt to describe the degree of productivity based on the number of restrictions does not lead to a quantitative measure.

There are three major problems with the restriction-based approach. First, it is not certain to what extent a given restriction limits a word formation process. Baayen and Renouf comment on this problem (1996: 87):

> [T]he idea that the degree of productivity decreases as the number of phonological, morphological, and semantic restrictions on a word formation rule increases is too imprecise to have any quantitative validity, for the simple reason that the number of different types removed from the input domain of an affix by such restrictions may vary widely from restriction to restriction. Without additional qualification of the restrictive weight of these restrictions, the claim

> that the degree of productivity and the number of restrictions are inversely related is simply vacuous.

The problem is also complicated by the fact that a given restriction may not hold absolutely (Bauer, 2001: 139).

Second, it is doubtful whether all restrictions can be exhaustively identified for each word formation process (Bauer, 2001: 127). The difficulty involved in identifying restrictions is summarized by Bauer (1988: 71) as follows:

> Since many kinds of limitation may apply to the bases available for a single morphological process, and the limitations have to be determined by the analysis of attested forms, the job of the linguist is an onerous one. Indeed, it is not clear whether anyone has ever stated in any clear way all the restrictions applying to the bases for any process: we simply do not know what such a statement would look like, and cannot tell whether all relevant variables might have been taken into account.

Third, although the restriction-based approach may lead to a (rough) prediction about how limited a given word formation process should be, such a prediction may differ from a description of the degree of productivity based on how word formation processes are actually used by speakers (Plag, 1999: 37). Bauer (2001: 139–143) points out that the number of words formed by a word formation process could be affected not only by restrictions but also by the need for those words. A given word formation process may be free of restrictions, but new words formed by that process may not be needed.

Although the consensus in the literature is that a restriction-based productivity measure is not viable, it cannot be doubted that restrictions on word formation do affect the degree of productivity. The point is, rather, that even if restrictions affect the degree

of productivity, it does not follow that such restrictions can be the basis for assessing the degree of productivity.

The restriction-based approach lacks empirical data as to how the word formation process for an affix is actually used to coin new words. The next section reviews empirical data that are available for assessing productivity.

2.2 Sources of Data

There are currently two major sources of data for a quantitative study of productivity: a dictionary and a corpus of texts. These data sources each have advantages and disadvantages for assessing productivity, which we review in this section. Productivity measures to be reviewed later in this chapter use either or both of these data sources.

2.2.1 Dictionary

Dictionaries have traditionally been a useful tool for studying word formation, making available a list of words for a specific morphological inquiry. A reverse dictionary, such as Walker (1936) or Muthmann (1999), sorts words by their word-ending patterns, and is handy for extracting a list of words with a particular suffix. Nowadays, many dictionaries are also published in electronic form (e.g., CD-ROM), and creating lists of words based on word-ending patterns is considerably facilitated by the search function accompanying an electronic version of a dictionary.

With respect to the coverage of words, the *Oxford English Dictionary* (*OED*) and *Webster's Third New International Dictionary* (*WD*) are currently considered the two most comprehensive and authoritative dictionaries. The OED represents British English and WD represents American English, although such dialect-specificity may be diminishing due to the global interaction of the two varieties of English.[8] In contrast to sources such as collegiate dictionaries, the OED and WD contain many rare or obsolete words that are rarely used by speakers (Baayen & Renouf, 1996), reflecting the fact that these dictionaries aim to achieve a fully comprehensive listing of attested usage. A list of words obtained from the OED or WD does not, therefore, represent the range of words familiar to any given speaker, or any given language community (Plag, 1999). Nevertheless, with an abstract, broadly defined English language community in mind, the OED and WD are often cited as authoritative sources for determining whether a given word exists in English.

The OED and WD, unlike other general dictionaries, present extensive etymologies for many of the listed words. The OED, in particular, is known for its rich etymologies. The etymologies of the OED tell us when and how a word was formed, and the electronic version of the OED allows a search for words based on those etymologies; for example, one can search for words that were formed with a given affix during a specific time period (e.g., between the years 1800 and 1900). Plag's (1999) research

---

[8] The OED and WD often list (or have information about) both British and American variants of a word (e.g., *colour* and *color*; *computerise* and *computerize*), but restrict the level of derivation for which the variants are listed: for example, the OED lists *colour*, *colourful*, *color*, but not *colorful*, whereas WD lists *color*, *colorful*, *colour*, but not *colourful*.

illustrates the use of etymologies in the electronic version of the OED in assessing productivity. We will discuss his findings in Section 2.6.

Because of their selectivity, dictionaries are often considered a less than optimal source of data for assessing productivity: they tend to omit words whose derivational patterns are clear (Anshen & Aronoff, 1988; Baayen & Renouf, 1996: 69). Anshen and Aronoff (1988) point out that many *-ity* words are typically listed in a dictionary as *headwords*, words with individual entries, whereas many *-ness* words are listed as *run-ons*, words that are listed under their bases. For example, the OED lists *probableness* as a run-on under *probable* while it lists *probability* as a headword.

In some cases, even the run-on listing is omitted; for example, the OED lists *stochasticity* as a run-on under *stochastic*, but does not list *stochasticness* at all, even as a run-on. Although this is perfectly a legitimate use of *-ness*, *stochasticness* is attested only 13 times on the Web (Google search engine, April, 2004). The omission of *stochasticness* makes sense given its strikingly low frequency of use. This illustrates the general tendency of a dictionary, even such a comprehensive dictionary as the OED, to list more frequent or idiosyncratic words than productively formed, infrequent words (Plag, 1999: 97).

For commercial reasons, dictionaries are likely to list words that are meaningful to the majority of dictionary users, instead of potential but non-existing words that presumably are of less interest (Baayen & Lieber, 1991; Baayen & Renouf, 1996; Plag, 1999: 97). Thus, we can expect lexicographers to show biases about which words should be listed in a dictionary (Bauer, 2001: 140; Baayen & Renouf, 1996: 81). In general, lexicographers are conservative in adding new words to a dictionary (Bauer, 2001: 35).

There could easily be words that speakers use but are not listed (Bauer, 1988: 62–63). A new word may not be in a dictionary either because the word was coined after the compilation of the dictionary, or because lexicographers were not convinced (on some unknown grounds) that the word merited a listing.

In summary, the general view of the literature is that a dictionary is limited as a data source for assessing productivity. Although there may be some advantages in the availability of etymologies, a list of words provided by a dictionary may not accurately reflect productively formed words in the actual use of language. The word list may be affected by unknown factors such as the decision of lexicographers to include (or not include) a given word.

2.2.2 Corpus of Texts

In any study that compiles statistics for a large corpus of texts, it is important to draw a clear distinction between *tokens* and *types*, so we will start our discussion by defining these terms which will be used throughout the rest of the present study in the contrast of *token* and *type frequencies*. Perhaps the simplest illustration of the distinction can be given for the word *the*. Any corpus is expected to have many occurrences of *the*,[9] and each of these occurrences counts as a *token*. A count of all occurrences of *the* in a corpus will give us its token frequency. Its type frequency will always be 1, because only one distinct form occurs.

[9] In the corpus of Kučera and Francis (1967), for example, this single word constitutes some 7% of the words sampled.

Let us now consider derivational affixation. Suppose that in a very small corpus we find 5 occurrences of *-ity* words, as listed in (1).

(1) {majority, opportunities, opportunities, possibility, possibilities}

Since we are concerned with the coinage of lexemes, not word forms, the inflectional endings of the words in (1) can be normalized as in (2):

(2) {majority, opportunity, opportunity, possibility, possibility}

In (2), the token frequency of *-ity* is 5, the count of all occurrences of *-ity*. The type frequency of *-ity*, on the other hand, is 3 because there are 3 distinct lexemes using *-ity*. In the present study of derivational affixation, we define token and type frequencies of a derivational affix (using the terminology developed in Section 1.2.3) as follows:

(3) a. The token frequency of a derivational affix is the sum of all occurrences of lexemes with that affix.
b. The type frequency of a derivational affix is the number of distinct lexemes with that affix.

The definitions of token and type frequency differ slightly for inflectional affixation. We will briefly discuss the corresponding definitions for inflectional affixes, although the productivity of inflectional affixes will not be dealt with in the present

study. In the case of the English plural suffix *-s*, for example, we will examine the words in (1) without normalizing them into (2). In (1), we have 3 tokens (i.e., *opportunities*, *opportunities*, and *possibilities*) and 2 types (*opportunities* and *possibilities*) of *-s*. We define token and type frequencies of an inflectional affix as follows:

(4) a. The token frequency of an inflectional affix is the sum of all occurrences of word forms with that affix.

b. The type frequency of an inflectional affix is the number of distinct word forms with that affix.

A large corpus of texts is available mainly in one of two types (Biber, 1993): the *domain-specific* type, and the *balanced* type. A *domain-specific* corpus is a collection of texts of more or less the same genre or register, such as texts drawn from a newspaper. The *Associated Press Newswire Corpus* of 44 million words used in Church and Gale (1991), for example, falls into this category. Since little attention needs to be given to the register variety, compilation is rather easy, and corpora of this type tend to be large (Biber, 1993). Many such corpora are available from the *Linguistic Data Consortium.*

A *balanced* corpus is a collection of texts of many different genres and registers. The *Brown Corpus* of 1 million words (Kučera & Francis, 1967) is an early and representative instance of a balanced corpus. The need for careful consideration of the text types included (their range, proportion, and so on) means that compilation of a balanced corpus is laborious, and that corpus size cannot be increased as easily as for a domain-specific corpus. One of the notable exceptions to the usual size limitation among

balanced corpora is the *British National Corpus* (*BNC*), which has as many as 100 million words. About 90 million words of the BNC belong to the written component, and 10 million words belong to the spoken component (Burnard, 2000). Taken as a whole, the BNC represents texts from a wide variety of genres and registers. Since texts from each unique source are sampled in an individual file, a domain-specific examination of sub-portions of the BNC is also possible. Among many corpora of large size available today, the BNC is one of the most sophisticated, and widely accepted, and for that reason, the present research uses the BNC. See Section 4.2 for more details of the BNC.

Corpus data are also available in the form of a lexical database, as seen in the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995). The CELEX database provides the frequencies of word forms (words with inflection) and lemmas (headwords without inflection), in English, Dutch, and German. The word frequency information of CELEX is based on the Cobuild Corpus (Renouf, 1987) of some 18 million words. The CELEX database is particularly useful in psycholinguistic research where calculations of word-form and lemma frequencies are often needed. Since frequency data are already made available, researchers do not need to concern themselves with the processing of raw corpus data.

Unlike a dictionary, a corpus of texts often contains productively formed words that are typically not listed in a dictionary (Baayen & Lieber, 1991: 803; Baayen & Renouf, 1996), and its compilation involves less filtering of words. Corpus data, therefore, reflect how words are actually used by speakers and writers (Aronoff & Anshen, 1998: 245; Baayen & Lieber, 1991: 803; Baayen & Renouf, 1996). Another important aspect of a corpus that a dictionary lacks is that it offers word frequency

information: using a corpus, we can assess how frequently a given word is used (Baayen & Lieber, 1991: 803).

Nevertheless, a corpus of texts has limitations. The range of words appearing in a corpus tends to be narrower than that in a dictionary (Bauer, 2001); for example, while the OED and WD widely cover terms used in different fields (e.g., medical, financial, and legal fields, etc.), how many of field-specific terms are found in a corpus depends crucially on the type and size of its texts. Therefore, while the data of a dictionary such as the OED or WD could be considered to belong to an abstract, broadly defined English language community, corpus data must be interpreted within the specific context represented by the texts of each corpus (or component of a corpus).

In summary, although special care must be taken about what types of texts are represented in a given corpus, a corpus of texts has advantages over a dictionary in studying productivity. In particular, corpus data serve as a good source for finding productive coinages of words seen in the actual use of language.

Before closing this section, we note that the Web can also be considered as a huge corpus of texts for morphological research (Baayen, 2003). The Web offers a corpus whose size has been given a lower-bound estimate of about 76 billion English words, as of March 2001 (Kilgarriff & Grefenstette, 2003: 339). Due to its rather unstructured nature, the use of the Web in linguistic research is still under development (see Kilgarriff & Grefenstette, 2003). Nevertheless, the use of the Web has already proved to be useful. In the preceding section, for example, we noted an interesting relationship between the fact that *stochasticness* in not listed in the OED, and its being attested only 13 times on the Web in an April 2004 Web-search.

2.3 Ratio of Actual to Possible Words

Aronoff (1976) was among the first to suggest a productivity measure. He first calls attention to the inadequacy of simply equating a large number of word types with an affix with high productivity of that affix. He argues that we must take into account the fact that the WFR of any given affix is subject to various restrictions, and that the WFR of one affix may be more restricted than the WFR of another. The suggested solution (Aronoff, 1976: 36) is to examine the ratio of actual words to possible words. Aronoff's proposed measure was later formalized by Baayen (1989: 28–31) as:

$$(5) \quad I = \frac{V}{S}$$

For a given affix, V is the number of actual word types with that affix, S is the number of possible word types with that affix, and I is the productivity index.

Some questions about the measure I have been raised in the literature (Baayen, 1989: 28–31; Baayen & Lieber, 1991: 803–804; Bauer, 2001: 189; Plag, 1999). The majority have to do with the lack of specification for the measure's implementation, namely how to obtain V and S. In Aronoff's original suggestion, the number of actual word types, V, is "the number of actually occurring words" formed by the relevant WFR, and the number of possible word types S is "the number of words which we feel could occur as the output" of the relevant WFR, the suggestion being that this could be obtained by counting "the number of possible bases for the rule" (see Aronoff, 1976: 36).

It would be difficult to estimate the numerator term *V* in an objective manner since it calls for a fixed list of words, some particular lexicon (Baayen & Lieber, 1991: 802–803). Since each individual may make a different judgment as to which are and are not actual words (Plag, 1999: 7), the number of actual words will be a matter of dispute, and the problem rests on how a word list is to be settled upon. Although a dictionary is a candidate that immediately suggests itself, we have already seen (Section 2.2.1) the difficulty in equating the range of words in a dictionary with the range of words familiar to any given speaker. Aronoff (1976: 36–37) himself points out a further problem: the notion of listing is not compatible with a very productive affix, such as the adverbial suffix *-ly*. An exhaustive listing of words derived with *-ly* would effectively be an exhaustive listing of adjectives, including adjectives which themselves are derived forms, because *-ly* attaches to almost any adjective to form an adverb. The suffix *-ness* has been offered as another example whose derived words cannot be exhaustively listed (Carstairs-McCarthy, 2002: 56).

The denominator term *S*, the number of possible words, is also problematic and difficult to estimate, in exactly the same way (Baayen, 1989: 28–31; Baayen & Lieber, 1991: 804). Again, a dictionary may be a candidate source, but Baayen (1989: 28–31) points out two problems. First, possible words by definition include words that have not been formed, and there could be many possible words that are not exhaustively listed in a dictionary. For example, in counting the possible bases for *-ion*, possible words with *-ize* and *-ify* must be included (*X-ize* → *X-ization*; *X-ify* → *X-ification*). Second, the number of possible words would be, in theory, very large for a very productive affix, which counterintuitively leads to a low index *I*. For example, given a productive affix such

as *-ness*, *S* would be increasingly large and the resulting index *I* increasingly small (Baayen & Lieber, 1991: 804).

Baayen (1989: 30) therefore suggests that Aronoff's measure *I*, can "ironically enough" be seen as expressing the *unproductivity* of the word formation process for an unproductive affix. For an unproductive affix, *S* would not be increasingly large, and as the number of actual words *V* approaches *S*, there would be arguably fewer bases left for coining new words with that affix.

2.4 Token Frequency and Hapax Legomena

Baayen (1989, 1992, 2001) proposes assessing productivity based on statistical data from a large corpus of texts. Of the several methods of assessing productivity proposed by Baayen, we will review the measure *P* in particular. (Other measures will be reviewed in Section 3.6.) *P* is formulated by Baayen (1992) as follows:

(6) $$P = \frac{n_1}{N}$$

For a given affix, *N* is the sum of tokens with that affix in a corpus, n_1 is the number of word types with that affix that occur only once in the corpus (the so-called *hapax legomena*, henceforth *hapaxes*), and *P* is the productivity index.

The primary interpretation of *P* is that it expresses the probability that a new word type with an affix will be encountered when *N* tokens with that affix have been sampled. A new word type is defined as one that has not yet been sampled in a corpus. Typically in

Baayen's writings, and also those of other researchers, such new words are often called *neologisms*. In citing these works, we will follow their practice and use the term *neologism*, in this sense.[10]

In addition to this probability-based interpretation, *P* can also be viewed as estimating the rate of vocabulary growth for an affix, that is, the rate at which the number of word types with an affix increases as corpus data are sampled (Baayen, 2001: 2–4; Baayen & Lieber, 1991). As words of a corpus are sampled one by one, a new word type with a given affix (a word type that is sampled for the first time) may be encountered. As more words of the corpus are processed, the number of sampled word types with that affix increases. Initially, many new word types with a given affix may be encountered at a high rate, but the rate at which new words with that affix are encountered will slow down as more words of the corpus are processed. The rate of increase in the number of word types with a given affix when *N* tokens with that affix have been sampled is estimated by *P*.

An important thing to note about *P* is that it is based on token frequency. Note that in addition to the fact that the denominator term *N* directly refers to (the sum of) token frequencies, the numerator term n_1 is also dependent on token frequency: a word is included in the n_1 count only if it occurs just once.

Baayen and Lieber (1991) report on the performance of *P* for a number of English affixes. They used an early version of the CELEX Lexical Database (Version E1.0) that

---

[10] Bauer (2001: 38–40) discusses some differences in meaning among *new word*, *nonce-word*, and *neologism*. While nonce-words are generally considered to be those that will not be part of the norm of the speech community, neologisms are considered to be those that will be part of the norm. New words are neutral between these senses.

provided word frequencies for the Cobuild Corpus of 18 million words (Renouf, 1987). The corpus represented both written (75%) and spoken texts (25%), drawn mainly from British English, and these texts were general and non-technical in their orientation, by and large. Among the obtained results, *P* indices for *-ness* (0.0044) and *-ity* (0.0007) accorded with linguists' intuitions about the difference in the degree of productivity between these suffixes. Baayen and Lieber take the *P* index for simplex nouns (0.0001) as a baseline against which an affix can be judged productive (or not). As compared with the *P* index for simplex nouns, the *P* index for *-ity* indicates that *-ity* has some degree of productivity. Other intuitively plausible *P* indices include those for *-ish* (0.0050 with noun bases; 0.0034 with adjective bases) and *-ous* (0.0006 with noun bases). A comparison with the *P* index for simplex adjectives (0.0001) indicates that *-ous* is also productive to some degree.

The consequence of the fact that the measure *P* is based on token frequency is exemplified in the findings for *-er* and *-ee*, which were described as "at first sight somewhat surprising" (Baayen & Lieber, 1991: 829). Although intuitively we may expect *-er* to be more productive than *-ee*, the *P* indices obtained for *-er* (0.0007) and for *-ee* (0.0016) would indicate the reverse: *-ee* is more productive than *-er*. The type frequencies for *-er* (682) and *-ee* (23) clearly do not play a role here.

A further example that shows the token-frequency basis of *P* is seen in the results for verbal affixes. Some strikingly low *P* indices were observed for *-ize* (0.0000710), *re-* (0.0000423), and *-ify* (0.0000000, i.e., no hapax), among others. Note that the *P* index for the verbal suffix *-ize* is substantially smaller than that for the nominal suffix *-ity* (0.0007). Very few hapaxes were found for *-ize* (2 with noun bases; 1 with adjective

bases) and *-ify* (0), as compared with hapaxes for other affixes, such as *-ness* (77). Nevertheless, when compared with the *P* index for simplex verbs (0.0000065), *re-* and *-ize* are still correctly identified as being productive. Baayen and Lieber (1991: 831) attribute the low *P* index for *-ize* partially to the fact that the texts of the Cobuild Corpus are not scientific or technical. Baayen and Lieber (1991: 836) also suspect that it may be relatively more difficult to coin a new verb than a new noun or adjective because verbs are more fundamental elements whose coinage involves establishing the predicate-argument structure.

As the corpus size keeps increasing, *N* for any affix would become increasingly large, and *P* would become vanishingly small (Baayen, 1992: 119; Baayen & Lieber, 1991: 837). That is, for any affix, *P* becomes zero in the limit as $N \rightarrow \infty$. It is important, therefore, to understand that *P* is dependent on corpus size; for example, *P* could be high in a small corpus where many word types still remain to be sampled for the first time, or low in a large corpus where most word types have been sampled more than once.

Undoubtedly, the most crucial element of the measure *P* is n_1, the number of hapaxes, which is used as an estimator of the number of new words. Baayen and Renouf (1996: 78) provide three reasons for giving hapaxes a crucial role in assessing productivity:

(7) a. Statistically, the proportion of hapaxes in a corpus offers an estimated probability of encountering new word types.

b. Most neologisms occur only once; that is, most neologisms are hapaxes.

c. Productive word formations occur unintentionally, and hapaxes tend to

represent those word formations.

Note that these three reasons need not be all justified simultaneously because each of them is an independent motivation for examining hapaxes.

The first of the reasons offered, (7a), means that the use of hapaxes is motivated by probability theory. We will review the underlying probability estimation method in Section 3.2.1.

With respect to (7b), Baayen and Renouf provide empirical evidence that the majority of neologisms are indeed hapaxes. They examined five English affixes, *-ly*, *-ness*, *-ity*, *un-*, and *in-*, in the (British) Times newspaper corpus of 80 million words. As a frame of reference for determining which words are neologisms among the sampled words, they used Webster's Third New International Dictionary (WD) and defined neologisms as those words not listed in the dictionary. For *-ness* and *-ity* in particular, the token frequencies (1 through 5) of neologisms so identified were analyzed, and the majority were found to occur among hapaxes. Although the majority of non-neologisms were also hapaxes,[11] a larger than expected proportion of the neologisms were hapaxes. Baayen and Renouf argue that among different token-frequency categories, neologisms are best represented among words of low frequency, and particularly among hapaxes.

[11] For both *-ness* and *-ity*, the largest number of words occurred with the frequency of 1, and the next largest number of words occurred with the frequency of 2, and so on; that is, the number of words occurring with a given frequency was a decreasing function of that frequency.

The claim in (7b) must be distinguished from its (unintended) inverse, namely that "most hapaxes are neologisms." The latter interpretation is undoubtedly tempting as a means of evaluating the measure. Plag (2003: 56), for example, examines the first 20 hapaxes for *-able* in alphabetical order extracted from the BNC, and having found that 13 of these hapaxes are not listed in WD, concludes that the number of neologisms is indeed high among hapaxes. However, it could just as well be the case that if we examined 20 *-able* words with token frequency 2, something like 13 of these words might also not be listed in WD. By examining hapaxes alone, we would not know whether most neologisms are represented among hapaxes. Going through a list of hapaxes one by one does not speak to the proper verification of the claim (7b).

Since not all hapaxes are claimed to be neologisms (Baayen & Lieber, 1991: 813), it follows that the measure *P* would not be invalidated even if, in an extreme case, none of the hapaxes for an affix were a neologism in a given corpus. Hapaxes function as an estimator of neologisms, and do not themselves need to be neologisms. Thus, the measure *P* relies on an estimated number of neologisms, which are not to be captured within sampled data but rather to be sought outside the corpus.

Among the three reasons for focusing on hapaxes, (7c) has been found to be the most controversial (Bauer, 2001; Plag, 1999). The claim that a productive word formation is carried out *unintentionally* is due to Schultink (1961). According to Schultink's view of productivity, intentional word formations are "creative" but not necessarily "productive." In the morphology literature, a distinction is often drawn between *productivity* (referring to a rule-based process of forming a new lexeme), and *creativity* (referring to a careful, intended coinage of a new lexeme, often depending on

analogy). The coinage of brand names, for example, is considered as a typical example of creativity. There is a general attitude among researchers to wish to exclude creativity from productivity based on the lack of involvement of rules (Bauer, 1983: 63; 2001: 64).

A view of productivity stipulating unintentionality has received some criticism (Bauer, 2001; Plag, 1999). Plag (1999: 14), for example, points out that many scientific terms are productively formed with *-ize*, and finds it difficult to believe that they are formed unintentionally. Bauer (2001: 67–68) raises questions as to whether a word formation process can be sometimes productive and sometimes creative, and whether the identification of intentional word formations is possible based on available data. Given the difficulty in distinguishing intentional coinages from unintentional coinages, Bauer (2001: 68, 71) suggests that productivity and creativity may be prototypes, rather than distinct categories. It is unclear whether it is useful or possible to include a notion as vague as "unintentionality" in word formation. It also remains unclear whether token frequency in corpus data (i.e., the fact that a word is a hapax) can usefully make a distinction between unintentional and intentional word formations.

We close this section by noting that the measure *P*, being dependent on token frequency, has found many applications in the psycholinguistic literature. The measure has been used in psycholinguistic studies investigating the relationship between productivity and lexical access (e.g., Schreuder & Baayen, 1995; Bertram, Schreuder, & Baayen, 2000), and morphological parsing (Hay & Baayen, 2002). That many extensions of the measure *P* are found in the psycholinguistic literature is understandable, given the important role that the word frequency effect has been found to play in processing (see Monsell, 1991, for an overview).

2.5 Corpus/Dictionary Comparisons

Bauer (2001: 156–161) proposes a productivity measure based on a comparison between a corpus and a dictionary. Under this measure, productivity is captured as the number of words that have been coined over some period of time. The more productive the word formation process for a given affix is, the more coinages formed by that process are expected to be found in that time period (Bauer, 2001: 156). Given two points in the history of a language, t_1 and t_2 (where t_1 precedes t_2), Bauer proposes associating t_1 with dictionary data and t_2 with corpus data. The comparison of the two gives us neologisms formed between t_1 and t_2. The rationale is as follows. A dictionary is generally conservative with respect to the inclusion of new words, whereas a corpus of texts includes whatever new words of necessity. Those words that are in a corpus at t_2 but are not in a dictionary at t_1 could be interpreted as new words formed between t_1 and t_2. Since t_1 precedes t_2, Bauer suggests a corpus/dictionary combination in which the publication date of the dictionary precedes the date of the materials sampled in the corpus.

Let a denote the number of word types with an affix found in a corpus, and let b denote the number of the subset of those word types that are found in a dictionary. Simple subtraction, $a - b$, gives us the number of word types with that affix that occur in the corpus but not in the dictionary, and this number could be interpreted as an index of productivity (Bauer, 2001: 159). Here, it is important to note that b is a subset of a. If b were mistakenly interpreted as the total number of word types with an affix in a dictionary, the subtraction, $a - b$, could result in a negative value. To avoid such a misinterpretation, we could use set notations to formalize the measure as follows. Let C denote the set of word types with a given affix that occur in a corpus, and let D denote the

set of word types with that affix that are listed in a dictionary. The productivity measure based on the comparison of C and D (originally, $a - b$) can be expressed as:

(8) $$P_{CD} = |C| - |C \cap D|$$

Bauer examined the productivity of *-ment*, *-ation*, *-isation*, and *-ificatioin* based on the 1.1-million-word *Wellington Corpus of Written New Zealand English* (created with texts from 1986) and the 1982 edition of the *Concise Oxford Dictionary*. The measure found 4 neologisms for *-isation* but none for *-ment*, *-ation*, and *-ification*. He concludes that of these patterns examined, only *-isation* is productive.

What is of interest about Bauer's approach is that he focuses on the gap between a dictionary and a corpus: the former is conservative and the latter liberal in including new words. However, two problems can be predicted for the proposed measure. First, the extent to which words are not listed in a dictionary could vary from affix to affix. We noted in Section 2.2.1 that some words that follow a clear derivational pattern may not be listed in a dictionary; for example, *-ness* may have more unlisted words than *-ity*. If that is the case, despite Bauer's original proposal to capture new words formed between two points in time, the proposed measure may also capture words that tend to be omitted in a dictionary, regardless of whether they are new. Thus, the degree of productivity based on this measure may also express the extent with which words with an affix are omitted in a dictionary.

Second, the obtained list of potentially new words could be much dependent on a particular corpus/dictionary combination, and there could be many possible

corpus/dictionary combinations involving variables such as the type and size of the corpus and dictionary. If different combinations yield different results, the interpretation of the results could be complicated.

A method similar to the corpus/dictionary comparison discussed above is a dictionary/dictionary comparison proposed by Bolozky (1999) in a study of productivity in Hebrew. Bolozky proposes comparing two dictionaries compiled at different times, or examining an addendum to a dictionary. In this approach, new words are those that are absent in an earlier dictionary but are present in a later dictionary, or simply those found in the addendum. In English, we could achieve a similar result by making use of etymologies in the OED, which is reviewed in the next section.

2.6 OED Etymologies

While (dated) etymologies drawn from the OED have been widely used in studies of productivity (e.g., Anshen & Aronoff, 1989; Bauer, 1992, 2001), Plag (1999: 91–118) demonstrates the use of the OED etymologies specifically in the context of assessing the degree of productivity. Using the search function of the electronic version of the OED, Plag (1999) sought neologisms for the verbal affixes *-ate*, *-ize*, *-ify*, *-en*, *be-*, and *en-/em-* formed between the years 1900 and 1985. The measure that he used can be expressed as:

(9) $P_{\text{OED}} = V_{MN}$

Based on the OED etymologies, V_{MN} is the number of new word types with a given affix formed between the years M and N.

The obtained neologisms indicated that *-ize* (346), *-ate* (87), and *-ify* (30) are productive while *-en* (2), *be-* (0), and *en-/em-* (7) are marginally productive or not productive at all. This finding matched Marchand's (1969: 364) claim that *-ize*, *-ate*, and *-ify* are the only productive verbal affixes in English. The numbers of neologisms for *-ize*, *-ate*, and *-ify* were also compared with the numbers of hapaxes for these suffixes in the Cobuild Corpus of 18 million words (Renouf, 1987). The numbers were somewhat similar for *-ate* (87 neologisms, 69 hapaxes) and *-ify* (30 neologisms, 18 hapaxes) but were substantially different for *-ize* (346 neologisms, 80 hapaxes).

The OED, however, has the limitation that its entries are dependent on lexicographers: there may be neologisms that lexicographers overlook for unknown (or unspecified) reasons. Plag (2003: 53) points out, for example, that despite the higher productivity expected for *-ness* relative to *-ize*, the number of *-ness* neologisms (279, Plag, 1999: 98) listed in the OED for the 20th century roughly equals the number of *-ize* neologisms for the same period.

Nevertheless, Plag (1999) remarks that, depending on how it is used, the OED, with its exceptional coverage and rich etymologies, could provide a useful resource for studying productivity.

2.7 Summary

In this chapter, we reviewed several quantitative productivity measures that have been proposed in the literature. The inadequacy of the restriction-based approach to describing the degree of productivity was discussed, and we reviewed dictionaries and corpora of texts as the two currently available major sources of data for a quantitative

study of productivity. The examination of different productivity measures reveals that the major task in assessing productivity lies in correctly identifying new words in given data.

Of the different approaches reviewed in this section, the general corpus-based approach (seen in the measure *P*) is one that will be taken up in the remaining chapters of the present study. Given the advantages afforded by corpus data, this approach is one that is currently more promising, and also one in which many possibilities for exploiting corpus data are still to be discovered. Despite the limitations of dictionaries discussed in this chapter, these remain useful in studying productivity. We will use a corpus of texts as the primary source of data for assessing productivity, and dictionaries as a source of supplementary information in analyzing those data.

Chapter 3 Defining New Words in Corpus Data

Comparing the performance of various productivity measures (most of which were reviewed in the preceding chapter), Bauer (2001: 204) concludes that "[d]ifferent productivity measures do not always agree, and there is thus still room for a generally-agreed measure of productivity." The absence of a uniform productivity measure may not be surprising if we think of the inherent difficulty of defining new words: a judgment as to whether a given word is new can vary from individual to individual. What is crucial is the criterion based on which we regard a given word as new. Different productivity measures adopt different criteria. In this chapter, we will focus on the corpus-based definition of new words. Borrowing insights from a technique used in Language Technology, a new productivity measure will be proposed.

3.1 Unseen Words as New Words

In a corpus-based approach, new words are defined as "unseen" words, words not previously sampled in a corpus. A good example of this concept is shown in the work associated with the AVIATOR Project (Renouf, 1987, 1993) which involves a *dynamic corpus* called the Cobuild Corpus. A *dynamic corpus*, as opposed to a *static corpus*, is a corpus whose size keeps increasing with a constant addition of texts. At certain time intervals, a chunk of texts is processed (e.g., part-of-speech tagged) and added to the corpus. As texts are successively added to the corpus, the words of those texts are compared with the words that are already in the corpus. If a word that has not previously

been sampled is encountered, it is considered a new word, and information about this first occurrence of the word is recorded. Since the successively added chunks of texts are viewed as a chronological flow of texts, new words in this approach are defined in opposition to old words that have been seen already in the sampled data. New words that have been formed during a given period can be collected based on the information about their first occurrences. Baayen and Renouf (1996) demonstrate the use of this information in a study of productivity.

In the corpus-based measure *P* that we reviewed in the preceding chapter, new words are also defined as unseen words after a given corpus has been sampled. The assumption underlying this definition of new words is that words that have not been sampled in a large corpus of texts tend to be the kind that may be new to the language community (Baayen & Lieber, 1991: 813).

An advantage of statistically estimating the number of unseen words (via count of hapaxes) is that it does not involve any judgment as to whether a given word is new. Assuming that the number of hapaxes is known in a given corpus, the measure *P* yields the same results, regardless of who uses the measure or how many hapaxes are new words. In other measures where a list of potentially new words is actually obtained, there is a temptation to go through the captured new words to determine which are in fact new. However, once a subjective judgment is involved, the results of a measure vary depending on the user, and this is undesirable. This is not the case with the measure *P*. One is not required to go through hapaxes and see if they are new words because hapaxes by themselves are not claimed to be new words.

There is, at the same time, an inherent problem with estimating unseen words: words that are unseen in a corpus remain unseen, so that only their number (and not their identity) are estimated. Hapaxes are seen words, and not themselves new words. We could, to some extent, examine hapaxes to see what types of words may be estimated, but the fact remains that new words are only estimated and never captured within corpus data.

The above considerations lead to a logical impasse, one which raises an interesting question: Is it ever possible to capture (not just estimate the number of) unseen words within a given corpus? In search of an answer to this question, we now turn to the discussion of so-called *smoothing* techniques in Language Technology.

3.2 Unseen Words and Smoothing

A previously unseen item in corpus data is a familiar concept in Language Technology (see Chen & Goodman, 1998; Manning & Schütze, 1999; Jurafsky & Martin, 2000). A given item will be previously unseen in a corpus if the corpus data are sparse. As discussed below, unseen items arising from sparse data pose serious problems in probabilistic language modeling.

In a probabilistic language model that uses corpus-based statistics, probabilities of word occurrence are calculated based on the relative frequencies of words in *training data* (henceforth, a *training set*). The obtained probabilities of word occurrence are used to process words in *test data* (henceforth, a *test set*). In an *n-gram* model, the next word w_n of an n word sequence is predicted by the previous $n - 1$ words that have been

processed. The probability P that w_n will occur is calculated as a conditional probability (Manning & Schütze, 1999: 196):

$$(1) \quad P(w_n \mid w_1 \ldots w_{n-1}) = \frac{C(w_1 \ldots w_n)}{C(w_1 \ldots w_{n-1})}$$

Here, $C(x)$ returns the frequency of the sequence x in the training set. In a bigram model (with $n = 2$), the occurrence of w_n is predicted by the preceding $2 - 1$ words, that is, the immediately preceding word only, so (1) reduces to:

$$(2) \quad P(w_n \mid w_{n-1}) = \frac{C(w_{n-1}\, w_n)}{C(w_{n-1})}$$

A serious problem that any probabilistic language model faces is that since the training set is finite, there will inevitably be sequences of words in the test set that are absent in the training set. For example, in a bigram model, the training set may have 3 occurrences of the sequence *captain steer* (with inflection normalized) but no occurrence of *captain drive*. In the training set, the term *drive* may appear with high frequency, but never preceded by the word *captain*. If *captain drive* is attested once in the test set, it will be assigned a zero probability of occurrence because C(*captain*, *drive*) would return a zero frequency. Also, if there is a word in the test set that is absent in the training set, any sequence of words that contains that word will necessarily be assigned a zero probability. For example, the training set may have no occurrence of the verb *sprint*. If the test set has

the verb *sprint*, any sequence of words involving *sprint* (whether it be *athlete sprint* or *man sprint*) must therefore be assigned a zero probability of occurrence.

No matter how large a training set is, the data of the training set will be sparse in the sense that at least some words or word sequences are non-occurring. Therefore, any probabilistic model requires a solution for the sparse data problem. Many techniques for solving that problem, called *smoothing* techniques, have been proposed in the literature (e.g., Good, 1953; Church & Gale, 1991; Jelinek & Mercer, 1985; Katz, 1987; Witten & Bell, 1991).[12] Among the different smoothing techniques, three prominent ones will be reviewed in this section: the Good-Turing estimation method (Good, 1953; Church & Gale, 1991) and the held-out and deleted estimation methods (Jelinek & Mercer, 1985).[13] In keeping with most such discussions in the literature, the three smoothing methods will be discussed in the context of a bigram model.

The basic aim of smoothing is to adjust observed frequencies in a training set so that no bigram in a test set, even if it is absent in the training set, will receive a zero probability of occurrence. A non-zero probability of an unseen bigram is achieved by passing some relative frequencies of seen bigrams onto unseen bigrams. That is, some seen bigrams are made less likely to occur so that unseen bigrams will have at least some probability of occurrence. Smoothing techniques differ mainly with respect to how much

---

[12] An evaluation of probabilistic language models that employ different smoothing techniques is available in Chen and Goodman (1998).

[13] Church and Gale (1991) report successful performances of these methods based on bigrams in the *Associated Press Newswire Corpus* of 44 million words.

probability mass (i.e., relative frequencies) is assigned to unseen words, and from which frequency categories the probability mass is collected.[14]

3.2.1 Good-Turing Estimation Method

To begin with, given N bigrams in total in a training set, the probability P of a bigram that has frequency r is:

$$(3) \quad \mathrm{P} = \frac{r}{N}$$

In all the smoothing methods under discussion, the probability P of a bigram that has frequency r will be calculated as:

$$(4) \quad \mathrm{P} = \frac{r^*}{N}$$

Here, r^* is an adjusted frequency. How to obtain an appropriate value of r^* for each r is where smoothing methods differ.

The smoothing technique based on Good's (1953) probability estimation method is widely known as the Good-Turing estimation method (Church & Gale, 1991; Gale,

[14] Notations used in the sections that follow are adopted (variously) from ones used in the literature (Church & Gale, 1991; Jurafsky & Martin, 2000; Manning & Schütze, 1999: 206, 210–213).

1994). Given data that are binomially distributed (Church & Gale, 1991; Gale, 1994), the method calculates the adjusted frequency r^* as:

$$(5) \quad r^* = (r+1)\frac{E(N_{r+1})}{E(N_r)}$$

Here, N_r is the number of bigrams that occur r times in the test set, and the function $E(x)$ returns the expected value of a random variable. The formula is typically used for frequencies $r < k$ with a constant k (e.g., $k = 5$ in Katz, 1987). Since low frequency items are numerous in corpus data, $E(x)$ can be reliably substituted with the actual value of x (Manning & Schütze, 1999: 212).[15]

Formula (5) cannot be used directly when $r = 0$, so Good (1953: 239) proposes the following formula for estimating the probability P when $r = 0$:

$$(6) \quad \mathrm{P} = \frac{n_1}{N}$$

where N is the total number of bigrams, and n_1 is the number of bigrams that occur once.

Note that formula (6) is the measure P when incorporated into a study of word formation. See Baayen (2001) for more detailed discussion of the relationship between the Good-Turing estimation method and the measure P.

[15] Alternatively, Good (1953) suggests replacing $E(x)$ with some smoothed value $S(x)$, whose value depends on which of several methods he proposes is used.

3.2.2 Held-Out/Deleted Estimation Methods

The held-out estimation method (Jelinek & Mercer, 1985) divides a training set into two sets: a *retained set* and a *held-out set*. The basic idea of the held-out method is to use the held-out set as a preview of the test set. Thus, frequencies in the retained set are adjusted based on the differences in frequency between the retained and held-out sets, so that the data of the test set will be processed more appropriately. The adjusted frequency r^* of formula (4) is obtained as follows:

$$(7) \quad r^* = \frac{T_r}{N_r}$$

N_r is the number of bigrams with frequency r in the retained set, and T_r is the sum of frequencies of the bigrams of N_r that are observed in the held-out set. Let $C_R(b)$ be the frequency of bigram b in the retained set and $C_{HO}(b)$ the frequency of b in the held-out set. The formula for T_r is:

$$(8) \quad T_r = \sum C_{HO}(b), \text{ for } b \text{ such that } C_R(b) = r$$

That is, for each bigram that has the frequency of r in the retained set, we examine how many times the bigram occurs in the held-out set, and obtain the sum of these bigram frequencies. What is of interest about the held-out method is that the calculation of r^* is based on actual differences in frequency between the retained set and the held-out set. Suppose there are two bigrams occurring with frequency $r = 10$ in the retained set, hence

$N_{10} = 2$. These bigrams occur with frequencies 10 and 9 in the held-out set, hence $T_{10} = 19$ by formula (8). According to formula (7), $r^* = T_{10}/N_{10} = 19/2 = 9.5$. This simple example shows how the calculation of r^* is based on actual differences in frequency between the two sets of data.

The held-out method achieves what has been called a "gold standard" estimation (Manning & Schütze, 1999: 210) when the retained set is the training set and the held-out set is the test set; that is, when the adjustment in frequency is determined by examining frequency differences between the training set and the test set. Training a model based on the test data of course subverts the usual purpose of building a probabilistic language model, but here it provides an upper bound against which the performances of models can be compared (Church & Gale, 1991: 25; Manning & Schütze, 1999: 210). The held-out method is regarded as providing an empirical answer to the question of how much adjustment in r is needed (Church & Gale, 1991: 24; Manning & Schütze, 1999: 205). In contrast to the Good-Turing method, which requires binomially distributed data, the held-out method requires only that the training and test sets are drawn from similar sources (Church & Gale, 1991: 25).

The deleted estimation method is an extension of the held-out estimation method that is "more parsimonious with the available data" (Jelinek & Mercer, 1985: 2593). Here, the training set is divided into two subsets identified as A and B, and we designate one as a *retained set* and the other as a *held-out set* (or a *deleted set*[16]). Similarly to the held-out method, we let N_r be the number of bigrams with frequency r in the retained set,

[16] In the original proposal, the held-out set is called a *deleted set* (Jelinek & Mercer, 1985: 2593), but we will retain the term *held-out set* in order to clarify the characteristic of the deleted method as an extension of the held-out method.

and let T_r be the sum of frequencies of the bigrams of N_r that are observed in the held-out set. One crucial difference between the held-out method and the deleted method is that under the deleted method, the roles of retained and held-out sets switch between the two sets, A and B. That is, once values of N_r and T_r that are needed for calculating r^* are obtained with retained set A and held-out set B, A becomes the held-out set and B the retained set, and a second set of values for N_r and T_r will be obtained. Using superscripts to refer to the two sets A and B, N_r^{A} is the number of bigrams with frequency r in A (as retained set), and T_r^{BA} is the sum of the frequencies of the bigrams of N_r^{A} that occur in B (as held-out set). After a switch of roles, we also obtain N_r^{B}, the number of bigrams with frequency r in B (now as retained set), and T_r^{AB}, the sum of the frequencies of the bigrams of N_r^{B} that occur in A (now as held-out set). The adjusted frequency r^* is obtained by the following formula:

$$(9) \quad r^* = \frac{T_r^{\mathrm{BA}} + T_r^{\mathrm{AB}}}{N_r^{\mathrm{A}} + N_r^{\mathrm{B}}}$$

The deleted estimation method is also known as the *cross-validation* method. A possibility of incorporating the deleted estimation method into a productivity measure will be discussed in the next section.

3.3 Productivity Measure Proposed

One basic mechanism of the deleted estimation method that is of interest to us in formulating a productivity measure is the cross-comparison of data sets to find differences between them.

The deleted estimation method is also of interest because, with some modifications that will be spelled out shortly, the method can be incorporated into a productivity measure based on type frequency. Before we proceed with the formulation of such a productivity measure, we take a moment to consider the motivation for using type frequency, rather than token frequency.

3.3.1 Use of Type Frequency

Equating the type frequency of an affix in a dictionary with its productivity has been claimed to have problems (Aronoff, 1976; Bauer, 2001: 21). Bauer (2001: 48–49, 144), for example, argues that type frequency in a dictionary can tell us only about the *past productivity*, and not the *present productivity* of a word formation process: a large number of word types with a given affix would indicates that the affix has been productive at some point(s) in the history of the language, but we cannot know whether the word formation process for that affix is still available. Bauer offers *-ment* as an example for which a dictionary may list a large number of word types, while new coinages with this suffix have been rare in recent years.

It has also been claimed that the type frequency of an affix in a corpus, V, often leads to counterintuitive results in assessing the degree of productivity (e.g., Baayen,

1992; Baayen & Lieber, 1991). For example, Baayen and Lieber (1991: 804) point out that the type frequencies of *-ness* (497) and *-ity* (405) in their corpus of 18 million words do not adequately express the fact that *-ness* is intuitively felt to be much more productive than *-ity*. (*P* indices obtained in their study for *-ness* (0.0044) and *-ity* (0.0007) seemed to capture the intuitive difference in productivity between these two suffixes more adequately.) A problem with *V* is that an affix that is not synchronically productive could still have a large number of word types if the word formation process of that affix was productive in the past and if the derived words remain current. Baayen (1992: 110–110) concludes that productivity cannot be measured in terms of type frequencies alone.

Baayen (1992: 123) sees *V* as representing the *extent of use*, that is, the extent with which the word formation process for an affix has been used to form words. It is not necessarily the case that *V* and *P* agree: for example, given *-er* and *-ee*, Baayen and Lieber (1991) find a markedly greater value of *V* for *-er* (682) than for *-ee* (23), but a modest reversal of *P* values for *-ee* (0.0016) and for *-er* (0.0007).

A possible use of *V* is suggested in an analysis of the *global productivity* of an affix, a two-dimensional analysis of overall productivity in which the *x*-axis shows *P* and the *y*-axis shows *V* (Baayen, 1992, 1993; Baayen & Lieber, 1991: 817–819). If two given affixes are tied with respect to *P*, along the *x*-axis, those affixes can be contrasted by *V* along the *y*-axis. Given a particular affix, *V* represents word types that we already have and *P* represents the probability of encountering new word types, and the claim is that both these elements are relevant to the overall productivity of that affix (Baayen, 1993). However, an analysis of global productivity suffers a limitation in that affixes that differ

with respect to both P and V cannot be contrasted in the absence of a clear method of weighing the relative importance of P and V (Baayen, 1993).

Another possible use of V is suggested by Baayen (1992) in combination with $\hat{S}$, the number of possible word types with an affix to be expected when the size of a corpus is increased theoretically infinitely. Baayen (1993) shows with examples from Dutch that productivity rankings of affixes can be obtained based on $\hat{S}$, or on the ratio of $\hat{S}$ to V. However, we will not examine these measures in detail for two reasons. First, a measure based on the ratio of $\hat{S}$ to V is claimed to express the pragmatic potentiality of a word formation process, the extent with which possible words ($\hat{S}$) are exhausted by actual words (V), but it is not considered as expressing the degree of productivity (Baayen, 1992: 122). Second, obtaining values of $\hat{S}$ involves an intricate calculation based on an extended version of Zipf's law, increasing corpus size infinitely in theory, and such an approach does not match the current attempt to capture new words within given corpus data.

The claims discussed above about the inadequacies of type frequency in a corpus hold only if type frequency is assessed and evaluated for a whole corpus. In the following section, it will be shown that type frequency in a corpus can indeed be effectively used to estimate the degree of productivity if word types in different corpus segments are compared. This will be possible by the deleted estimation method.

What is the motivation for pursuing a productivity measure based on type frequency? Despite the evident disadvantages associated with any type frequency in general (whether in a dictionary or corpus), the number of word types with an affix has traditionally been the kind of information that is conceptually easy to grasp, which

explains the "temptation" to equate the number of word types with an affix with the degree of productivity of that affix. What we will pursue in the present study is the possibility that analyses with word types and type frequency in a corpus have not been explored fully, and there are ways to make good use of type frequency in a corpus in assessing productivity. One such example will be shown in the present study in conjunction with the deleted estimation method.

From a technical point of view, a method based on type frequency also has an advantage, as pointed out by Plag (1999: 100), that a decision to include or not include a particular word in an analysis would have less impact on type frequency than on token frequency. For example, consider the question of whether *business* is a word type with *-ness*. Etymological information in the OED suggests that *business* was originally formed with *-ness*, but the synchronic validity of the word formation may be questioned. If we exclude *business* on the basis of the lack of synchronic validity, we also need to reconsider the treatment of many other words such as *ability* and *activity* that are originally French loans. Many *-ity* words are said to be lexicalized (e.g., Anshen & Aronoff, 1988), and are known to have idiosyncratic meanings. Whichever way the decision falls, the inclusion or exclusion of *business* would affect the type-frequency count for *-ness* by only 1, whereas it could affect the token-frequency count for *-ness* substantially, depending on the source of texts sampled.

3.3.2 Productivity Measure Based on Deleted Estimation

We will now proceed with incorporating the deleted estimation method into a productivity measure.[17] We focus on formula (8), repeated here as (10):

(10) $T_r = \sum C_{HO}(b)$, for b such that $C_R(b) = r$

For all bigrams that have frequency r in the retained set, T_r is the sum of frequencies of those bigrams in the held-out set. What we are interested in achieving based on the underlying mechanism of (10) is that for all words that have frequency 0 in the retained set, the number of those words present in the held-out set should be obtained. Formula (10) is therefore transformed into (11):

(11) $V_0 = \sum 1$, for w such that $C_R(w) = 0$ and $C_{HO}(w) \geq 1$

Here, $C_R(w)$ and $C_{HO}(w)$ return the token frequency of word w in the retained set and the held-out set, respectively, and V_0 is the number of word types in the held-out set that are unseen, $r = 0$, in the retained set. The summation is conducted over word types, not word tokens.

Using formula (11), a productivity measure based on the deleted estimation method, which we will call the P_{DE} measure, is formulated as follows. We begin by

[17] Many of the notations used in the remainder of this study are adopted from and/or influenced by ones used in a series of studies by Baayen and his colleagues (see Baayen, 2001: Appendix A, for a list of symbols).

preparing two corpus segments that have the same size (m) and text type. For the moment, we simply assume that two such corpus segments are available (see Section 4.5 for further discussion), and that these are labeled A and B. Using superscripts to refer to A and B, given a particular affix, V_0^{AB} is the number of word types with that affix in A (as the held-out set) that are unseen in B (as the retained set), and V_0^{BA} is the number of word types with that affix in B (as the held-out set) that are unseen in A (as the retained set); in addition, V^A is the total number of word types with that affix in A, and V^B is the total number of word types with that affix in B. P_{DE} is calculated as:

$$(12)\quad P_{DE} = \frac{V_0^{AB} + V_0^{BA}}{V^A + V^B} = \frac{\left(V_0^{AB} + V_0^{BA}\right)/2}{\left(V^A + V^B\right)/2} = \frac{V_N}{V}$$

For a given affix, V is an average number of word types with that affix in a corpus segment of size m, V_N is an average number of *otherwise-unseen* (new) word types with that affix in a corpus segment of size m, and P_{DE} expresses the degree of productivity of that affix. Words contributing to V_N (V-New)[18] are called *otherwise-unseen* words because the status of a word being unseen depends on a particular segmental relationship. When we average V_0^{AB} and V_0^{BA} as in (12), the segmental relationships are obscured, and it becomes more appropriate to regard words contributing to the V_N count as having the "potential to be unseen."

[18] This symbol should not be confused with $V_{(N)}$, the number of word types when N tokens have been sampled, which appears in the literature but not in the present study. In the present study, the subscript N in V_N stands for New.

Our attempt to capture unseen words within a corpus resulted in capturing otherwise-unseen words, words that have the potential to be unseen, depending on a particular segmental relationship. Assuming for the moment that otherwise-unseen words (similarly to unseen words) are good candidates for new words, P_{DE} expresses the degree of productivity of an affix based on the likelihood that a given word type with that affix will be new. These interpretations of otherwise-unseen words and P_{DE} are tentative ones. Full interpretations of them, which require an introduction of some new concepts, will be given in Sections 3.7 and 3.8.

A Venn Diagram usefully illustrates the elements involved in the P_{DE} measure, as shown in Figure 3-1. Symbols in the figure are explained in the following text.

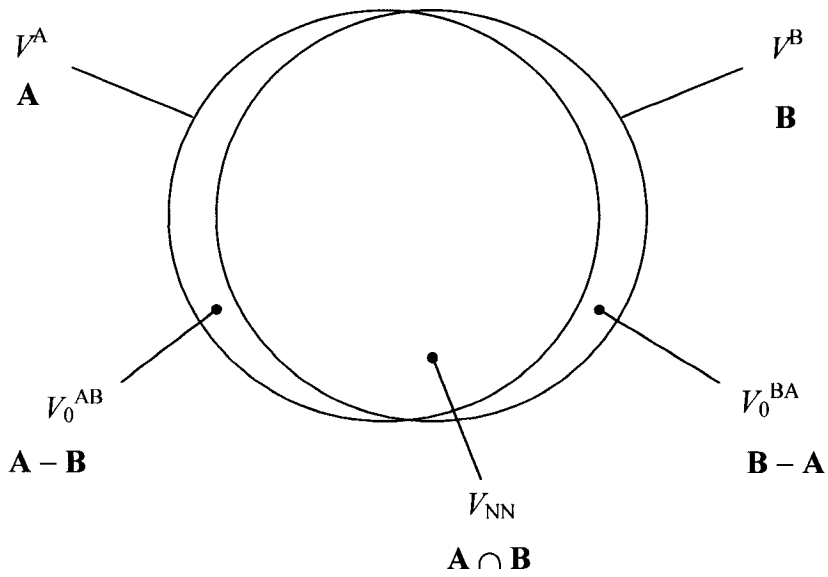

Figure 3-1. Venn Diagram illustration of elements of the P_{DE} measure.

Given a particular affix, **A** is a set of word types with that affix found in corpus segment *A*, and **B** is a set of word types with that affix found in corpus segment *B*. The

word types that are shared by the two segments (i.e., **A** ∩ **B**) will be called *common* word types. In contrast to otherwise-unseen word types, common word types are those that do not have the potential to be unseen; hence, they are not likely to be new. V_{NN} (V-Non-New) denotes the number of common, non-new word types. The relationship $V = V_N + V_{NN}$ holds in each application of the measure. V_{NN} does not directly participate in the calculation of P_{DE}, but it is one of the elements identified by the P_{DE} measure.

Implicit in formula (12) is a procedure in which the elements of the numerator and of the denominator are averaged separately, yielding V_N and V respectively, before the ratio of V_N and V is taken to give P_{DE}. It may be questioned why segment-specific ratios, V_0^{AB}/V^A and V_0^{BA}/V^B, are not first calculated and then averaged to obtain a value for P_{DE}. In fact, averaging two such ratios might be considered to lead to a more stable value of P_{DE}. However, this way of obtaining P_{DE} would be a choice for consideration only if our interest lay in P_{DE} as the sole element relevant to assessing productivity. Instead, our primary interest lies in the demarcation of new words in a given set of words, namely how V is separated into V_N (new word types) and its complement V_{NN} (non-new word types). P_{DE} as a productivity index is a way of presenting the information on what proportion of V is identified as new (V_N). Formula (12) attempts to reduce variance in V_N and V because obtaining stable values of V_N and V is of primary importance. An appropriate productivity index would naturally fall out from a proper identification of new words.

Another question that may be raised about the ratio V_N/V is why the basic unit for which productivity is measured is one corpus segment (of size m). Indeed, if we were to make maximal use of available data, an alternative to the averaging of values over two

segments could be to regard the sum of the two segments as the unit for which productivity is to be assessed. For example, we might introduce an element called V_U (V-Union) that refers to the number of word types in the two segments taken together, that is, the union of the two sets of word types, $\mathbf{A} \cup \mathbf{B}$, in Figure 3-1. However, such an approach has a serious problem in that the unit for which productivity is assessed will be unclear. It has been shown in the literature (e.g., Baayen, 2001: 2–4) that the size of V is dependent on the size of N (i.e., the sum of token frequencies) in sampled data. The value of V for an affix in a corpus segment of 10 million words, therefore, is normally expected to be smaller than the value of V for the same affix in a segment of 20 million words if the words are sampled from the same source. Suppose that each of the two segments A and B in Figure 3-1 has 10 million words. In our original procedure, it follows that V^A and V^B each pertains to a segment of 10 million words; V_0^{AB} and V_0^{BA} will be identified based on the comparison of A and B; and after averaging these values, the productivity of the relevant affix will be assessed for a segment of 10 million words. There is no element in this procedure that is defined with respect to a combined segment of 20 million words.

The identification of V_0^{AB} and V_0^{BA}, as presented in Figure 3-1, refers only indirectly to the total of 20 million words, and what yields V_0^{AB} and V_0^{BA} is a comparison of two segments of 10 million words. Now let us consider the consequence of introducing V_U. Once V_U is introduced, it will be unclear for what segment size the productivity of an affix is being assessed. V_U is the number of word types with an affix in a combined segment of 20 million words; V_0^{AB} and V_0^{BA} are identified via the cross-comparison of corpus segments, each 10 million words. A productivity index should not be based on an

interaction of V_0^{AB}, V_0^{BA}, and V_U because these are elements that belong to different-size samples.

It may also be questioned why the ratio V_N/V, instead of V_N only, is examined. In fact, V_N does offer information relevant for productivity, and as will be demonstrated in Section 5.1.2, V_N will prove useful in analyses of a particular kind. The advantage of examining the ratio V_N/V is that we take into consideration the fact that a different number of bases may be available for each affix. The ratio could be interpreted as posing the question: "Of all the derived words that have been formed with an affix, what proportion is new?"[19]

Consider affix *x* that has a V of 1000 and a V_N of 100, and affix *y* that has a V of 100 and a V_N of 20. If bare V_N were to be used as a productivity index directly, affix *x* (100) would be found to be more productive than affix *y* (20). On the other hand, if the ratio of V_N to V were to be used as a productivity index, affix *y* (index = 20/100 = 0.2) would be more productive than affix *x* (index = 100/1000 = 0.1). Finding affix *y* to be more productive than affix *x* (as in the latter case) is the desired outcome of the P_{DE} measure. An affix will not be regarded as being less productive simply because the word formation process for that affix is used less often to form a word (leading to smaller values for both V and V_N). Rather, an affix will be considered to be highly productive if a

[19] There is a simplification in interpreting V as the number of bases available for the word formation process for an affix because V may contain words, such as French loans, that were not formed with the relevant word formation process (see Bauer, 2001: 144). We may, therefore, need to interpret V_N/V as posing the less specific question, "Of all the words with an affix, what proportion is new?", abstracting away from whether each word with an affix was really formed by the word formation process for that affix.

derived word with that affix is highly likely to be new, regardless of how often the word formation process for that affix is used.

All the questions that have been raised above for the procedure implicit in (12) will be resolved by discussion in Sections 3.7 and 3.8, where further explanation will be given as to the implications of V_N and V_{NN}. Before proceeding with a detailed explanation of the design features of the P_{DE} measure, it is perhaps desirable to examine the performance of the P_{DE} measure on some actual corpus data. It is worth knowing whether the measure leads to promising or totally counterintuitive results. In the next section, the performance of the P_{DE} measure with some Chinese suffixes will be reviewed.

3.4 Productivity of Mandarin Chinese Suffixes

The performance of the P_{DE} measure in an analysis of five Mandarin Chinese suffixes is reported in Nishimoto (2003). The suffixes examined are listed in Table 3-1.

Table 3-1. Mandarin Chinese suffixes examined in Nishimoto (2003).

Suffix	Category	Prediction	Remarks
-hua	Verbal	Productive	Similar to English *-ize/-ify*
-men	Plural	Productive	Pluralizes human nouns only
-r	Nominal	Context-dependent	Diminutive, phonological
-zi	Nominal	Not productive	Historical ⎫ More *-zi* types
-tou	Nominal	Not productive	Historical ⎭ than *-tou* types

Predicted differences in productivity of these suffixes, as shown in the third column of Table 3-1, are based on characteristics of the word formation process associated with

each suffix and observations in the literature, both of which will be discussed shortly. The five suffixes will be introduced briefly below.

The verbal suffix *-hua* is similar to English *-ize* or *-ify*:

(13) *xiàndài* 'modern' + *-huà* → *xiàndàihuà* 'modernize'

Verbs formed with *-hua* can also be used as nouns (Baxter & Sagart, 1998: 40), so *xiàndàihuà* in (13) can also be interpreted as 'modernization' depending on the context. Analogous to English *-ize*, *-hua* quite regularly attaches to a noun to form a verb, such as *gōngyè* 'industry' + *-huà* → *gōngyèhuà* 'industrialize'; *jìsuànjī* 'computer' + *-huà* → *jìsuànjīhuà* 'computerize'.

The attachment of *-men* to a noun pluralizes it:

(14) *xuésheng* 'student' + *-men* → *xuéshengmen* 'students'

The status of *-men* as an equivalent of English plural suffix *-s* is questionable on three points (Lin, 2001: 59; Norman, 1988: 159; Ramsey, 1987: 64). First, the base is generally limited to human nouns; second, the suffix is obligatory in pronouns but not in nouns; third, the suffix is incompatible with numeral classifiers. Because of these characteristics, Lin (2001: 58) notes that *-men* may not be as frequently used or as "productive" as the English plural suffix *-s*. Nevertheless, *-men* has many possible bases among human nouns to which it can attach, as in *jìzhě* 'reporter' + *-men* → *jìzhěmen* 'reporters'; *kèrén* 'guest' + *-men* → *kèrénmen* 'guests'; and *shìzhǎng* 'mayor' + *-men* → *shìzhǎngmen* 'mayors'.

The suffix *-r* forms a noun from a verb or adjective, or it forms a diminutive noun (Ramsey, 1987: 63; Lin, 2001: 57–58):

(15) a. *huà* 'to paint' + *-r* → *huàr* 'painting'

b. *niǎo* 'bird' + *-r* → *niǎor* 'small bird'

The attachment of *-r* could also occur for stylistic (phonological) reasons, and the use of *-r* in this manner is abundant in the colloquial speech of local Beijing residents. In both Mainland China and Taiwan, the use of *-r* is generally avoided, especially in broadcasting (Chen, 1999: 39; Ramsey, 1987: 64), so the productivity of *-r* would be context-dependent.

The suffixes *-zi* and *-tou* formed nouns from bound morphemes when, in the history of the language, Mandarin underwent a transition from largely monosyllabic words to more disyllabic words (during the Han dynasty, 206BC to AD220; Packard, 2000: 265):

(16) a. **mào* + *-zi* → *màozi* 'hat'

b. **mù* + *-tou* → *mùtou* 'wood'

Li and Thompson (1981: 42–43) state that *-zi* and *-tou* in words of this type are considered to be suffixes only in the historical sense, and they are not synchronically productive. In addition to bound morphemes, these suffixes also formed nouns from free morphemes (Lin, 2001: 58–59; Packard, 2000: 84):

(17) a. *shū* 'to comb' + *-zi* → *shūzi* 'comb'

b. *xiǎng* 'to think' + *-tou* → *xiǎngtou* 'thought'

It is synchronically difficult, if not impossible, to form words such as those in (17).

Considering the relatively limited discussion of productivity in English, it is perhaps not surprising that discussion of productivity in Chinese is highly limited. Few explicit attempts have been made in the literature to quantify productivity in Chinese (e.g., Nishimoto, 2003; Sproat & Shih, 1996). The term "productive" occurs occasionally in the literature on Chinese morphology, but in most cases, it is not clearly stated in what sense the term is used. In fact, productivity often seems to be associated with the number of word types in a dictionary. Ramsey (1987: 63), for example, states that *-zi* is felt to be more productive than *-tou*, but it is not the case that a new word is easily coined with *-zi*. Ramsey's observation may perhaps correlate with the fact that there are more *-zi* words than *-tou* words in the present-day Mandarin. The productivity of *-zi* and *-tou* is predicted to be limited because it is quite difficult, if not impossible, to form a new word with these suffixes. Lin (2001: 57) views *-r* as the most productive suffix but does not offer any clear reason. Since the use of *-r* is expected to be limited in public broadcasting, the productivity of *-r* would be context-dependent. Both *-hua* and *-men* are predicted to be productive if we consider the fact that the word formation processes for these suffixes can be regularly and actively used (Spencer, 1991: 48–49), and also that a large number of bases is available for these processes. The productivity of *-hua* is also predicted to be high by analogy to English *-ize* and *-ify*.

The corpus used in Nishimoto (2003) is a revised version of the *Mandarin Chinese PH Corpus* (Guo, 1993; henceforth, the *PH Corpus*) made available by Hockenmaier and Brew (1998). The corpus has 2.4 million segmented Chinese words from XinHua newspaper articles. The use of the PH Corpus, with its relatively small size, was due to the limited availability of corpora of segmented Chinese texts.[20]

Results for the five target suffixes when the P_{DE} measure is applied to the PH Corpus of Mandarin Chinese are presented in Table 3-2. Figure 3-2 presents a graphic illustration of a productivity ranking of the suffixes, based on their P_{DE} indices. The PH Corpus was split in the middle to form the two corpus segments required by the P_{DE} measure.

Table 3-2. P_{DE} measure and its components (Nishimoto, 2003).

Suffix	V	V_N	P_{DE}
-men	149.0	70.0	0.470
-hua	144.0	65.0	0.451
-r	24.5	10.5	0.429
-zi	130.5	46.5	0.356
-tou	29.5	6.5	0.220

Note: Non-integer values for V and V_N arise from averaging across corpus segments.

[20] Chinese texts lack word delimiters, and their segmentation is a contested issue (see Sproat et al., 1996). A larger corpus, the *Academica Sinica Balanced Corpus* (1998; henceforth, the *Sinica Corpus*) of 5 million words would be an alternative corpus for use. The texts of the Sinica Corpus are syntactically parsed while those of the PH Corpus are not.

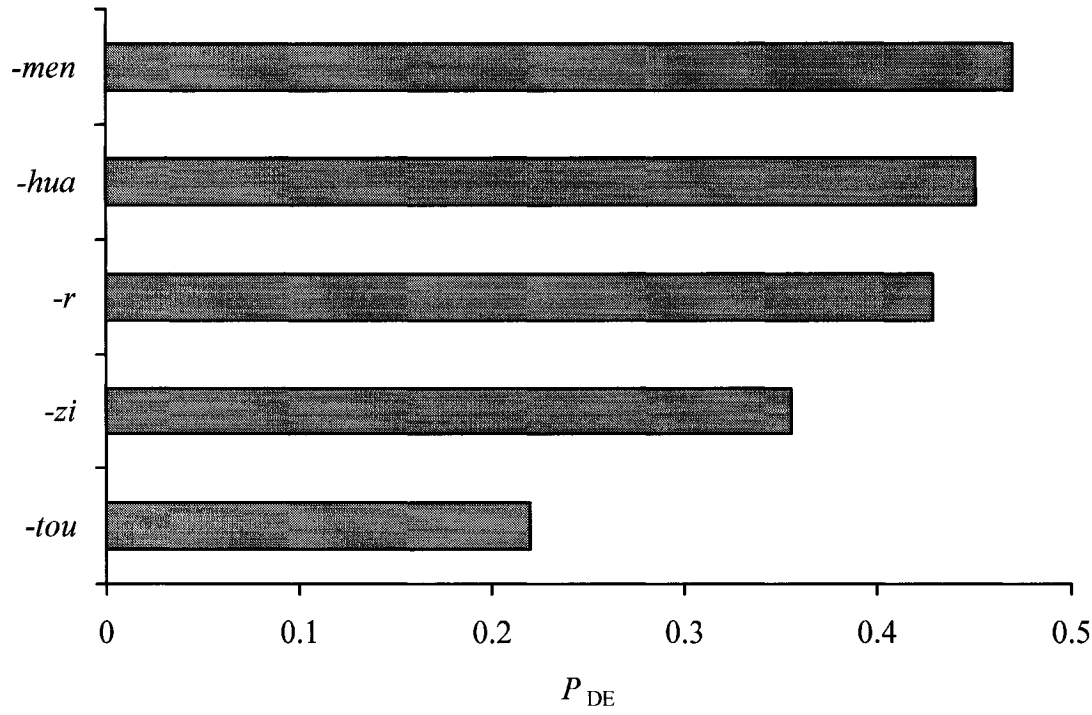

Figure 3-2. Productivity ranking of Mandarin Chinese suffixes.

The productivity ranking of suffixes based on P_{DE} meets our expectations, by and large. Certainly *-men* and *-hua* are shown to be more productive than *-zi* and *-tou*. The finding that *-zi* is remarkably more productive than *-tou* may have plausibility. The results for *-r* are somewhat unexpected: since the PH Corpus is a collection of newspaper articles, we might have expected to find limited use of *-r*. Yet the P_{DE} index for *-r* is close to that for *-hua*. The data in Table 3-2 indicate that although *-r* does not have a wide range of word types (V is small), when an *-r* word is encountered, it is reasonably likely that the word will be new.

Although overall the results for the P_{DE} measure on Chinese corpus data seem reasonable, four major limitations prevent a full evaluation of the performance of the measure. First, since the texts of the PH Corpus are not syntactically parsed, it was

impossible to distinguish *-hua* verbs from *-hua* nouns (i.e., both cases of *-hua* were included in the data). If only *-hua* verbs had been examined, the values for V and V_N may have been smaller, and it is unknown how the value for P_{DE} might have been affected. Second, a corpus of 2.4 million words is relatively small by today's standards. A larger corpus would be desirable, and particularly so if we wish to examine the stability of a measure over corpora of different sizes. Third, the PH Corpus was simply split down the middle to form two segments; this is due to the fact that the corpus consists of one single, very large file. Two segments created in this manner could well be dissimilar in text type and the number of word types (see Baayen, 2001: 210). As will be shown in Section 4.5, the most desirable way of creating corpus segments for the P_{DE} measure is to randomize documents (articles) of a corpus into two segments. Fourth, and most importantly, it is difficult to evaluate the performance of the P_{DE} measure because there is no objective method of judging whether a given description of productivity is acceptable.

In the next section, we will discuss how the difficulty in evaluating a productivity measure also holds in English, despite the fact that much more is known about English word formation than Chinese word formation.

3.5 Difficulty in Evaluating Results

Having conducted a preliminary test of the P_{DE} measure, we face one important question: How should a productivity measure be evaluated? We noted in the previous section that what is crucially missing is an objective method of determining whether given results of a productivity measure are acceptable.

Using WD, there is a way to examine how new words are captured by a productivity measure. Baayen and Renouf (1996) identify new words as those among their sampled words that are not listed in WD. They examined the token frequency of these new words for *-ness* and *-ity*, and found the majority of them to be hapaxes. We will test this WD-based method in Section 5.1.3. Examining how new words are captured only indirectly evaluates a given productivity measure; this is particularly so if productivity indices are not simply numbers of new words.

There seem to be two major factors that play into our evaluation of a productivity measure. One is intuitive judgments on productivity, and the other is observations in the literature. Baayen (1992: 110) suggests that for a given productivity measure to be of linguistic interest, one of the requirements that must be satisfied is that the measure "should provide a ranking of word formation processes that is in general correspondence with a ranking based on linguistic intuitions." He stipulates, for example, that *-ity* being ranked higher than *-ness* is "clearly unsatisfactory" (Baayen, 1992: 110).

It is also possible to evaluate results of a productivity measure based on whether the results meet observations in the literature. In fact, it is common for a study of productivity to refer to other researchers' statements about the productivity of a given affix. We must note, however, that those statements are themselves based on some (formal or informal) productivity measure, and they would also be subject to evaluation based on intuition.

In a sense, observations in the literature could be considered as a collection of views about the productivity of affixes that are to be agreed upon or refuted by researchers. If there is a view that is generally agreed upon, referring to such a view

provides us with a means, objective to some extent, for evaluating a productivity measure. Fundamentally, however, since observations in the literature are also subject to intuition-based validation, there seems to be no escape from linguistic intuition in evaluating a productivity measure. In the next section, we will discuss the nature of intuitions about productivity.

3.6 Intuitions about Productivity

Intuitions about productivity, despite their apparent role in evaluating productivity measures, have received only limited attention in the literature (Baayen, 1993).[21] Note that while linguistic intuitions are commonly discussed for grammaticality judgments in syntax, when it comes to intuitions about productivity, we must begin with questioning whether such intuitions are available to speakers. It seems likely that speakers that we refer to must be familiar with the basic concepts of morphology and the coinage of new words.[22]

The limited attention that intuitions about productivity have received in the literature is understandable if we consider the fact that intuitions may be unreliable and certainly lack quantitative precision. For example, speakers of English presumably

---

[21] Aronoff (1976: 37) states that "speakers of a language have intuitions about productivity," but his focus is on speakers' intuitive judgments on whether a given word is acceptable, as discussed in Section 1.3. The intuitions currently under discussion are those that might correlate with a productivity measure's finding for a set of words, those with some particular affix.

[22] Wheeler and Schumsky (1980) found in their experiment that native English speakers' ability to identify a suffix in a derived word varies dramatically. For example, while nearly all subjects correctly identified *-ness*, many failed to identify *-er* in *baker*.

cannot, on the basis of intuition, say to what extent *-ness* is more productive than *-ity*. However, with an appropriately guided introspection, speakers may be able to estimate that *-ness* should be "much more" productive than *-ity*. It is ironic that a quantitative productivity measure is crucially evaluated by intuitions that themselves are likely to lack quantitative precision. However, it would be premature to conclude that little can be learned from inquiries into the nature of intuitions on productivity. Baayen (1993) discusses the relationship between productivity and word frequency, and in his discussion, intuitions about productivity are often referred to. There are useful insights that we can glean from his survey.

Baayen (1993) first discusses intuitive evaluations of a productivity measure in his response to Van Marle's (1992) criticism of the data reported in Baayen (1992); Van Marle claims that some of those data are counterintuitive. Questions are raised on two points. The first question concerns the Dutch suffix *-er* (forming gender-neutral personal nouns and *-ster* (forming female personal nouns). While *-er* is usually held to be more productive than *-ster* on grounds of its semantic flexibility (Van Marle, 1992: 154), the *P* indices showed that *-ster* (0.231) is much more productive than *-er* (0.076). Second, the *P* indices found *-erd* (0.444), the supposedly non-productive suffix that is used to coin pejorative personal nouns (e.g., *gek* 'crazy' → *gekkerd* 'fool'), to be much more productive than noun compounding (0.225) that is associated with "a more or less 'automatic' kind of productivity" in Dutch (Van Marle, 1992: 155).

Van Marle questions the relevance of word frequency to productivity and the validity of *P* as a productivity measure. He sums up the situation as follows (1992: 156):

> On a general level of theoretical reflection I do not see what kind of direct relationship there is between the chance that a given rule is put into action and the frequency with which the words that have already been produced by that rule are used. Once a word is coined, the frequency in use of that word, it seems to me, is more or less irrelevant to the degree of productivity of that rule.

In response, Baayen (1993) argues that Van Marle's (1992) claim that the results are counterintuitive holds only if *P* were by itself expected to express the *global productivity* of an affix, which in his analysis is to be estimated taking into account values for both *P* and *V*. That is, Baayen's position is that *P* indices for the suffixes in question are not counterintuitive if the corresponding *V* values are also taken into account; whether a given set of findings is intuitive should not be determined on the basis of *P* alone. In fact, *V* is much higher for the Dutch suffix *-er* (299) than for *-ster* (30); and to a markedly greater extent, *V* is also higher for nominal compounding (4277) than for *-erd* (6). Baayen (1993: 187) claims that intuitions about degrees of productivity are primarily linked to global productivity. However, he also acknowledges that it is difficult to weigh the relative importance of *P* and *V* in any evaluation, and that a ranking of affixes cannot be obtained based on global productivity.

Baayen (1993) introduces a further productivity measure, *P**, formulated as follows:

$$(18)\quad P^* = \frac{n_1}{h}$$

Given a particular affix, n_1 is the number of hapaxes with that affix, and *h* is the sum of all hapaxes in the sampled corpus. *P** therefore expresses the contribution of a given

affix to the growth rate of the vocabulary as a whole. Since h will be constant in a given corpus, in effect, the productivity of affixes can be compared by n_1 alone. The measure P^* seems to lead to more intuitive results for the suffixes whose P indices were questioned by Van Marle (1992): n_1 is higher for the Dutch suffix *-er* (128) than for *-ster* (18), so P^* finds *-er* to be more productive than *-ster*; n_1 is also higher for nominal compounds (2591) than for *-erd* (4). The statistics P and P^* are distinguished as the *category-based degree of productivity* and the *hapax-based degree of productivity*, respectively.

The measure P^* appears to be suited for ranking productive affixes (Baayen, 1993: 193). For example, limiting our focus to English suffixes that we will examine later in the present study, the values of n_1 in Baayen (1993) provide the following productivity ranking: *-ness* (77), *-ation* (47), *-er* (40), *-ity* (29), *-ment* (9), *-ee* (2).

A more explicit reference to intuitions about productivity is made by Baayen (1993) in his discussion of the measure A, a productivity measure that is based on the *morphological race model* proposed by Frauenfelder and Schreuder (1992). According to that model, a given word may in principle be processed by either of two routes: one route which uses whole-word representations in the lexicon, and another which involves morphological parsing, and thus utilizes the representations of stems and affixes in the lexicon. Under the measure A, the extent of involvement of morphological parsing translates to the degree of productivity. The statistic A for an affix is calculated as the sum of token frequencies of the word types with that affix that fall below a token frequency threshold θ. The measure is formalized as:

(19) $$A = \sum_{r=1}^{\theta-1} r \times n_r$$

Given a particular affix, n_r is the number of word types with that affix that occur with frequency r, and $r \times n_r$ is the sum of token frequencies of the n_r word types. Baayen (1993) sets the threshold θ to 8 in English,[23] so that the value of A for *-ness*, for example, is the sum of token frequencies of *-ness* word types occurring 7 or less times in the corpus. In terms of the morphological race model, A for an affix is intended to capture the total number of times that the morphological parsing route was invoked in processing words of a given type, that is, the "activation level" of the parsing route for that affix. A higher value of activation level A indicates more involvement of morphological parsing, and greater involvement of morphological parsing, in turn, indicates a higher degree of productivity. With respect to the implications of A for intuitions on productivity, Baayen (1993: 204) remarks on its grounding in psychological notions:

> Interestingly, A is the only productivity statistic that is psychologically motivated. The kind of knowledge tapped into when speakers make intuitive productivity judgments appears to be closely linked to the resting activation levels of affix access representations. This might explain why such intuitions are of an ordinal rather than of an interval nature, that is, why we have intuitions about whether affix a_1 is more productive than affix a_2—given that there is a substantial difference in productivity—while we do not have intuitions concerning the exact number of actual or probable types with which a_1 exceeds a_2. Given the present theory, our ability to make these ordinal judgments is to be traced to the positive correlation between high activation levels and high numbers of actual (or possible) types.

[23] The threshold is determined in part based on corpus size and considerations on semantic transparency (Baayen, 1993). The rationale for choosing a particular threshold value is the consideration that frequent, semantically opaque words should not be subject to morphological parsing. See Hay and Baayen (2002) for further discussion.

Baayen (1993: 204) proposes that the measure *A* be used to rank productive word formation processes according to their degree of productivity. Again, limiting our focus to English suffixes that we will examine later in the present study, the values of *A* in Baayen (1993) offer the following productivity ranking: *-ness* (791), *-ity* (337), *-ish* (156), *-ment* (154).

Let us summarize Baayen's (1993) discussion about the relationship between productivity measures and intuitions about productivity. The primary use of the category-based measure of the degree of productivity, *P*, is to distinguish between productive and unproductive word formation processes (Baayen, 1993). In the light of Van Marle's (1992) criticism, we learn that *P* indices, considered in isolation, should not be evaluated against intuitions. Global productivity, taking into consideration both *P* and *V*, is more closely tied with intuitions, but its interpretation cannot be precise because no means of weighting *P* and *V* components has been formulated. The hapax-based measure of the degree of productivity, *P**, appears to lead to intuitive results, but the measure *A* is directly psycholinguistically motivated and can be used to rank word formation processes according to their degree of productivity (Baayen, 1993: 204). The measure *A*, therefore, has the closest link to intuitions about productivity. A further development of the measure *A* is seen in Hay and Baayen (2002) where they study in more detail the relationship between productivity and psycholinguistic mechanisms of morphological parsing.

For Baayen (1993), in short, we find that different productivity measures capture different aspects of productivity, and that these tap into intuitions about productivity in different ways. A finding that some measure yields results which are intuitive (or counterintuitive) leads to meaningful inquiries about the nature of intuitions about

productivity, and in turn, to the development of new measures. Although intuitions must in the end play a primary role in evaluating any productivity measure, it would not be constructive to accept or reject any productivity measure, solely on the basis of our current understanding of intuitions. As long as intuition remains imprecise, and its mechanism unknown, it cannot provide any "gold standard" against which to assess the validity of productivity measures.

The most useful research agenda would appear to be one in which a variety of productivity measures is proposed, so that a clear account of what aspect of productivity each measure captures can be developed. Then, some systematic map between what results are intuitive or counterintuitive can emerge: a well-understood measure providing intuitive results and another, counterintuitive, are both valuable in that either may lead to appropriate and motivated refinements in the underlying theory. This approach may lead to insights about the nature of intuitions (what information is or is not available to speakers' introspection).

Ultimately, a productivity measure that we seek may not necessarily be one that merely satisfies intuitions, but one with a clear account of what aspect of productivity the measure captures, so that we learn from asking *why* its predictions are intuitive or counterintuitive. The responsibility of a researcher proposing a productivity measure should be to ensure that the sense in which productivity is captured is made conceptually clear. In this respect, the P_{DE} measure requires further explanation. Our formulation of the P_{DE} measure has taken a strictly technical viewpoint, that of capturing unseen words, without any linguistic interpretation of just what it is that is depicted by the measure. In

the next two sections, it will be made clear what the P_{DE} measure is intended to capture, from a linguistic point of view.

3.7 Usage Commonality of Words

To this point, the ratio V_N/V of the P_{DE} measure has been said to express the likelihood that a given word type with an affix will be otherwise-unseen, or new. However, we need a justification for associating otherwise-unseen words with new words. Otherwise-unseen words do not fit the usual corpus-based definition of unseen words: more standard, unseen words are those that have not yet been sampled in a corpus; otherwise-unseen words are sampled within a given corpus.

What does V_N really represent? We first take note of the fact that a corpus as a whole represents words used by many speakers.[24] If we think of two cross-compared segments as two groups of sampled speakers, words that are not shared by the two segments are those that are not used in common by the two groups of speakers, whereas words shared by the two segments are those that are used in common by the two groups of speakers. Although the distinction being made between shared and unshared words is a crude one when only two segments are compared, the point being made here is that if we are to characterize words in terms of their usage by speakers, words not shared by the two segments are more likely to be those not used by many speakers. In terms of the newness

[24] Since corpus data are drawn from written texts in general, we might use the term *authors* rather than *speakers*. However, for consistency with the literature, we will continue to refer to text-sources as *speakers*.

of words, words not used in common by the two speaker groups are more likely to be new.

To consider two segments as representing two essentially similar groups of speakers, it does not suffice to split a given corpus at its middle point: the text type of a corpus could well differ among different components of the corpus, so that the first and second segments created by a split of a corpus could represent markedly different word usages (e.g., the first segment might be drawn from technical journals, and the second segment from fiction). Randomizing the words of a corpus into two segments is not advisable for our current purposes because words used by a single speaker are highly likely to be distributed across segments. Randomization-by-words would undermine the point of cross-comparing segments, if that comparison is to be between speaker groups.

Given, then, that the set of words used by any speaker is to be sampled in one or other (but not both) of two segments, an alternative is to randomize the corpus by documents. Document (article) boundaries in a corpus (assuming that such boundaries are available) enable us to identify a set of words that belongs to a speaker (or co-speakers, in the sense of co-authors). By randomly distributing documents into segments, we can form groups of randomly sampled speakers (and words that they used). Section 4.5 elaborates on the issue of corpus-segment creation, with actual corpus data.

To gain a better understanding of the P_{DE} measure, we now turn to develop what will be called the *Usage-Commonality* (*UC*) measure. The number of segments created from a corpus and cross-compared to distinguish shared/unshared words is not limited to two. If we increase the number of cross-compared segments, some words will appear in fewer segments, and others in more segments. Unlike the 2-segment case where words

are either shared or unshared, an *n*-segment case ($n > 2$) can identify varying degrees to which words are shared by corpus segments. The usage commonality of words then refers to the occurrence of words across groups of speakers (i.e., corpus segments), and the UC measure expresses usage commonality in this sense based on how many corpus segments share a given word.

Let us examine how we proceed with collecting data relevant to the UC measure. Given *n* segments of a corpus (all segments having the same size in token count), we let *Sr* represent the number of word types that appear in *r* segments, for $r = 1...n$. Figure 3-3 identifies the elements involved in a 3-segment case; identifying symbols are defined in the text following the figure.

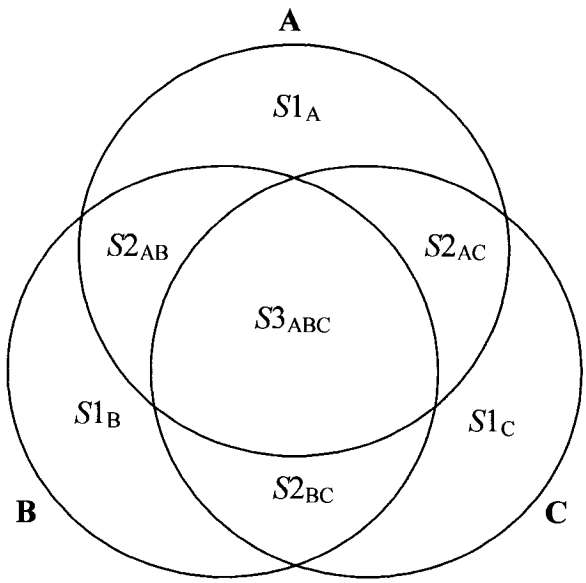

Figure 3-3. Elements involved in cross-comparing 3 corpus segments.

In Figure 3-3, symbols **A**, **B**, and **C** represent sets of word types with a particular affix identified in segments *A*, *B*, and *C*, respectively. Symbols *S*1, *S*2, and *S*3 represent

the number of word types appearing in exactly 1, 2, and 3 of those segments, respectively, and an additional letter subscript, as in $S1_A$ and $S2_{AB}$, identifies the segment(s) in which those word types occur. There are two important points to note in Figure 3-3. First, the basic unit for which we collect the data of *Sr* is one segment (of size *m* tokens) and not the sum of all segments. Second, the *Sr* data for each segment will be collected irrespective of which other segments share the word types.

Figure 3-4 illustrates how *Sr* data are collected for **A**. In obtaining *S*2 data, for example, $S2_{AB}$ and $S2_{AC}$ are summed to give an *S*2 value for **A**. In other words, *S*2 for **A** is the sum of word types that **A** shares with either **B** or **C**. As indicated in Figure 3-4, *S*sum denotes the sum of *Sr* values for an affix in a segment, hence:

(20) $S\text{sum} = \sum_{r=1}^{n} Sr$

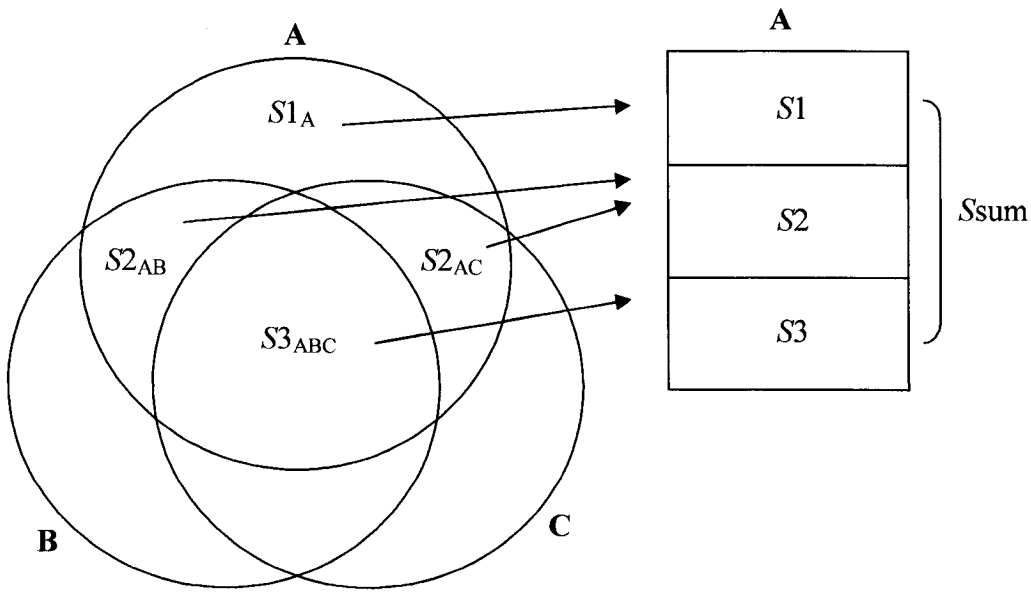

Figure 3-4. Illustration of *Sr* data collection.

To make maximal use of available data and to obtain stable prototypical *Sr* values for a segment of size *m*, we repeat the procedure in Figure 3-4 for **B** and **C**, and average the obtained values for *S*1, *S*2, *S*3, and *S*sum over the three segments. With an assumption that this averaging is always carried out, we will henceforth refer to the mean values as *S*1, *S*2, *S*3, and *S*sum (without explicit indication that these are mean values).

As was the case in the P_{DE} measure, the basic unit for which the relevant data are collected is a single segment, not the sum of segments. A disadvantage of referring to the entire set of available data has already been discussed for the P_{DE} measure, and that disadvantage is even clearer in the case of the UC measure. Similarly to our previous discussion of the same issue for the P_{DE} measure, let us use V_U to refer to the union of the three sets of word types **A**, **B**, and **C** (i.e., **A** ∪ **B** ∪ **C**) in Figure 3-3, and consider that each of **A**, **B**, and **C** draws from a segment of 10 million words. When the basic unit for collecting *Sr* data is one segment (as in the original proposal), there will be no element that pertains to a combined segment of 30 million words. *S*sum is the total number of word types with an affix in a segment of 10 million words, and *S*1, *S*2, and *S*3 are identified based on a comparison of 10-million-word segments. Thus, *S*1, *S*2, *S*3, and *S*sum all reflect information with respect to segments of 10 million words. By contrast, V_U refers to a combined segment of 30 million words, while *S*1, *S*2, and *S*3 are still identified based on the cross-comparison of 10-million-word segments. For this reason, an analysis that uses *S*1, *S*2, *S*3, and V_U in a mixture would be difficult to evaluate. By referring to the entire available corpus (as in V_U), we would lose sight of what corpus-segment size a given analysis of *Sr* belongs to.

Moreover, once we refer to the entire set of available data (as in V_U), increasing the number of segments will not make sense. As the number of segments is increased to 4 and 6, say, V_U will belong to a combined segment of 40 million words and 60 million words, respectively, although the *Sr* data are still defined over segments of 10 million words. The relationship among the different-number cases will not be clear. Having the size of V_U constant and changing the size of each segment as the number of segments increases offers no solution: while V_U may refer to a combined segment of a certain size, *Sr* data will be defined over segments of different sizes as the number of segments increases.

If we fix the size of each segment, and take the size of one segment as the unit for which *Sr* data are collected (as in the original proposal), values for *S*1 to *S*4 in the 4-segment case and values for *S*1 to *S*6 in the 6-segment case will be consistently defined relative to 10-million-word segments, whether the number of segments is 4 or 6. In this manner, a comparison among cases differing in the number of segments will be meaningful.

Let us now examine how the usage commonality of words is assessed, for the 6-segment case. After we obtain values for *S*1 to *S*6 and *S*sum (averaged over 6 segments), we can consider the segment-frequency categories *S*1 through *S*6 as forming a scale, as shown in Figure 3-5.

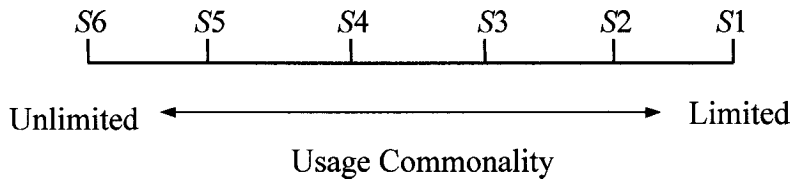

Figure 3-5. Illustration of usage-commonality scale.

The scale in Figure 3-5 reflects varying usage commonality among words, that is, the distribution of the usage of words across 6 groups of randomly sampled speakers. Word types contributing to the *S*6 value are unlimited in usage commonality (i.e., their usage is maximally widespread), since such words are shared by all 6 groups of speakers. In terms of the newness of words, word types contributing to the *S*6 value are least likely candidates for new words. As we move to *S*5, *S*4, and so on, the usage distribution becomes narrower, that is, the usage commonality of words becomes more limited. When we reach the end of the scale, word types contributing to the *S*1 value are the maximally limited in usage commonality, since they are used by one speaker group only. These are most likely candidates for new words.

The idea that we may wish to ultimately associate with usage commonality may be the likelihood that any speaker is familiar with a given word. The varying number of speakers to whom words may be familiar is seen in Bauer's (2001: 36) use of the term *item-familiar*:

> Let us say that a word is an EXISTING WORD from the moment it is first coined. The word may be ITEM-FAMILIAR to individual speakers, without having become part of the norm of the language. A word is ESTABLISHED once it becomes part of the norm, that is, once it is item-familiar to a large enough sub-set of the speech community to make it worth listing in reference works.

An item-familiar word is a word that is recognized by a speaker as a lexeme with a fully specified meaning; for example, *telephone box* typically means one particular thing for any speaker to whom it is item-familiar (Bauer, 1983: 48).[25] It is evident from the passage quoted above that, in Bauer's (2001) view, there are degrees to which a word can be item-familiar. At one end of the continuum, a word might be item-familiar to only one individual speaker (who coins the word); at the other end, a word might be item-familiar to every speaker. It is conceptually attractive to associate new words with words that are item-familiar to the least number of speakers. However, as we confine our data to a corpus of texts in which only the use of words (not the perception of words) can be examined, there is still an inferential gap between the usage commonality of words, and the number of speakers to whom words are item-familiar. For this reason, we will maintain the term "usage commonality," and simply note that the motivation for associating new words with words that are most limited in usage commonality is similar to the motivation for associating new words with words that are least likely to be item-familiar.

The consequence of cross-comparing many segments instead of just two is that we can characterize words in a corpus by the (varying) likelihood that they could be new, based on usage commonality. New words are no longer identified categorically, and the newness of words is instead expressed in a gradient manner. Earlier, we noted that a fundamental problem with defining new words lies in how a judgment is to be made about just which words are new. Identifying new words in a graded, non-categorical

---

[25] See Meys (1975: 61–77) for a discussion of the term *item-familiarity* at the time of its introduction to the literature.

manner is a well motivated response to the difficulty of categorically delimiting new words.

There are two elements to be analyzed in the *Sr* data: (i) the number of word types for any *Sr*, and (ii) the ratio of *Sr* to *S*sum. The choice between (i) and (ii) resembles the choice we had earlier between V_N and V_N/V in the P_{DE} measure. We will pursue the ratio analysis (ii), just as we primarily use V_N/V in the P_{DE} measure. The ratio *S*1/*S*sum, for example, tells us the proportion of word types with an affix that are most likely to be new. For notational simplicity, we will refer to *Sr*/*S*sum as %*Sr* (the % of *Sr* in *S*sum).

When %*Sr* data are analyzed along the usage-commonality scale, a simple prediction is that a productive affix will be characterized by %*Sr* values that increase toward *S*1. This is illustrated in Figure 3-6. In other words, a productive affix is expected to be one that has an increasing percentage of word types that are new (or likely to be so). Another prediction to be made about %*Sr* is that the values of %*Sr* as *S*1 is approached on the usage-commonality scale should be greater for a more productive affix. Section 5.2.2 will examine these predictions with actual corpus data.

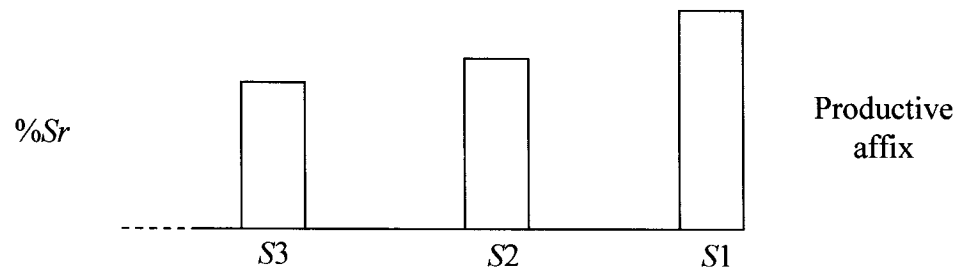

Figure 3-6. Predicted %*Sr* values as a function of usage commonality.

3.8 Productivity Measure Revisited

Having introduced the UC measure to reflect the usage commonality of words, we are now in a better position to understand what the P_{DE} measure captures. Figure 3-7 shows the usage-commonality scale in the 2-, 4-, and 6-segment cases.

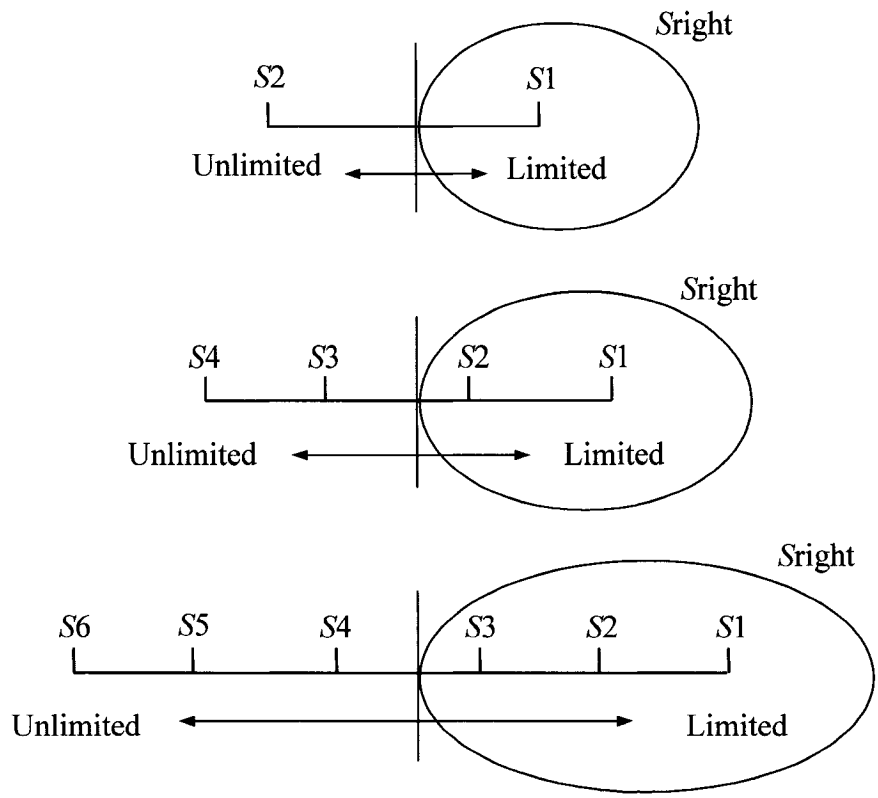

Figure 3-7. Usage-commonality scales for 2-, 4-, and 6-segment cases.

The 2-segment case in the top panel of Figure 3-7 corresponds straightforwardly to the P_{DE} measure where the following relationships hold: (i) *S*1 is V_N; (ii) *S*2 is V_{NN}; (iii) *S*sum is V; and (iv) *S*1/*S*sum (or %*S*1) is V_N/V or P_{DE}.

Where the UC measure involves more than 2 segments (as in the middle and lower panels of Figure 3-7), the predicted relationship with P_{DE} is as follows. As the

number of segments increases, there will be finer degrees with which the usage commonality of words can be determined. Let *S*right denote the sum of *Sr* values for the right half of the usage-commonality scale (i.e., the circled elements in Figure 3-7). When the size of each corpus segment is constant (e.g., 10 million words), we predict that values of *S*right will be highly similar among the 2-, 4-, and 6-segment cases. If that is the case, the 2-segment case reveals at coarse grain what could be obtained in more finely differentiated treatments involving more segments. Despite the inherent crudeness in determining usage commonality based on only 2 segments, the computational simplicity of the 2-segment case is suitable for generating overall productivity indices. Hence, the 2-segment case of the UC measure receives a special status as the P_{DE} measure. With the above concept in mind, P_{DE} is redefined as:

(21) P_{DE} expresses the likelihood that a given word type with an affix will be limited in its usage distribution over groups of speakers, and hence will be likely to be new.

We have now obtained a clearer interpretation of new words captured by the P_{DE} measure: new words are those that are limited in their usage distribution. Now that we have pinned down a less technically oriented interpretation of what is captured by V_N in the P_{DE} measure, words contributing to V_N will henceforth be called new words (rather than otherwise-unseen words), and the newness of those words must be interpreted in the sense that has been explained in the preceding sections of this chapter.

3.9 Summary

In search of a method with which to capture unseen words within corpus data, we reviewed smoothing techniques in Language Technology. The deleted estimation method was incorporated into a productivity measure based on type frequency. That productivity measure, P_{DE}, is formulated to cross-compare two corpus segments, and designates unshared word types with an affix as potentially new words with that affix. The performance of the P_{DE} measure on some Mandarin Chinese suffixes was reviewed. In the light of difficulty in evaluating the results of the measure, we discussed the roles that intuitions play in evaluating findings for any productivity measure, and stressed the importance of clarifying the design features of any such measure. To gain a clearer understanding of what it is that the P_{DE} measure captures, the mechanism of the P_{DE} measure was extended into the UC measure, where the usage commonality of words is assessed by cross-comparing multiple segments. Taking corpus segments to represent groups of randomly sampled speakers, words shared by fewer segments are those more limited in usage commonality, hence more likely new. The relationship between the P_{DE} measure and the UC measure was explained. The methods of assessing productivity proposed in this chapter will be put to test in Chapter 5.

Chapter 4 English Suffixes in the British National Corpus

The main purpose of this chapter is to describe in detail the ways in which the data of the BNC are processed for analysis in the present research. The fact that the entire chapter is devoted to describing the processing of the corpus data is an indication of its importance. A corpus-based productivity measure is dependent on frequency data, and frequency data are sensitive to many properties of a corpus, some of these idiosyncratic, that can all too easily distort findings for the measure. Technical issues that arise in conducting a corpus-based study of productivity are discussed in the literature (Evert & Lüdeling, 2001; Plag, 1999: 107–110; 2003: 74–78), and many of these issues are addressed in this chapter.

An advantage of having a detailed account of how corpus data are processed is that if certain results for a productivity measure are found to be peculiar, their cause could potentially be traced to particular procedures used in processing corpus data. If the procedures for processing corpus data are obscure, there will be little room for refining the present findings in future research.

4.1 Twelve Target Suffixes

Table 4-1 lists 12 derivational suffixes of English and 1 non-suffix control, which are to be examined in the present study. Where appropriate, the suffixes are grouped into rival pairs to facilitate later analysis of data. At least one suffix is examined for each of the four major lexical categories: noun, adjective, verb, and adverb. Predicted differences

in productivity in the third column of Table 4-1 are largely based on views expressed in the literature.

Table 4-1. English derivational suffixes and a non-suffix control, utilized in the study.

Suffix	Category	Prediction
-ness vs. *-ity*	Nominal	*-ness* > *-ity*
-er vs. *-ee*	Nominal	*-er* > *-ee*
-ion vs. *-ment*	Nominal	*-ion* > *-ment*
-th	Nominal	Unproductive
-ize vs. *-ify*	Verbal	*-ize* > *-ify*
-ish vs. *-ous*	Adjectival	*-ish* > *-ous*
-ly	Adverbial	Productive
ch#	Non-suffix, noun ending	Unproductive

We limit our focus to suffixes for two reasons: (i) suffixes (as opposed to prefixes) play a major role in English derivational morphology (Carstairs-McCarthy, 2002: 55), and (ii) limiting the focus to suffixes has some technical advantages in processing words in a corpus, as will be clear in later sections. Below, we review discussions in the literature on the productivity of these suffixes. Plag (2003: 86–101) offers concise descriptions of a number of English affixes including those examined here.

Nominal Suffixes *-ness* and *-ity*. In general, *-ness* is intuitively felt to be more productive than *-ity* (Aronoff, 1976: 43). The pairing of *-ness* and *-ity* has provided a prototypical comparison for which the degree of productivity is known to differ, and for which many discussions have been offered in the literature (see, e.g., Aronoff, 1976: 37–45). The suffix *-ness* is considered as one of the most productive affixes in English (Plag,

2003: 92). The word formation process for *-ity* is subject to the Latinate Restriction (Aronoff, 1976: 51–52; Plag, 1999: 57–60; see Section 1.2.2), which requires the base to have the [+Latinate] feature. By contrast, *-ness* freely attaches to both Latinate [+Latinate] and Germanic [−Latinate] bases. Also, as discussed in Sections 1.2.1 and 1.2.2, Aronoff (1976: 43–45) points out that a derived word with *-ness* is semantically more coherent, and that many word forms with *-ity* are blocked by existing words. Both *-ness* and *-ity* attach to an adjective to form a noun. In addition, *-ness* may also attach to a noun (e.g., *thingness*), a pronoun (e.g., *us-ness*), or a phrase (e.g., *over-the-top-ness*), as discussed in Plag (2003: 92).

It has been pointed out (e.g., Aronoff, 1976; Aronoff, 1983; see Section 1.3) that the productivity of *-ness* and *-ity* varies with the class of bases that the suffixes attach to. In particular, *-ness* is claimed to be more productive than *-ity* given *-ive* bases, and *-ity* is more productive given *-ible* bases. In the present study, the productivity of *-ness* and *-ity* will first be assessed, overall, irrespective of the class of bases that the suffixes attach to, and then the data will be further refined to examine these particular claims about *-ness* and *-ity* sub-patterns.

Based on *P* indices, Baayen and Lieber (1991) find *-ness* (0.0044) to be more productive than *-ity* (0.0007). Nevertheless, as compared with the *P* index for simplex nouns (0.0001), *-ity* was still found to have some degree of productivity.

Nominal Suffixes *-er* and *-ee*. Intuition suggests that *-er* is more productive than *-ee*. This may be in part affected by the larger number of word types we find for *-er* (e.g., 682 for *-er* vs. 23 for *-ee* in the study reported by Baayen & Lieber, 1999). A distinction is

sometimes drawn between *-er* words that denote agentive/human nouns (e.g., *baker*) and *-er* words that denote instrumental/non-human nouns (e.g., *heater*), but that distinction is not clear-cut in many cases (Bauer, 2001: 199–203). An example of such an ambiguous case is *walker*.[26] Ryder (1999) views the two types of *-er* words as resulting from the same word formation process. In addition, Ryder stresses the importance of considering *-er* words derived from non-verbs, despite the approach taken by many studies to consider only *-er* words derived from verbs. In the present study, we will treat all *-er* words collectively, without distinguishing agent/human nouns from instrumental/non-human nouns, and by including cases where *-er* attaches to a non-verb, such as a noun (e.g., *islander*), an adjective (e.g., *foreigner*), a preposition (e.g., *downer*), and a phrase (e.g., *two-footer*), as discussed in Ryder (1999: 270).[27] The rationale for this inclusive treatment of *-er* words is twofold. First, some suffixes such as *-ness* and *-ish* also attach to bases drawn from different lexical categories, and in the present study we will collectively treat those *-ness* and *-ish* words without making distinctions among subtypes. Second, later we wish to ask questions about what implications follow when results for a productivity measure seem intuitive or counterintuitive, and in asking such questions, a collective treatment of *-er* words seems appropriate; we are not certain whether intuitions are available for particular subclasses of *-er* words. As an initial approach, we may include all *-er* words in our analysis, including *-er* words derived from

---

[26] This word could mean "a person who walks" or "a device that helps a person walk."

[27] Comparative forms in *-er* (e.g., *happier*) are, of course, excluded from our analysis, based on the word-class (part-of-speech) tags of the BNC, as will be discussed shortly. For ambiguous forms such as *cleaner*, we also rely on word-class tags to identify the nominal suffix *-er*.

non-verbs. Many researchers (e.g., Carstairs-McCarthy, 2002: 51; Plag, 2003: 89) join with Ryder in regarding various *-er* words as belonging to the same suffix *-er*.

Excluded from our analysis of *-er* are *-or* (as in *conductor*)[28] and *-eer* (as in *auctioneer*).

Various *-ee* words will also be treated collectively. It has proven difficult to describe the word formation process for *-ee* as a uniform process; for example, we can contrast *employee* "a person who is employed" and *escapee* "a person who escapes," and *-ee* also attaches to a noun or adjective, as in *patentee* and *absentee*. Barker (1998) offers a semantic account, in place of a syntactic account, of words derived with *-ee*. The suffix *-ee* is called "episodic *-ee*" and denotes a sentient entity that is involved in an event as a non-volitional participant. Since the derived word must have a sentient referent, *amputee* is someone whose limb is amputated, not the limb that is amputated (Plag, 2003: 88). The only *-ee* words that will be excluded are those where *-ee* is a variant of the diminutive suffix *-ie*, as in *goatee*.

Baayen and Lieber (1991) examined *-er* and *-ee* words derived from verbs, and found *-ee* (0.0016) to be more productive than *-er* (0.0007), which they initially found to be somewhat surprising. However, Baayen and Lieber (1991: 828) attribute the high productivity of *-ee* to the fact that *-ee* is a "vogue" suffix, as suggested by Marchand (1969: 210).

---

[28] The suffix *-or* is an orthographical variant of *-er* that occurs mainly with Latinate bases ending in /s/ or /t/, as in *compressor* and *conductor* (Plag, 2003: 89).

Nominal Suffixes *-ion* and *-ment*. These suffixes attach to a verb to form a noun. Unlike *-ness* and *-ity*, it is not immediately clear what the intuitive estimates of the productivity of these suffixes are. The suffix *-ion* is subject to the Latinate Restriction (Aronoff, 1976: 36), and involves a stress shift (Plag, 2003: 90–91), while no stress shift is involved in *-ment* word formation. Aronoff (1976: 53) points out that *-ment* may take a verb base formed with *en-* (e.g., *encourage* + *-ment* → *encouragement*) and *be-* (e.g., *besiege* + *-ment* → *besiegement*), although it has been suggested that these prefixes are themselves no longer productive (Marchand, 1969: 332).

An OED-based analysis by Bauer (2001: 8) suggests that new words in *-ment* have rarely been coined in recent years.[29] Bauer's analysis indicates that *-ment* had two peaks in productivity in the history of English, the early 17th century and the early 19th century, and that the productivity of *-ment* has since been in decline.

Excluded from our analysis of *-ion* are words such as *accordion*, *onion*, *companion*, *million*, and so on (see Aronoff, 1976: 99).

Baayen and Lieber (1991) examined *-ation* and *-ment* words and found *-ation* (0.0006) to be more productive than *-ment* (0.0002). The *P* index for *-ment* is very close to the baseline *P* index (0.0001) provided by simplex nouns, so by their definition, *-ment* was found to be barely productive.

Nominal Suffix *-th*. Suffix *-th* is treated almost unanimously in the literature as a prototypical example of an unproductive affix (e.g., Anshen & Aronoff, 1989; Aronoff &

[29] Bauer (2001: 8) finds *underlayment*, and Plag (1999: 77), in addition, finds *endistancement* and *tracklement* as the only *-ment* words formed after 1950.

Anshen, 1998; Bauer, 1988: 59–60; Matthews, 1991: 70, 79; Spencer, 1991: 49). Aronoff and Anshen (1998: 243) point out that *-th* has not been successful in coining a new word for some 400 years, despite occasional attempts in terms like *coolth*.[30] Baayen (2003: 235–236), nevertheless, points out that a search on the Web reveals some uses of *-th* words such as *greenth* and *gloomth* that show "the residual degree of productivity of *-th*" (Baayen, 2003: 236). Sporadic uses of the word formation process for *-th* could be described as being *individually productive* (Bauer, 1988: 65); that is, a particular speaker may find a productive use of a given affix in a limited occasion, although that affix may not be productive for the language community more generally. Matthews (1991: 70, 79) argues that new coinages with *-th* are possible only by analogy, and Spencer (1991: 49) suggests that it is doubtful whether *-th* could be regarded as a genuine morpheme in contemporary English.

Excluded from our analysis of *-th* are ordinal words with *-th*, as in *5th*, *9th*, or *100th*.

Verbal Suffixes *-ize* and *-ify*. The literature treats suffix *-ize* as being more productive than suffix *-ify* (e.g., Bauer, 1983: 222–223), and intuition agrees. Plag's (1999: 104) OED-based analysis indicates that there were more new words coined with *-ize* (346) than with *-ify* (30), between the years 1900 and 1985. The suffix *-ify* prefers Latinate bases (Plag, 2003: 84–85). Plag (1998, 1999: 94; 2003: 93–94) finds a diversity of semantic patterns in derived words with *-ize* and *-ify*. We treat the British spelling *-ise* as *-ize*, as will be discussed later. Bauer's (2001: 184–185) OED-based analysis of the

[30] According to the OED, *coolth* was first attested in the mid 16th century.

patterns *-ization* and *-ification* shows that more new coinages for *-ization* than for *-ification* have been attested in the last two centuries. Based on this finding, Bauer (2001: 184) observes that "*-ify* seems to be becoming less important as a verbaliser, while *-ise* is increasing in use."

Baayen and Lieber (1991) find *-ize* (0.0000710) to be more productive than *-ify* (0.0000000, i.e., no hapax), albeit with very low *P* indices for both suffixes. The fact that the *P* index for *-ify* is lower than the *P* index for simplex verbs (0.0000065) indicates, by their definition, that *-ify* is not productive at all.

Adjectival Suffixes *-ish* and *-ous*. The suffix *-ish* is intuitively felt to be very productive. Our analysis of *-ish* includes not only cases where the base is a noun (e.g., *babyish*) but also other cases where the base is an adjective (e.g., *clearish*), a numeral (e.g., *thirtyish*), an adverb (e.g., *soonish*), and a phrase (e.g., *out-of-the-wayish*), as discussed in Plag (2003: 96). The suffix *-ous* is subject to the Latinate Restriction, and attaches to nouns and bound roots of Latinate origin (Plag, 2003: 97). Aronoff and Schvaneveldt (1978: 107) note that only a limited number of *-ous* words have been coined since 1800 (e.g., *magnetiferous*, *edacious*, and *scrumptious*).

Somewhat problematic for our corpus statistics is that *-ish* attaches to a numeral base, as in "He looks *30ish*" or "Let's meet at *10:30ish*?" As far as *10:30ish* is concerned, we certainly do not wish to accept all other variants of the same type, such as *11:00ish* or *12:30ish*, as distinct word types with *-ish*, so we transform *10:30ish* and other variants into *xx:xxish*, where *xx:xxish* will represent all the variants of this type. In actuality, the BNC has only one instance of *10:30ish* (written as *ten-thirty-ish*), and it does not have

any other word of this type. However, this example shows how important it is to have a principled treatment of certain types of words to prevent an unwarranted increase in form variations and the number of word types with an affix.[31] The same treatment applies to words such as *30ish*, *40ish*, or *13ish*, and so on; these words are transformed into *xxish*. There are 9 words of this type.[32] Note that the number of word types with *-ish* would have increased by 9, instead of by 1, if not for the transformation of these words into *xxish*.

Baayen and Lieber (1991) find *-ish* (0.0050) to be much more productive than *-ous* (0.0006). Nevertheless, the fact that the *P* index for *-ous* is substantially greater than the *P* index for simplex adjectives (0.0001) suggests that *-ous* may not be completely unproductive.

Adverbial Suffix *-ly*. The adverbial suffix *-ly* is regarded as one of the most productive affixes in English (Aronoff, 1976: 36; Anshen & Aronoff, 1989: 197). Almost any adjective takes *-ly* to form an adverb, although in a limited number of cases, an adjective base that already ends in the adjectival suffix *-ly* (e.g., *friendly* or *lovely*) is generally avoided as it would lead to a sequence of *-lily* (Aronoff, 1976: 37 fn 4; Bauer, 1992; 2001: 6, 130; Baayen & Renouf, 1996: 82–83). Some researchers suggest that *-ly* might even be considered to be an inflectional affix (Aronoff & Furhop, 2002: 481–482;

---

[31] We certainly do not wish to regard the form variations in *10:30ish*, *11:00ish*, *12:30ish*, and so on, to be relevant to the productivity of *-ish*, although the fact that *-ish* attaches to a time expression to form a word should be acknowledged. This is why we let *xx:xxish* represent all the words that follow this pattern.

[32] Attested in the BNC are: *9ish*, *11ish*, *13ish*, *20ish*, *30ish*, *40ish*, *50ish*, *60ish*, and *90ish*.

Haspelmath, 1996: 49–50; Plag, 1999: 113 fn 2; 2003: 97), since its appearance is syntactically triggered and obligatory (Plag, 2003: 97).

The adjective suffix *-ly*, as in *friendly* and *fatherly*, is excluded from our analysis. Words derived from an ordinal number, as in *firstly*, *secondly*, or *ninthly*, are transformed into *xxthly*.

While the *P* index for *-ly* is not available in Baayen and Lieber (1991), Baayen and Renouf (1996: 86) report finding a large number of hapaxes for *-ly* (1278, cf. 739 for *-ness*) in a corpus of 80 million words.

Non-Suffix Control *ch#*. The non-suffix *ch#* is merely a word ending of a noun, as in *watch* and *church*. This control case will provide a baseline condition for an affix to be considered productive (or not). The productivity index for *ch#* could arise from, for example, compounding or coinages of simplex nouns. We also expect some sources of noise, such as the occurrence of rare or obsolete words, to be represented in the productivity index for *ch#*.[33]

Baayen and Lieber (1991: 815) use simplex words for each lexical category to generate a baseline condition against which values of their productivity index could be interpreted. Coinages of simplex words are expected to be limited in productivity as compared with affixational processes.

---

[33] For their measure *P*, Baayen and Lieber (1991) predict that hapaxes will be unlikely to be obsolete words in a large corpus because even obsolete words are likely to be attested more than once. In the P_{DE} measure, due to its mechanism of identifying new words, we do not rule out the possibility that obsolete words may be captured as potentially new words.

In the present study, we focus on nouns ending in *ch#* because the number of word types ending in *ch#* is expected to fall in a range similar to the number of word types for some of our target suffixes. A substantially larger number of word types resulting from examination of all simplex nouns may complicate the evaluation of the productivity indices. The control *ch#* may also make a better case of unproductivity than *-th* because the total number of word types with *-th* is expected to be low.

4.2 Data in the British National Corpus

The performance of the P_{DE} measure will be tested based on the BNC, a corpus of 100 million words. The BNC represents modern British English, and draws its texts from a variety of sources covering different genres and registers (both written and spoken); it stands as one of the largest and most widely accepted corpora available for computational research. Below, we will review how data are organized in the BNC.

At the topmost level, the BNC is divided into 5 text-type categories, as shown in Table 4-2.

Table 4-2. Text-type categories of the BNC (Burnard, 2000).

Text Type	Words	Documents
Spoken Demographic[a]	4,206,058	153
Spoken Context-Governed[b]	6,135,671	757
Written Books and Periodicals	78,580,018	2,688
Written-to-be-Spoken	1,324,480	35
Written Miscellaneous	7,373,707	421

Notes: [a] This mostly represents spontaneous conversations. [b] This mostly represents interviews, speeches, and meetings.

The majority of texts fall in the "Written Books and Periodicals" component (henceforth, *WBP*), and we will be making an exclusive use of this component in the present study (except in Section 5.1.5). What the third column of Table 4-2 refers to is the number of contributing documents (files on a CD-ROM), each of which contains texts drawn from a single unique source. The size of such documents varies: the target sample size is 40,000 words for books, and no document exceeds 45,000 words (Burnard, 2000). The reference manual of the BNC (Burnard, 2000: 8–9) explains that:

> Text samples normally consist of a continuous stretch of discourse from within the whole. A convenient breakpoint (e.g., the end of a section or chapter) was chosen as far as possible to begin and end the sample so that high-level discourse units were not fragmented. Only one sample was taken from any one text.

In implementing the P_{DE} measure, we consider each of the 4,054 documents as sampling words used by one speaker (author) or a set of co-speakers (co-authors). A subset of these documents (e.g., 2,688 documents of WBP) will be randomly distributed to form the segments on which the P_{DE} and UC measures depend. As it is documents that will be randomly distributed into segments, words within any document will never be dispersed across segments. This point will be discussed in more detail in the next section.

The texts of the BNC are parsed by a *word-class* (i.e., part-of-speech) tagger, *CLAWS*.[34] A word-class tag assigned to each word by CLAWS identifies its grammatical features. Word-class tags are of great advantage in collecting relevant words: we will examine only nouns, adjectives, verbs and adverbs, distinguishing these by the tags. The tags are also useful in distinguishing a noun and a verb that share the same orthographical

[34] For more details about the BNC and CLAWS, see Burnard (2000), Garside, Leech, and McEnery (1997), Leech and Smith (2000).

form; for example, a word like *audition* can be used as a noun or as a verb, and for *-ion*, we will restrict attention to the noun usages of *audition.*

Of the 57 word-class tags (plus 4 punctuation tags) assigned by CLAWS, those of interest to us are listed in Table 4-3.

Table 4-3. Word-class tags relevant to the current study.

Tag	Feature Description	Category	Suffix
AJ0	Adjective (general or positive)	Adjective	*-ish*, *-ous*
AJ0-NN1	Ambiguous, more likely AJ0		
NN0	Common noun, neutral for number	Noun	*-ness*, *-ity* *-er*, *-ee* *-ion*, *-ment* *-th*, *ch#*
NN1	Singular common noun		
NN2	Plural common noun		
NN1-AJ0	Ambiguous, more likely NN1		
VVB	Finite base form of lexical verbs	Verb	*-ize*, *-ify*
VVD	Past tense form of lexical verbs		
VVG	The *-ing* form of lexical verbs		
VVI	Infinite form of lexical verbs		
VVN	Past participle form of lexical verbs		
VVZ	The *-s* form of lexical verbs		
AV0	General adverb	Adverb	*-ly*

A combination of two tags, as in AJ0-NN1 and NN1-AJ0, is used for an ambiguous word to which CLAWS could not assign the word-class tag with certainty. The first of the pair of tags is the one that CLAWS considers to be more likely. We will adopt the practice of taking the more likely tag as definitive and therefore treat AJ0-NN1 as an instance of AJ0, and NN1-AJ0 as an instance of NN1.

Plag (1999: 108–109), in his experiment with corpus data, points out that words ending in *-izing* and *-ifying* can sometimes be used (and identified by a tagger) as a noun or adjective. Such cases do exist in the BNC (Leech & Smith, 2000). However, we will not expand our analysis of *-ize* and *-ify* to include such nominal and adjectival uses, and instead examine only verb uses of *-ize* and *-ify*, specifically.

There are two major reasons that the BNC is well-suited for the present research. First, the BNC makes available a large amount of data that have been carefully processed (e.g., word-class tagged),[35] and testing the UC measure with many corpus segments requires a large number of words in total.[36] Second, the BNC offers document boundaries in the form of files on a CD-ROM, and this organization of the corpus facilitates corpus-segment creation.

4.3 Word-Type Database

In order to assess token and type frequencies in the BNC, we must define relevant word types for our 12 suffixes and the non-suffix control. A database of word types for the suffixes was built for this purpose. The word-type database has three major components: (i) a list of relevant words, (ii) a list of irrelevant words, and (iii) a list of modification rules. Modification rules, as will be discussed shortly, deal with any

[35] The tagging error-rate for the entire BNC has been estimated to fall in the range 0.71% to 1.15% (Burnard, 2000).

[36] Word frequencies for the entire BNC are made available elsewhere (e.g., Leech, Rayson, & Wilson, 2001), but these are not suited to testing the currently proposed measures, in which token and type frequencies must be assessed for data appropriately organized into segments.

necessary transformation of a word (e.g., correction of misspelled words) in order to obtain a form which can be classified as relevant or irrelevant. In this section, we will discuss the processes involved in building the word-type database. At various stages of the processes, we will make use of the OED and WD.

Following the conventions of the BNC, we consider a *word token* to refer to any *word* within the pattern "<w TAG>*word*" in the texts of the BNC, where each *word* is associated with a word-class tag, <w TAG>, assigned by CLAWS.[37] Reference to number of tokens (e.g., in creating segments of a specified size) is based on the count of word tokens defined in this way. Punctuation marks that appear as *punct* within the pattern "<c TAG>*punct*" are excluded from any word count.

Another important convention adopted is that only one token for a suffix will be identified in any given word token. For example, the word *institutionalization* will be counted only as a token for *-ion*, and not as a token with any embedded suffixes such as *-ize*.[38] Since the present research concerns itself only with suffixes, we will essentially be examining the word ending of each token, identifying instances via their outermost

---

[37] Special handling is required for those forms appearing in the BNC as "multiwords"; the BNC presents such forms as *multiword* within the pattern "<w TAG>*multiword.*" These refer to the occurrence of a space or a slash within the word, for example, *of course*, or *electrical/electronic* (Burnard, 2000). A multiword is considered to be one token by CLAWS, and receives one word-class tag. For the purposes of the present study, we divide a multiword into separate tokens, assigning the original word-class tag to each part; for example, <AJ0> *electrical/electronic* will be divided to give separate tokens <AJ0> *electrical* and <AJ0> *electronic*. This convention attempts to maximize our capture of relevant words; it adds to the token pool the otherwise overlooked non-final elements of any multiword.

[38] Plag (1999: 108) notes that acknowledging only the rightmost affix in each token may weaken the psycholinguistic validity of a corpus-based productivity measure. At the same time, acknowledging every suffix in a multiply affixed word adds considerably to the technical complexity of processing corpus data.

derivational suffix. Note that this convention means that no token is ever treated as instantiating more than one suffix, since there is no orthographic ambiguity within the set of target suffixes.

The search for relevant words is first narrowed down by focusing only on nouns, adjectives, verbs, and adverbs, using the word-class tags shown earlier in Table 4-3. Words with any other word-class tags are entirely ignored. The search for relevant words is further narrowed down by finding words whose inflectional ending matches any of the normalization patterns listed in Table 4-4.[39]

Table 4-4. Normalization of inflectional and spelling variants.

Before	After
-ness, -nesses	*-ness*
-ity, -ities	*-ity*
-er, -ers	*-er*
-ee, -ees	*-ee*
-ion, -ions	*-ion*
-ment, -ments	*-ment*
-th, -ths	*-th*
-ize, -izes, -ized, -izing *-ise, -ises, -ised, -ising*	*-ize*
-ify, -ifies, -ified, -ifying	*-ify*
-ish	*-ish*
-ous	*-ous*
-ly	*-ly*
ch, ches	*ch*

[39] The possessive *'s* as in *government's* is treated as a separate token in the BNC, as in <NN1> *government* <POS> *'s*.

As shown in Table 4-4, the British form *-ise* is transformed into the American form *-ize*.[40] For consistency, British/American form variations will be eliminated by transforming the British form into the American form.[41] For example, British forms, such as *colourfulness* and *sombreness*, are transformed into their American equivalents, *colorfulness* and *somberness*, respectively. WD is particularly useful in dealing with form variations of this kind. When there are spelling alternants available (e.g., *cancellation ~ cancelation*), WD specifies which spelling is the main entry. Since we should not allow mere spelling variation to inflate the number of word types with a suffix, form variations are eliminated by transforming any variant form to the form that constitutes WD's main entry.

At the stage of normalizing inflectional endings, all letters of a word are lowercased, and any special characters (e.g., accented characters) are replaced with ASCII-equivalent characters. If a suffix is hyphenated (as in *me-ness*), the hyphen is deleted, but hyphens not adjacent to a suffix are preserved (for use in the procedures discussed in Section 4.4).

---

[40] While the automatic transformation of *-ise* into *-ize* successfully transforms some words (e.g., *computerise → computerize*), it also leads to inappropriate overcorrection (e.g., *promise → promize*; *franchise → franchize*). Some modification rules are defined for these cases so that the overcorrected forms are maintained in (transformed back to) their appropriate *-ise* forms (e.g., *promize → promise*; *franchize → franchise*); such words with *-ize* are then excluded from analyses referring to suffix *-ize*. The serial transformations involved in the overcorrected cases may appear to be inefficient, but as a practical matter, it turns out to be more efficient than *selectively* transforming *-ise* forms into *-ize*; cases like *promise* and *franchise* are few in number.

[41] The choice of American forms over British forms reflects no more than the fact that the present research is conducted in the United States.

Aside from eliminating form variations, by far the greatest majority of modification rules are needed for correcting spelling errors; when the corpus is large, the amount of spelling errors is also large. Spelling errors in a corpus are common but must not be overlooked because each error typically occurs only once and greatly distorts the number of hapaxes (Evert & Lüdeling, 2001). In the spoken component of the BNC, many idiosyncratically transcribed words also need correction (e.g., *dona-a-a-ation* → *donation*). As for spelling-error corrections, omitted portions of taboo words also need to be spelled out.

Each word to which one of the normalization patterns in Table 4-4 was applicable has been scrutinized to determine whether it contains one of our target suffixes. The majority of words were automatically sorted into a group of likely relevant words and a group of likely irrelevant words based on various types of information (described below), but to minimize errors at any stage, no word was exempt from repeated item-by-item inspection. To facilitate an item-by-item inspection that can be extremely laborious, each word is supplemented with a wealth of information. For each word entry in the word-type database, various types of information as listed in (1), were automatically collected. Each of these elements and the use to which it is put is described below, in turn.

(1) a. Token frequency

b. OED/WD labels

c. Word definitions

d. Probable bases and base definitions

e. Contexts (10 at maximum)

f. Spelling suggestions

Token frequency. This is simply how many times the word appears in the entire BNC.

OED/WD labels. The word is checked against word lists created from the OED and WD,[42] and the word receives the label "OED" and/or "WD" to signal that it is present in those dictionaries. We also use the eOED/eWD word lists in which listed words are only those for which relevant etymologies are available in the OED/WD. If an etymology is available for the target word, the labels OED and/or WD will be replaced with eOED and/or eWD. Any word receiving either eOED or eWD is considered to be a relevant word. Thus, according to this procedure, *business* is accepted as a word type suffixed with *-ness*. A simple rule that we follow is that a given word will count as an instance of some suffix if there is a relevant etymology in either the OED or WD.

Word definitions. Word definitions, if any, are gathered from two electronic dictionary resources: the 1913 edition of Webster's Third New International Dictionary (Project Gutenberg, 1999), which we will refer to as WD†, and the WordNet lexical database (Princeton University, 2003, Version 1.7.1). These dictionary resources together offer word definitions for 198,475 words (a combined count excluding any overlap).

[42] The OED/WD lists are created by searching electronic versions of the OED and WD for words ending with the target suffixes.

Probable bases and base definitions. The word is put through a series of tests that modify the word ending to identify a probable base of the word and check whether such a base is listed in WordNet and/or WD†.[43] As a result, a list of probable bases and their word definitions, if any, will be available for the word.

Contexts (10 at maximum). Contextual information plays an important role in identifying the word, especially in the case of rare or low frequency words. A context consists of 31 words or punctuation elements: the target word itself, and 15 additional words (or punctuation elements) preceding and following that target. An example of such a context is shown below, with its accompanying (tag-free) translation:

> of {PRF} the {AT0} CTP {NP0}) {PUR} , {PUN} is {VBZ} that {CJT} it {PNP} can {VM0} not {XX0} explain {VVI} the {AT0} intentionality {NN1} or {CJC} &bquo; {PUQ} ***aboutness*** {NN1} &equo; {PUQ} that {CJT} connects {VV2} the {AT0} neural {AJ0-NN1} events {NN2} with {PRP} the {AT0} object {NN1} they {PNP} are {VBB} supposed {AJ0} to {T00} be {VB1} perceptions {NN2}
>
> ...of the CTP), is that it can not explain the intentionality or "aboutness" that connects the neural events with the object they are supposed to be perceptions...

Each word in the context is followed by its word-class tag, and the target word is enclosed in asterisks for emphasis. The 31-element window is in almost all instances sufficient to understand how the target word has been used. Examining a maximum of 10 contexts generally proves sufficient, because those words for which contextual information is needed are typically words of low frequency.

---

[43] These are not checked against the OED/WD word lists because the OED/WD word lists are for words ending with the target suffixes only.

Spelling suggestions. If the word received neither an OED nor WD label, suggested spelling corrections are obtained using the DICT server/client (a dictionary query/response system, Dict.org, 2003) on a Linux machine.

Although the majority of words are easily identified with the types of information discussed above, the inspection of some words is also augmented by searches on the Web (Google search engine), and additional etymologies from Ayto (1990) and Chantrell (2002).

The word-type database was compiled to contain 17,347 word types in total, for the 12 suffixes and the non-suffix control. There are 6,797 modification rules (mostly automatically generated, but some inevitably defined manually for cases like *dona-a-a-ation* → *donation*). The number of modification rules is quite large, but it must be noted that some of them are needed to obtain correct forms for irrelevant words (so that they can appropriately be deemed irrelevant), and that a given word can be misspelled in quite a number of ways. The next section will discuss the major theoretical and technical difficulties encountered in obtaining the total number of word types.

4.4 Difficulties in Defining Word Types

Prefixation and compounding are two major obstacles in defining distinct word types with any suffix, since the number of word types with a given suffix depends crucially on how these other word formation processes are handled. In this section, we discuss our treatment of prefixation and compounding. What needs to be kept in mind as we proceed is that given the enormous size of the corpus to be processed, any procedure

that is adopted must be one that is principled, and at the same time one that can be largely automated.

4.4.1 Prefixation

Although the productivity of English suffixes is the focus of the present study, prefixes in the data cannot be simply ignored. The kind of question that typically needs to be answered is the following: Are *familiarity* and *unfamiliarity* to count as two distinct word types with *-ity* or variant forms of one *-ity* type? This issue is often encountered in the literature (e.g., Bauer, 2001; Plag, 1999, 2003: 75–78; Plag, Dalton-Puffer, & Baayen, 1999), and decisions are inevitably difficult to make. For example, in testing the measure *P*, Bauer (2001: 151–152) notes that the number of *-ment* hapaxes in his data set differs by 12, depending on whether the prefix is removed from words such as *under-development*. He therefore presents two results: one that removes the prefix ($P = 0.0083$) and the other that does not ($P = 0.011$). He notes that the two *P* indices lead to slightly different impressions about the productivity of *-ment*. In testing his own measure, Bauer (2001: 196) excludes words such as *defamiliarisation* and *dehumanisation* from the pool of potential neologisms, noting that they are simply prefixed versions of their counterparts (*familiarisation* and *humanisation*), which are not innovative and are listed in the *Concise Oxford Dictionary*. Bauer suggests that the prefixed word forms should be considered in discussing the productivity of prefixes but not the productivity of suffixes. What concerns us here is the potential arbitrariness in the treatment of prefixes. Unless the treatment of prefixes is uniform and principled, it is difficult to evaluate findings for a productivity measure.

At the outset, we take note of the fact that prefixes could dramatically increase the number of word types with a suffix. Some example prefixes are:

(2) *a-*, *anti-*, *be-*, *co-*, *counter-*, *de-*, *dis-*, *en-*, *ex-*, *in-*, *mis-*, *non-*, *pre-*, *re-*, *sub-*, *un-*

A given word may be compatible with more than one prefix; for example, given *humanize*, the OED has entries for *unhumanize*, *dehumanize*, and *dishumanize*. It is difficult to estimate (in the abstract) how prefixes affect the number of words because prefixes also differ in productivity; for example, *non-* and *un-* are more productive than *dis-* and *in-* (Matthews, 1991: 71–74).

The so-called neoclassical elements (Bauer, 1998a; Plag, 2003: 135, 155–159), such as *bio-*, *photo-*, *psycho-* or *neuro-*, are not considered prefixes, but are included among the compounds to be discussed in the next section. Distinguishing neoclassical items from prefixes can sometimes be difficult. For consistency, we adopt a distinction based on the treatment of such elements in WD, where neoclassical items are identified as *combining forms*.

Our goal here is to devise a principled treatment of prefixes. We consider five options, discussing each of these in detail:

(3) a. Maintain all prefixes.

b. Remove all prefixes.

c. Remove prefixes appearing after the relevant suffixation.

d. Remove all hyphenated prefixes.

e. Remove a prefix based on OED/WD criteria.

Maintain all prefixes. An immediate problem with this option is that each and every form variation created by prefixation adds a distinct word type with a suffix. Consider, for example, the word *anti-institution*, occurring just once in the BNC. A possible analysis of this word would be [*anti-* [*institution*]] or [[*anti-institute*] *-ion*]. To the extent that the first analysis is preferred, it seems altogether unsatisfactory to allow the prefix *anti-* in this particular case to increase the number of word types with *-ion*; that is, the form variation between *anti-institution* and *institution* should not affect the counts through which the productivity index for *-ion* is calculated. It is clear from this example alone that, rather than allowing an uncontrolled increase in word-type counts, what is necessary is to restrict the treatment of prefixation in some way.

Remove all prefixes. This option, opposite to the one just considered, has a simple appeal, but engenders its own problems. Treating *anti-institution* as *institution* may be acceptable as a way of blocking unwarranted inflation of word-type counts. However, in some cases, the removal of a prefix does not seem justifiable. Let us consider *disagreement*. We feel somewhat uncomfortable in treating *disagreement* as an instance of *agreement* (removing *dis-*) because the more natural analysis of this word would be [[*disagree*] *-ment*], rather than [*dis-* [*agreement*]]. Moreover, the removal of a prefix is clearly unacceptable in cases such as *encouragement* → **couragement*.

Considering these first two options, together, we conclude that some cases of prefixation seem to be more relevant to the productivity of a suffix than others. If we are

to impose restrictions on accepted prefixations, we may need to identify and selectively block prefixations that are irrelevant to the productivity of a suffix.

Remove prefixes appearing after the relevant suffixation. We saw in *anti-institution* that the prefixation of *anti-* likely occurs after the suffixation of *-ion*, while in *disagreement*, the prefixation of *dis-* likely precedes the suffixation of *-ment*. Since our focus is on the word formation process for a suffix, we may wish to allow a prefix only if the prefix is already present as part of the base when the relevant suffixation occurs. We will then need to order affixes by the order of their occurrence. A possible solution for this situation may be sought in the *Level Ordering Hypothesis* (e.g., Kiparsky, 1982; Siegel, 1979, among others), which provides a phonologically motivated order in which affixes attach to a base (or a root). Affixes that cause a phonological change (e.g., a stress shift) in the base that they attach to have been referred to as *Level I affixes* (with *morpheme boundary* +, e.g., *+ity*, *+ion*, *+ous*), whereas affixes that do not cause such a phonological change have been referred to as *Level II affixes* (with *word boundary* #, e.g., *#ness*, *#er*, *#ment*, *#ize*, *#ly*). Level I affixes are considered to precede Level II affixes when they attach to a base. Ordering affixes appears to enable us to ignore any prefixation that occurs after the relevant suffixation.

Unfortunately, however, Level Ordering does not offer us a straightforward solution to our problem due to an issue known as the bracketing paradox (see, e.g., Spencer, 1988; Sproat, 1988). A well-known example is seen in the word *ungrammaticality*. There are two possible ways of bracketing this word to specify its internal structure:

(4) a. [$_N$ un- [$_N$ grammaticality]]

b. [$_N$ [$_A$ ungrammatical] -ity]

Analysis (4a) follows the principles of Level Ordering where the Level I affix *-ity* attaches to the base *grammatical* before the Level II affix *un#* attaches to *grammaticality*. However, this ordering of the affixes runs against the observation that *un-* does not normally attach to a noun (Spencer, 1991: 180). Semantically, *ungrammaticality* should be a nominalization of *ungrammatical*, as in (4b).

Whether we follow Level Ordering or semantic plausibility, there are more problematic cases for deciding whether to remove a prefix. Consider *nonrandomness*, which can be analyzed as:

(5) a. [$_N$ non- [$_N$ randomness]]

b. [$_N$ [$_A$ nonrandom] -ness]

Each of *non-* and *-ness* is a Level II affix, and these two analyses of *nonrandomness* appear to be equally semantically acceptable. We cannot decide whether or not to remove *non-* in this example.

Without immersing ourselves in deep theoretical issues (and inevitable case-by-case argumentation), we conclude that selective removal of prefixes based on their order of occurrence with respect to suffixation is by no means a straightforward matter.

Remove all hyphenated prefixes. A hypothesis we might formulate given the comparison of *anti-institution* and *disagreement* is that the appearance of a hyphen in *anti-institution* somehow seems to signal a prefix that can be removed. However, any such hypothesis is difficult to evaluate given a certain arbitrariness about which words are hyphenated in English (Bauer, 1988: 101; Plag, 2003: 5). For example, the difference between *demystify* and *de-mystify* could be a matter of style or of dialect. We note, nevertheless, that there seems to be a general tendency for novel cases of prefixation to be hyphenated, as seen in the following examples attested in the BNC (none of which is listed in the OED or WD):

(6) *dis-beautify*, *counter-regulation*, *de-contextualize*, *co-movement*

There are, however, some problems with claiming that all hyphenated prefixes are novel cases of prefixation. A prefix may be hyphenated customarily, as perhaps in *co-operation*. Another problem is that, in the context of corpus data, any writer may have arbitrarily (idiosyncratically) decided to spell a word with a hyphen, and this will not give us any principled treatment of prefixes. The option of removing hyphenated prefixes needs further refinement.

Remove a prefix based on OED/WD criteria. Reference to the OED and WD[44] can bring a certain kind of principled distinction in our prefix treatment. With an assumption that the OED and WD are conservative in accepting and listing new word forms, we can take advantage of that conservatism and regard a prefixed form not listed in the OED or WD as a novel case of prefixation, and hence as a case from which the prefix can safely be removed. Thus, we will consider the word forms in the OED and WD as attested cases of prefixation, and any other cases of prefixation as novel. Since our aim is to restrict form variation due to prefixation, we will also incorporate the preceding option by removing the prefix from a word if the prefixed form cannot be spelled without a hyphen in either the OED or WD. The consequence of this treatment is that *anti-institution* will be *institution*, while *disagreement* will remain as *disagreement*.

The procedure involved is as follows. Given an inherent arbitrariness in whether a word is hyphenated by individual authors, a prefixed word may appear in the BNC in either hyphenated or unhyphenated form. We first eliminate this form variation by transforming all hyphenated forms into unhyphenated forms. Then, the word is checked against the OED and WD word lists to see if it is listed in an unhyphenated form. If either (or both) of the OED or WD allows the unhyphenated spelling, the prefix is accepted. Otherwise, the prefix is removed, and the word is normalized in its unprefixed form.[45]

---

[44] For this purpose, the OED refers to a combination of the *Oxford English Dictionary* and the *Shorter Oxford English Dictionary*. There are slight differences between these two dictionaries in whether a given word is spelled with a hyphen. The purpose of combining the two dictionaries is to encompass more variations in how words are spelled. WD, as before, refers to *Webster's Third New International Dictionary*.

[45] Ideally, we should also take one more step by checking whether the word minus the suffix (i.e., the prefixed base) is listed in the OED or WD in an unhyphenated form, but

We find this to be the most effective and principled method of preventing novel cases of prefixation from inappropriately increasing the count of word types with a suffix. Its additional advantage is that the procedure can be automated.

4.4.2 Compounding

Compounds can be spelled in three different ways, *girl friend*, *girl-friend*, and *girlfriend* (Bauer, 1988: 101; Matthews, 1991: 94), variation being apparently random (Bauer, 1998b: 69). If a compound is spelled with a character space, as in *girl friend*, its components are treated as two individual tokens in the BNC (Leech & Smith, 2000). The fact that these two tokens constitute a compound is ignored in the present study because the procedures identify only one token at a time. We will, therefore, be concerned with compounds spelled either as *girl-friend* or as *girlfriend*. The type that concern us most are *synthetic compounds*, where the head element has a verb as (part of) its base and the modifier is an element that could function as an argument of that verb in a sentence (Bauer, 1988: 36). Some examples are *bookkeeper*, *dishwasher*, and *taxi-driver*. Although the discussion that follows focuses on synthetic compounds (seen primarily among *-er* words in our data) the rule we arrive at will be applicable to all types of compounds. Compounding poses a kind of question similar to that posed by prefixation: Are *bookkeeper*, *dishwasher*, and *taxi-driver* to be considered word types with *-er* distinct from *keeper*, *washer*, and *driver*? In light of the discussion of the treatment of prefixes, we will consider three options as listed in (7), examining each in turn below:

---

this is not currently possible because our OED/WD word lists only contain words that end with the study's target suffixes.

(7) a. Maintain all compounds.

b. Remove all modifying elements.

c. Remove a modifying element based on OED/WD criteria.

Maintain all compounds. This option would have a larger impact on the number of form variations than its equivalent for prefixes because the range of elements that constitute a modifying element in a compound is presumably much wider than the range of prefixes. By a *modifying element*, we refer to the *book* of *bookkeeper*, the *dish* of *dishwasher*, the *taxi-* of *taxi-driver*, or more generally, any initial element in a compound. The consequence of maintaining all compounds is that words such as *agreement-seeker* or *attention-getter* will be treated as distinct word types (different from *seeker* and *getter*) in analyzing the productivity of *-er*. There are certainly grounds for concern about maintaining any compound (grammatical or ungrammatical, acceptable or unacceptable) that appears in corpus data. As was the case with prefixation, a better approach would be to impose some restriction on accepted compounds.

Remove all modifying elements. A problem with this option is that in many cases it is not clear whether the removal of a modifying element can be justified. For example, removing *auction-* from *auction-goer* may be acceptable, but removing *baby* from *babysitter* seems doubtful.

As our first attempt at refining our definition of a modifying element, we may examine whether the base is listed in a dictionary. In the case of babysitter, for example, removing *baby* from *babysitter* may be objected to because the verb *babysit* (a back-

formation from *babysitting*) is listed in a dictionary. That is, *baby* may not strictly be a modifying element of a compound, [[*baby*][*sitter*]], because it is already part of the compound verb *babysit*. However, this means of determining what properly constitutes a modifying element in the nominal compound leads to seemingly inconsistent treatments, as shown in Table 4-5.

Table 4-5. Variable OED/WD listing of compound forms.

Example	Change	OED/WD Listings		
		Verb	Verb + *-er*	Verb + *-ing*
auction-goer	(*auction-*) + *goer*	**auction-go*	**auction-goer*	**auction-going*
churchgoer	(*church*) + *goer*	**churchgo*	*churchgoer*	*churchgoing*
childbearer	(*child*) + *bearer*	**childbear*	**childbearer*	*childbearing*
babysitter	*babysitter*	*babysit*	*babysitter*	*babysitting*

All word forms in the first column of Table 4-5 are attested in the BNC, and the parenthesized elements in the second column are modifying elements, perhaps subject to removal. The remaining column shows related word forms listed in either or both of the OED and WD. Note that the modifying element would be subject to removal in the first three cases, because no (bare) verb base is attested that contains the modifying element. On the other hand, *babysitter* would remain as it is, unchanged, because of the existence of the verb base *babysit*. However, treating *churchgoer* and *babysitter* differently seems unsatisfactory for two reasons. First, the fact that *babysitter* has a compound verb base which is the result of back-formation does not seem to have any relation to the productivity of *-er*. Second, the above rule would reject *churchgoer* even though *churchgoer* itself is a word listed in the OED/WD.

Remove a modifying element based on OED/WD criteria. Just as was done in our treatment of prefixes (see Section 4.4.1), we can rely on the OED/WD to obtain a consistent method of identifying novel cases of compounding. A modifying element will be removed from a compound whenever the compound is not listed in at least one of the OED and WD in an unhyphenated form. The procedure is implemented as follows. First, all hyphenated compounds will be transformed into unhyphenated form to eliminate any idiosyncratic variation (e.g., *auction-goer* → *auctiongoer*). Then, each compound will be checked against the OED/WD word lists. If either the OED or WD lists the compound in an unhyphenated form, the compound is accepted. If it is not, we obtain (working leftwards through the word) the longest unhyphenated constituent that is allowed in the OED or WD. Here, analogy from attested *-ing* forms enters into determining what constituent is acceptable.

Among the compounds listed in Table 4-5, *churchgoer* and *babysitter* undergo no modification (and are accepted as distinct *-er* word types), because they are listed in the OED/WD. In addition, *childbearer* is also accepted because, even though the word itself is not specifically listed in the OED/WD, the word *childbearing* is listed. By analogy with that form, we accept *childbearer*. By contrast, *auction* will be removed from *auctiongoer* because no entry in the OED/WD suggests that *auctiongoer* is acceptable.

Before closing the discussion of the treatment of compounds, we must ask what status *goer* in Table 4-5 (resulting from *auctiongoer*) has, among the word types with *-er* (see Fabb, 1998: 69, for remarks on this point). One possibility is that *goer* should be regarded as representing all novel cases following the *X-goer* pattern. The same consideration applies to *mindedness*, resulting from word forms such as *fair-mindedness*,

broad-mindedness, and *light-mindedness*. In this case, it may initially appear implausible to uniformly reduce these forms to *mindedness*. However, more frequent words such as *absentmindedness* are indeed listed in the OED/WD in unhyphenated form,[46] so our proposed compound treatment (seeking the largest unhyphenated constituent, on a case-by-case basis) is effective in preventing instances of *X-mindedness* from gratuitously inflating the count of *-ness* word types, but is tempered rather than monolithic in its application. The same point applies to phrasal cases (treated as compounds in the present study), such as *other-worldish* and *end-of-the-worldish*; these are reduced to *worldish*.

Bringing together the foregoing discussion of prefix and compound treatments, we note that the procedures decided upon effectively limit the extent to which novel cases of prefixation and compounding increase the number of word types with a suffix in an uncontrolled way. However, the procedures are certainly not perfect, particularly for compounds, because of idiosyncrasies in the OED/WD word lists: clear cases of compounds are accepted whenever they are listed in the OED/WD. In a theoretical discussion of the productivity of a suffix, we can choose to discuss only pure cases of suffixation. However, faced with actual corpus data, it becomes evident that isolating "pure" suffixation is difficult both theoretically and technically, and that difficulty is acute for prefixed forms and compounds. Neither ignoring or accepting all prefixed forms and compounds seems plausible, and a reasonable compromise must be sought. A treatment of prefixed words and compounds that differs from the one adopted here might yield different numbers of word types with the target suffixes. What is most important, of course, is that such a treatment should be principled.

---

[46] WD lists this word as *absentmindedness*, and the OED, as *absent-mindedness*.

4.5 Random Sampling of Documents

In this section, we will discuss how the corpus segments required by the P_{DE} and UC measures should be created from the BNC. Of particular interest is how different methods affect the identification of new words in the P_{DE} measure. All the data in this section are based on WBP as this is the component that we exclusively use in Chapter 5 (except Section 5.1.5).

4.5.1 Randomization by Documents or Words

We proposed in Section 3.7 that a preferred method of corpus-segment creation in the P_{DE} measure (and the UC measure) involves random sampling at the level of documents (henceforth, *RD*), where the documents (rather than the words) making up a corpus are the unit of randomization. The motivation was that considering each document as a set of words used by a speaker (or co-speakers) allows each corpus segment to be viewed as representing the words used by a group of randomly sampled speakers. A non-preferred alternative method of corpus-segment creation involves random sampling at the level of words (henceforth, *RW*), where the words of a corpus are randomized into segments, with document boundaries ignored.

A fundamental difference between RD and RW is that in RD, words used by the same speaker(s) are grouped together as they are sampled into a segment, whereas in RW, words originating with a given speaker are freely dispersed across segments. Recall from Chapter 3 that new words in the P_{DE} measure are those that appear in only one of two segments. For a given word type to be new, then, all its tokens must appear in one

segment only. Supposing that a word might be repeatedly used by the same speaker in a document, we can predict that RW will lower the likelihood for that word to be captured as new, relative to RD.

Whether there is a major advantage in RD over RW clearly rests on how common it is for a new word to be repeatedly used in a document. Church (2000) argues that once a content word (e.g., proper nouns, technical terminology, and good keywords for information retrieval) appears in a given document, it is highly likely that the same word will be repeated in the same document. If the probability of having a given word in a corpus is p, the probability of having the same word occur twice is p^2, if those occurrences are independent. Church finds, however, that once a document has an instance of *Noriega*, say, the probability of finding another instance of *Noriega* in the same document is close to $p/2$, a probability that is much higher than p^2. Although Church's discussion does not focus on words newly coined with an affix, it is worthwhile considering the possibility that Church's finding may apply to such usages, given that many technical terms are coined with suffixes such as *-ity* and *-ize*. Even if occasional clustered usages of new words are seen, it would not necessarily be counterevidence to the claim (e.g., Baayen & Renouf, 1996) that most new words occur only once.[47] It could be the case that *most* new words do occur only once, while *some* occur more than once because of their usefulness to a speaker's current topic. An important question, obviously, is how much is most. If 90% of new words are hapaxes in sampled data, for example, an argument for ignoring the remaining *non-hapaxes* may be acceptable, but the

[47] Baayen and Renouf (1996: 79) acknowledge the possibility that when a new technical term is introduced, it could be used more than once. Their claim, nevertheless, is that most new words are used only once.

situation is quite different if, say, only 60% of new words are hapaxes. We will return to this issue in Section 4.5.3.

Abiding by RD has the advantage that we allow both hapaxes and non-hapaxes (within documents) as candidates for new word status. As a measure not based on token frequency, the P_{DE} measure should not rule out a word as not productively coined simply because the speaker who coined it uses the word repeatedly. If there exists a non-hapax new word in one document of the BNC, we do not wish the tokens of that word to be distributed into two segments, thereby forcing the word to be treated as non-new. RD ensures that all tokens of a word in the same document will be distributed into only one of two segments. In the next section, we will examine differences that arise between RD and RW based on the actual data of the BNC.

4.5.2 New Words and Binomial Probability

Only in the case of RW is V_N expected to be fully predictable based on the distribution of token frequencies and binomial probability. Results of the P_{DE} measure when RW is followed will be obtained in the following manner. Let us assume that the P_{DE} measure has been applied to WBP following RD—as in the results to be reported shortly. We will use the term *whole corpus* to refer to the union of documents that have been distributed into two segments in this experiment with RD. The question to be asked about RW is how results obtained under RD would differ if the whole corpus were redistributed by words. Thus, for each outcome of RD, a corresponding outcome of RW will be calculated.

We will estimate the outcome of RW based on token frequency in the whole corpus and binomial probability, as described below. In the P_{DE} measure, words that are new are those that appear in only one of two segments. For a given word to be captured as new, all tokens of that word in the whole corpus must be distributed to only one of two segments. Given corpus segments *A* and *B*, let W_A denote the event that all *r* tokens of word *w* will be in *A* ($\{w_1...w_r\} \in A$). If any given token is equally likely to be distributed into segment *A* or *B*, the probability of W_A is:

(8) $$P(W_A) = P(w_1 \in A) \times P(w_2 \in A) \times ... \times P(w_r \in A)$$
$$= 0.5 \times 0.5 \times ... \times 0.5$$
$$= 0.5^r$$

Similarly, if we let W_B denote the event that all *r* tokens of word *w* will be in *B* ($\{w_1...w_r\} \in B$), the probability of W_B, $P(W_B)$, is 0.5^r. Since W_A and W_B are mutually exclusive events, the probability of word *w* with *r* tokens in the whole corpus to be new is:

(9) $$P(w\text{: new}) = P(W_A) + P(W_B) = 2(0.5^r)$$

Figure 4-1 plots the probability that *w* will be new as a function of token frequency *r*.

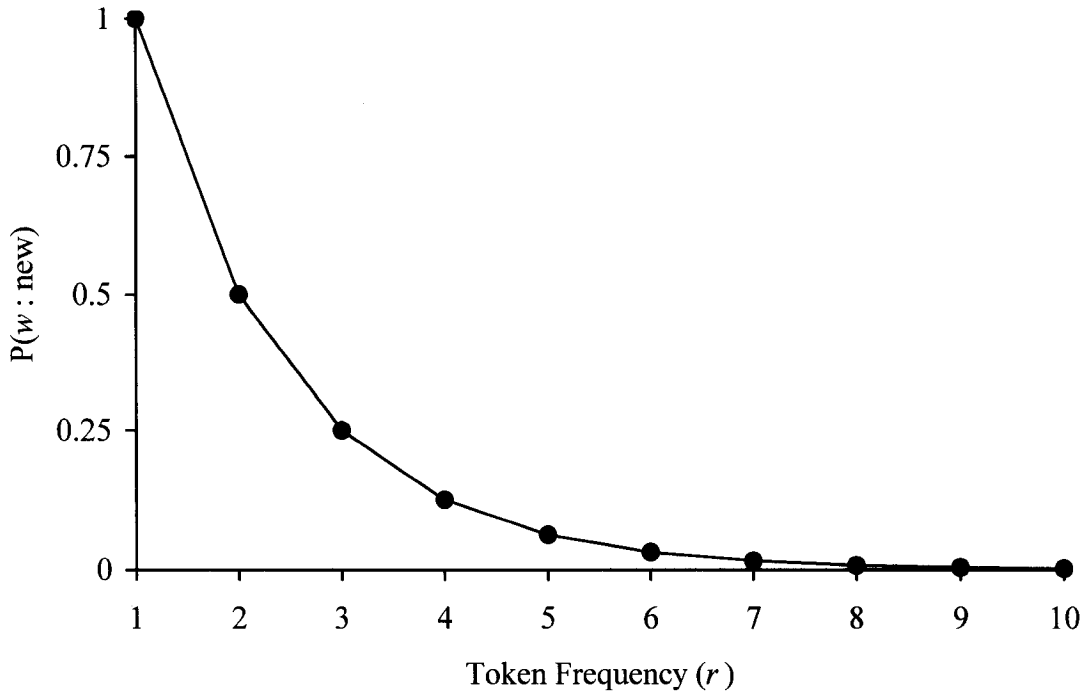

Figure 4-1. Probability that *w* is new, as a function of token frequency.

What is evident in Figure 4-1 is that while every hapax legomenon in the whole corpus is (necessarily) guaranteed to be captured as a new word, the probability that a word will be called new quickly diminishes as its token frequency increases. In fact, any word that occurs more than several times in the whole corpus are very unlikely to be called new. Whether we follow RD or RW, hapaxes will never be shared by two segments; our focus here, however, is on non-hapaxes. RD and RW procedures are predicted to differ with respect to how many non-hapaxes they identify as new words.

Given token frequencies of word types with a particular affix in the whole corpus and the binomial probability function plotted in Figure 4-1, the expected value of V_N under RW can be estimated in the following manner. We note first that the probability in Figure 4-1 is for the two mutually exclusive events W_A and W_B taken together. However,

V_N is a value for one segment, that is, V_N is based on an average of the two events W_A and W_B:

(10) $(P(W_A) + P(W_B))/2 = 2(0.5^r)/2 = 0.5^r$

Suppose that *-ness* has 100 hapaxes in the whole corpus. The probability that these hapaxes enter into the V_N count is $0.5^1 = 0.5$; and $100 \times 0.5 = 50$ hapaxes are expected to be in the V_N for *-ness*. Let us say that *-ness* also has 100 *two-token* words in the whole corpus. The probability that these words enter into the V_N count is $0.5^2 = 0.25$; and $100 \times 0.25 = 25$ two-token words are expected to be in the V_N for *-ness*. If we continue in this manner and sum all the expected values, we can estimate the value of V_N for an affix. An expected value of V_N, $E(V_N)$, for a given affix when following RW is calculated as:

(11) $$E(V_N) = \sum_{r=1}^{n} N_r(0.5^r)$$

Here, N_r is the number of word types with that affix that occur r times in the whole corpus.

To compare V_N and $E(V_N)$, the P_{DE} procedure was applied repeatedly to WBP, with 30 million words in each of two segments (i.e., a total of 60 million words was used in each application).[48] Values for V_N, $E(V_N)$, and the number of hapaxes within V_N were calculated for each of the 12 suffixes and the non-suffix control which are the focus of

[48] Since our focus here is only on new words, under each of RD and RW, we postpone discussion of V and P_{DE} to Chapter 5.

the current study. In each of 100 independent applications, V_N and the number of its hapaxes were obtained under RD; for each, $E(V_N)$ was then calculated using the formula in (11). Table 4-6 presents values for each suffix, averaged over applications.

Table 4-6. Mean values for V_N, $E(V_N)$, and the number of hapaxes in V_N.

Suffix	V_N	$E(V_N)$	Hapaxes in V_N
-ness	431.2	401.4	310.3
-ity	234.4	186.5	140.5
-er	558.6	463.9	346.4
-ee	26.1	18.3	15.2
-ion	348.7	270.6	202.4
-ment	61.6	49.4	36.6
-th	3.5	2.8	2.3
-ize	114.5	101.4	81.5
-ify	21.1	18.5	14.5
-ish	90.6	85.9	72.4
-ous	107.1	89.1	66.7
-ly	754.3	707.0	546.1
ch#	29.7	23.0	15.3

In Table 4-6, we see that each value of $E(V_N)$ is consistently less than the corresponding value of V_N. Figure 4-2 presents a scatter plot comparing the proportion of hapaxes in V_N with $E(V_N)$ as a proportion of V_N.

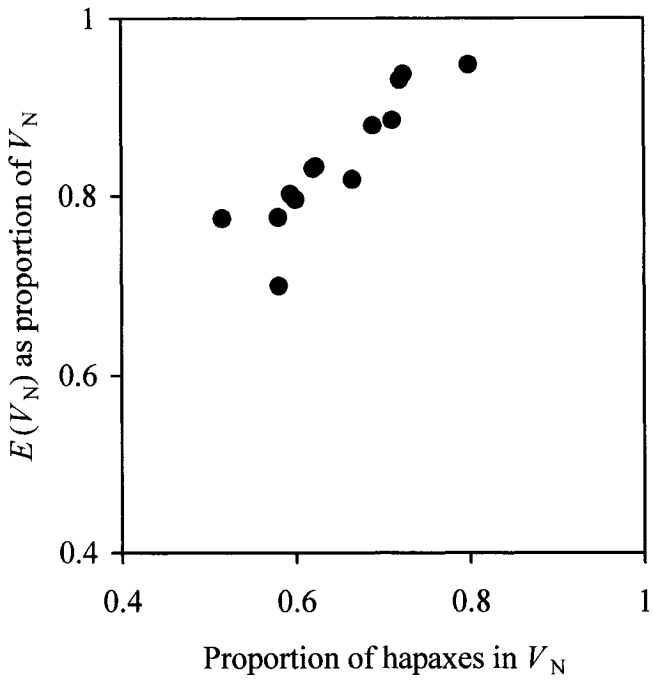

Figure 4-2. Scatter plot relating hapaxes in V_N and $E(V_N) \sim V_N$ estimations.

It is clear that where higher proportion of words identified as new are hapaxes, $E(V_N)$ is a closer estimate of V_N. A Spearman's test indicates that these values are highly correlated, $r_s = .962, p < .01$. To put the matter conversely, we can say that marked differences between V_N and $E(V_N)$ arise to the extent that *non-hapaxes* are identified as new words, only in RD.

4.5.3 Token Frequency of New Words

In this section, we will examine the proportions of hapaxes and non-hapaxes among new words. The comparison of V_N and $E(V_N)$ in the previous section revealed that V_N includes non-hapaxes that cannot be accounted for simply by token frequency and

binomial probability considerations. There are two questions to be asked: (i) To what extent hapaxes are the majority?; and (ii) Is the proportion of hapaxes consistent among suffixes? In discussing token frequency distributions, we use n_r to represent the number of word types that occur r times; for example, n_1 is the number of hapaxes, n_2 is the number of word types that occur twice, and n_{4+} is the number of word types that occur 4 times or more.[49] Figure 4-3 shows what percentage n_r contributes to total V_N count, for each of the token frequency categories.

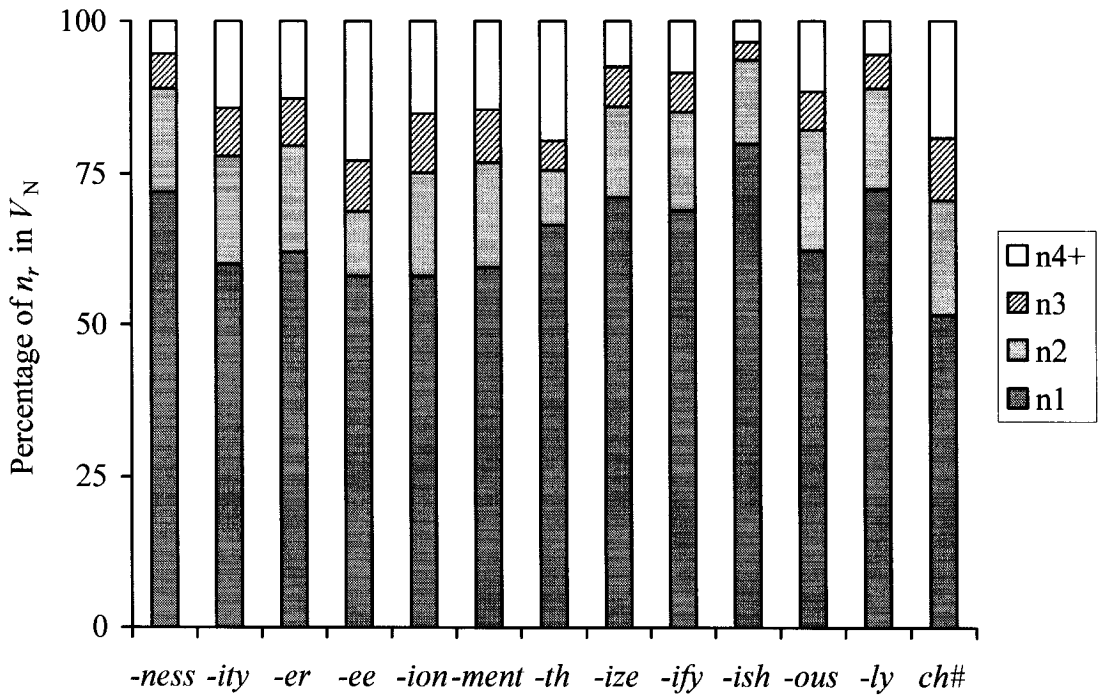

Figure 4-3. Token-frequency distributions within V_N words.

[49] These notations are used in the literature to discuss token frequency distribution in a corpus, but here, we will use them in discussing token frequency distribution specifically within new words.

Focusing first on n_1 (i.e., hapaxes), it is clear that the percentage of n_1 varies from suffix to suffix. Among the sampled suffixes, the highest percentage of n_1 is seen for *-ish* (79.9%) and the lowest percentage for *-ion* (58.0%). For all sampled suffixes taken together, 65.9% of new words are n_1. We may suspect that the percentage of n_1 may correlate with the number of new words, but a Spearman's test shows no significant correlation between V_N and the % of n_1 in V_N, $r_s = .236$, $p > .10$.

Turing to non-hapaxes (n_2 to n_{4+}) in Figure 4-3, we observe that the percentage of each n_r also varies among the suffixes. In particular, the percentage of n_{4+} varies markedly: the highest percentage is seen for *-ee* (22.8%) and the lowest percentage for *-ish* (3.4%).

Based on the findings above, we can summarize the situation as follows. As expected, hapaxes do provide the majority of new words captured. However, they do not constitute the bulk of new words, averaging only 65.9% overall. Moreover, the percentage contribution of hapaxes varies among the suffixes, in a way that has yet to be examined. Any proposal that non-hapaxes are sufficiently marginal in number to be ignored must clearly be abandoned for the currently proposed measure. With the percentage of hapaxes varying among affixes, new words captured by V_N cannot be reduced to words of any specific token frequency. It follows that new words defined by the P_{DE} procedure are fundamentally different from new words defined on the basis of token frequency.

4.5.4 Non-Hapax New Words and Document Frequency

In the preceding section, the percentage of non-hapax new words (34.1% on average) was found to be not so marginal as to be neglected.

What types of words are non-hapaxes in the BNC that are plausibly viewed as new? Let us consider *causee*, a productively formed word type with *-ee* that is unfamiliar to the majority of English speakers (except perhaps syntacticians). In WBP, *causee* appears 9 times in a single document that samples a book titled "The English Infinitive."[50] If we follow RD, all the 9 occurrences of *causee* will be distributed into only one of two segments, and since no other document in the whole corpus offers an instance of this word, *causee* is guaranteed to be called new. The situation is quite different under RW. According to Figure 4-1, occurring 9 times in the whole corpus, *causee* will be called new only with probability 0.002. In effect, we expect *causee* to be non-new under RW.

To examine how many words like *causee* there are in WBP, we examine what will be called the *document frequency* of a word, that is, a count of the number of documents of the whole corpus in which the word appears. Since we take each document to represent a speaker, the document frequency can also be considered to index the number of speakers in the whole corpus who used a given word. When we follow RD, the probability of a word being new as a function of document frequency follows exactly the pattern shown in Figure 4-1: token frequency is simply replaced by document frequency.

---

[50] The following number pairs specify (i) a sequential identifier for the sentence in which *causee* occurs and (ii) the number of times *causee* is used in that sentence: (#750, 1), (#795, 3), (#796, 2), (#802, 1), (#803, 1), and (#805, 1). These data show that the 9 occurrences of *causee* in this document were seen within the context of about 55 sentences, and that 2 sentences used *causee* more than once.

Thus, a word that appears in only one document will always be new, a word that appears in two documents has probability 0.5 of being new, and so on.

Based on document frequency, we can estimate the number of new words for our suffixes in WBP. Among all word types for the 12 target suffixes, there are 3,935 hapaxes in total. Obviously, each of these hapaxes has document frequency 1. Additionally, 658 non-hapaxes each occur in only one document (i.e., document frequency = 1). Moreover, 1,766 non-hapaxes each occur in two documents (i.e., document frequency = 2). Figure 4-1 indicates that the probability that a word with document frequency 2 will be called new is 0.5; we therefore expect $1,766 \times 0.5 = 883$ of these word types to be identified as new, also. If we continue in this manner, there are 5,875 words in total in WBP that will be identified as new, of which 3,935 words are hapaxes and 1,940 words are non-hapaxes, which means that about 33% of new words are non-hapaxes. This finding converges with the average percentage of non-hapax new words for the 12 suffixes we found in the previous section.

Figure 4-4 illustrates distributions of document frequencies (1 through 10) in WBP, for word types with suffixes *-ness* and *-ity* only. For both suffixes, words occurring in a single document are the majority. (Note that these words with document frequency 1 may or may not be hapaxes because we are not concerned with how many times a given word is repeated within a document.) It is interesting to find that the distribution of document frequency resembles a general token frequency distribution in a corpus (see Baayen, 2001): the distribution is characterized by a large number of low frequency items. We also find that *-ness* consistently has more word types with low document frequencies than *-ity*, and based on this fact, we would expect *-ness* to have a larger

number of new word types in general than *-ity*. In fact, values of V_N in Table 4-6 earlier show that more new words are found for *-ness* (431.2) than for *-ity* (234.4).

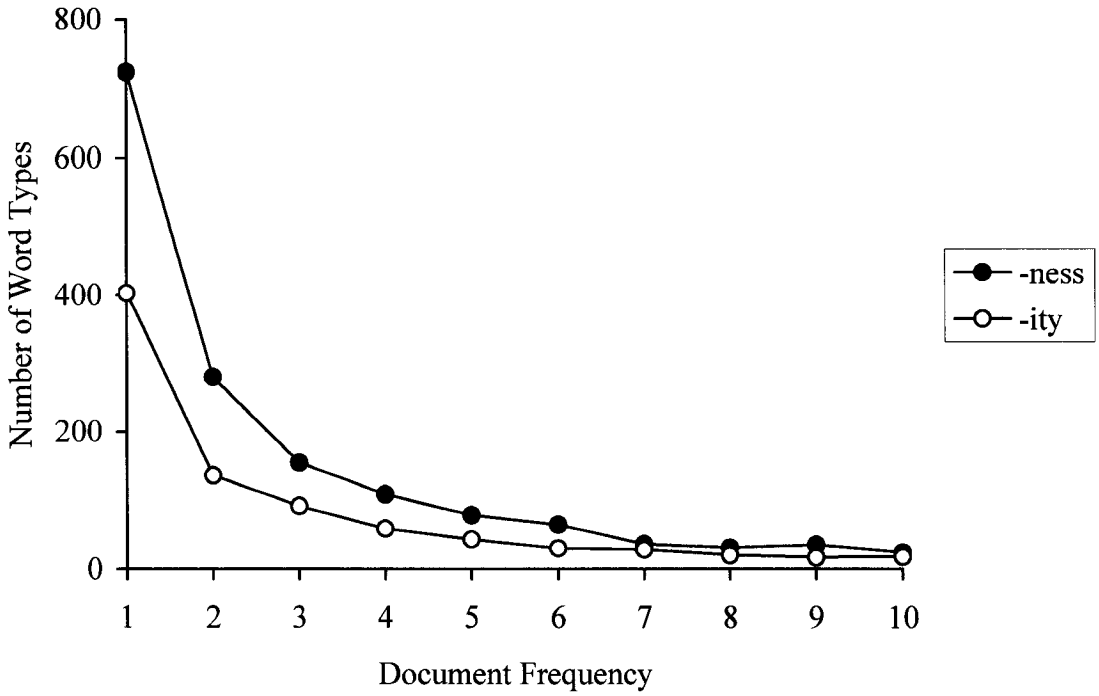

Figure 4-4. Document frequency distributions for word types in *-ness* and *-ity*.

Token frequency and document frequency provide us with different types of information. While token frequency tells us the total number of times a given word was used, document frequency tells us the (estimated) total number of speakers who used that word. To put the matter conversely, token frequency obscures how many speakers in total used that word, while document frequency obscures the total number of times that word was used. In the current approach where we define new words as those used by the least number of speakers, only document frequency gives us the information of interest. Consider a word with token frequency 100. The 100 occurrences of this word could have

been generated by 2 speakers (each speaker repeating the word 50 times) or 20 speakers (each speaker repeating the word 5 times). Only by examining document frequency can we distinguish between these two situations.

There are two ways to implement the P_{DE} measure. One approach is to actually simulate the processes of corpus-segment creation and collect relevant data. This is the approach that we take in the present study, as it is conceptually in line with how the measure was formulated, and part of our goal is to demonstrate how the measure works. In this approach, we do not need to directly deal with document frequency. However, the important point to note is that when the data are collected, it is the underlying document frequency that is affecting the results of the measure. The other approach is to assess document frequency only and estimate expected results of the P_{DE} measure based on document frequency and binomial probability considerations. We will leave this latter approach to future studies. We believe that the concept and relevant data of document frequency are worthy of future extension.

4.5.5 Number and Size of Corpus Segments

We close our discussion of corpus-segment creation by specifying the actual procedure involved in sampling documents of the BNC. There are two variables that can be manipulated in creating corpus segments: the number of segments, and the size of each segment (in millions of words). WBP that we will use in Chapter 5 provides a base corpus of about 78 million words. To make the maximal use of those data, we will follow a scheme shown in Table 4-7. A check in a cell in Table 4-7 indicates that the combination is possible. With number of segments set to 2, we may examine the effect of

segment size on the P_{DE} measure; these studies are reported in Section 5.1.4. Then, with segment size set to 10 million words, we may examine the effect of the number of segments on the UC measure; see Section 5.2.2 for data report and discussion.

Table 4-7. Number and size of corpus segments in corpus-segment explorations.

	Number of Segments			
Segment Size		2	4	6
	10 million words	√	√	√
	20 million words	√	–	–
	30 million words	√	–	–

The documents of the BNC are randomized into segments as follows. WBP has about 78 million words in total, in 2,688 documents. We set the number of segments n and the size of each segment m for the case at hand. Let us consider creating $n = 6$ segments, each of which has $m = 10{,}000{,}000$ words. With the number of tokens in each document known in advance, we randomly choose one of the 2,688 documents, and randomly assign that document to any of the 6 segments whose total count of tokens is less than 10,000,000. A document that has been assigned to a segment in this way is removed from the available pool so that the next document is randomly chosen from the remaining 2,687 documents; in other words, documents are sampled without replacement. We repeat the steps of randomly choosing a document and randomly assigning it to a segment until all 6 segments achieve a size equal to or greater than 10,000,000.

It follows that when segments are created in this way, it will not be the case that all n segments will have exactly the same number of tokens. Since a segment stops

receiving further documents when its size is equal to or greater than m, and no document of the BNC exceeds 45,000 words (Burnard, 2000), two given segments may differ by up to 45,000 words. For example, segment #1 may have 10,000,300 words, and segment #2 may have 10,044,800 words. A difference of up to 45,000 words may appear to be a sizable gap, but we will consider it to be acceptable for three reasons. First, segments are typically very large: the minimum size that we use (in Section 5.1.5) will be 5 million words, in which case variation in the maximum token count is less than 1%. In other cases, we use segments of 10- or 30-million words, so that the maximum possible size variation constitutes a still smaller percentage of the total. Second, since we are examining word types, the impact of the size "gap" will be less than if token counts were directly at issue. Third, the averaging of values over segments in the P_{DE} and UC measures reduces whatever variability is caused by small differences in segment size.

4.6 Summary

This chapter has been devoted to explaining how the data of the BNC are processed for analysis. We selected 12 suffixes and 1 non-suffix control as the targets for analysis, and reviewed observations expressed in the literature about the productivity of these suffixes. We discussed the structure of the BNC and the way in which word tokens are individuated in the corpus. We also reviewed the processes involved in building a word-type database that can be used to identify relevant words for the analysis targets. Some theoretical and technical difficulties in defining distinct word types with a suffix were addressed, with particular reference to the treatment of prefixation and compounding. Finally, the procedure for creating corpus segments was spelled out, and

arguments presented that the currently proposed measures require corpus segments to be created by random sampling within the BNC at the level of documents, not at the level of words.

Chapter 5 Analysis of Data

The two methods of assessing productivity that were proposed in Chapter 3 will be put to test in this chapter using the BNC, the data of which were processed for analysis as specified in Chapter 4. Section 5.1 presents results for the P_{DE} measure; Section 5.2 presents results for the UC measure; finally, Section 5.3 offers an intuition-based interpretation of the findings reported in Section 5.1.

5.1 Results for the Productivity Measure

We begin our analysis of data by examining outcomes when the P_{DE} measure is applied to WBP. We examine various aspects of the measure, making maximal use of the large number of tokens, some 78 millions words in total, that WBP offers.

5.1.1 Key Concepts and Notations

The P_{DE} measure, a measure of productivity based on the deleted estimation method, was formulated in Sections 3.3 as:

$$(1) \quad P_{DE} = \frac{V_0^{AB} + V_0^{BA}}{V^A + V^B} = \frac{\left(V_0^{AB} + V_0^{BA}\right)/2}{\left(V^A + V^B\right)/2} = \frac{V_N}{V}$$

where given a particular affix and two corpus segments A and B (both of size m), V^A and V^B are the total number of word types with that affix that are present in A and B, respectively, and V_0^{AB} and V_0^{BA} are the numbers of word types with that affix that are present in A but not in B, and in B but not in A. After separate averaging within the denominator and numerator terms (see Section 3.3), V is the average number of word types per segment with that affix, and V_N (V-New) is the average number of word types with that affix that are captured as new.

When each corpus segment is interpreted as a group of randomly sampled speakers (see Sections 3.7 and 4.5), words not shared by the two segments are seen as limited in usage commonality, and thus are likely candidates for new words (see Sections 3.7 and 3.8). Thus, V_N is interpreted as the number of word types with an affix that are new. P_{DE} expresses the degree of productivity of an affix based on the likelihood that a given word type with an affix will be limited in usage commonality, hence new. We also define V_{NN} (V-Non-New) as the number of word types with an affix that are likely non-new. In each application of the P_{DE} measure, the following relationship holds: $V = V_N + V_{NN}$. That is, the P_{DE} measure divides the total number of word types with an affix (V) into the number of new word types with that affix (V_N) and the number of non-new word types with that affix (V_{NN}). It is the proportion of V_N in V that we use to estimate the productivity index for an affix.

5.1.2 Productivity Indices

The result of primary interest is the productivity ranking provided by the P_{DE} measure, among 12 English suffixes and 1 non-suffix control. Table 5-1 presents mean

values for V, V_N, and P_{DE}, averaged over 100 runs, where each run is an independent application of the P_{DE} measure with two segments each having 30 million words sampled from WBP (i.e., 60 million words in total were used in each run). Standard deviations are shown in parentheses. Based on the P_{DE} indices in Table 5-1, Figure 5-1 depicts a productivity ranking of the suffixes; the non-suffix control case (word-ending *ch#*) is also represented.

Table 5-1. Mean values for the P_{DE} measure and its components.

Suffix	V		V_N		P_{DE}	
-ness	1354.9	(12.6)	431.2	(8.4)	0.318	(0.006)
-ity	1008.5	(8.5)	234.4	(6.6)	0.232	(0.006)
-er	2517.8	(14.1)	558.6	(8.7)	0.222	(0.004)
-ee	88.6	(2.5)	26.1	(2.0)	0.295	(0.023)
-ion	2152.9	(15.6)	348.7	(9.6)	0.162	(0.004)
-ment	424.2	(3.6)	61.6	(3.2)	0.145	(0.008)
-th	40.9	(0.9)	3.5	(0.8)	0.085	(0.020)
-ize	437.6	(5.4)	114.5	(4.2)	0.262	(0.009)
-ify	105.8	(2.1)	21.1	(1.8)	0.199	(0.017)
-ish	261.3	(4.2)	90.6	(3.5)	0.347	(0.012)
-ous	639.1	(6.7)	107.1	(5.3)	0.168	(0.008)
-ly	3585.0	(22.2)	754.3	(14.4)	0.210	(0.004)
ch#	213.6	(2.8)	29.7	(2.1)	0.139	(0.011)

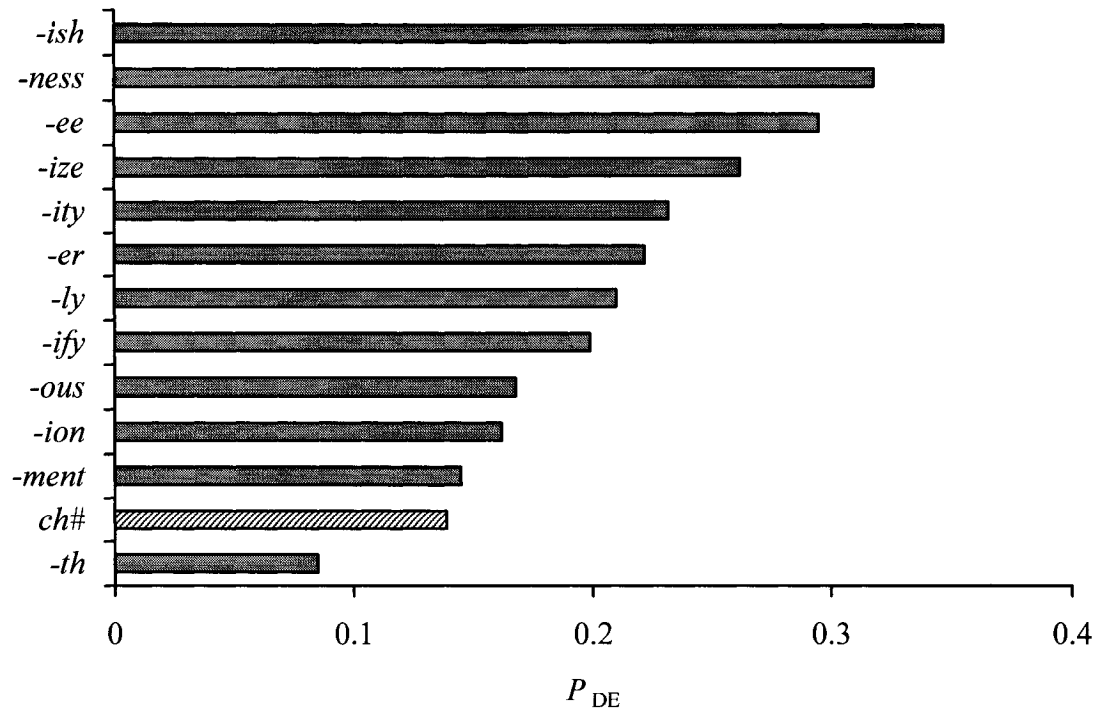

Figure 5-1. Productivity ranking of suffixes based on the P_{DE} measure.

There is a conceptually straightforward way of reading a P_{DE} index: each index value can be multiplied by 100 and understood as the percentage of V that is captured as new (V_N). For example, the P_{DE} index for *-ness* (0.318) can be interpreted as saying that 31.8% of the word types with *-ness* are considered (potentially) new.

Before proceeding further, a question clearly raised by Figure 5-1 is whether it makes sense to compare the productivity of suffixes of different lexical categories on a single scale. Words of different lexical categories serve different functions in a sentence, and exhibit notable differences in overall token frequency in a corpus; for example, Baayen and Lieber (1991) found that the mean token frequency is higher for verbs than for nouns and adjectives, meaning that each verb tended to be used more frequently than

each noun or adjective.[51] We will, nevertheless, propose that across-the-board comparison of suffixes is not only possible but also meaningful. There seems to be nothing inherent in what the P_{DE} measure is designed to capture that prevents us from comparing suffixes of different lexical categories. A comparison of suffixes across lexical categories may even be desirable if we wish to show that a measure is *generally* applicable. We also note that if a productivity ranking of all suffixes were not allowed, it would be impossible to evaluate findings for an affix such as *-ly*, which lacks any directly comparable rival.

Examining the ends of the ranking in Figure 5-1, we find that *-ish* and *-ness* are among the most productive, while *-th* and the non-suffix control *ch#* are among the least productive. These results accord generally with our expectations, although on some grounds we might have expected *-ness* to be more productive than *-ish*—Plag (2003: 44–45), for example, sees *-ness* as more productive than *-ish*. In Section 5.2, we pursue a finer-grained characterization of the difference between *-ish* and *-ness*.

The finding that non-suffix control *ch#* (P_{DE} = 0.139) ranks higher than *-th* (P_{DE} = 0.085) is of considerable interest. We know that the P_{DE} index for *ch#* must necessarily be considered as arising from processes other than derivational affixation, such as the coinage of new simplex forms, compounding, and so on. The P_{DE} index for *ch#* also represents sources of noise that lead to inadvertent capture, such as the occurrence of rare or obsolete words. Under the P_{DE} measure, we do not rule out the possibility that a

[51] This had an effect in their study that *P* indices for affixes that take a verb base, such as *-ee* (0.0016) and *-er* (0.0007), were considerably lower than those that take a non-verb base, such as *-ness* (0.0044) and *-ish* (0.0050).

measure mis-identifies rare or obsolete words as potentially new, since rare or obsolete words are likely to be used by fewer speakers.

In Figure 5-1, we could consider the P_{DE} index for *ch#* as a baseline condition against which a judgment can be made about the extent to which any suffix is to be considered productive. We would find, for example, that P_{DE} value for *-ish* and *-ness* still show strong evidence for high productivity, even when that portion of their productivity values indicated by *ch#* is discounted. With *ch#* taken in this way as the affixational productivity baseline, *-th* would certainly be interpreted as having zero productivity. The V_N for *-th* counts word types such as *coolth*, *handbreadth*, *lukewarmth*, and *tilth*.[52]

We will now turn to the evaluation of individual suffixes, dividing the suffixes into pairs or groups where appropriate. In some cases, we go on to explore the results for a suffix in more detail based on a breakdown of word formation patterns. It is suggested in the literature that productivity varies with the class of bases to which an affix attaches (e.g., Aronoff, 1976; Bauer, 2001; Van Marle, 1985; also see Section 1.3); that is, the *domain of productivity* has to be taken into consideration in assessing productivity of an affix. Ignoring the domain of productivity in some cases carries the risk of giving a false impression of the overall productivity of an affix (Bauer, 2001: 194). Unfortunately, the current word-type database does not provide detailed lexical information about the base of each derived word. However, in many cases, we can rely on orthographic patterns to divide word formations with a given suffix into subtypes (e.g., *-ion* can be divided into *-ization*, *-ification*, and so on).

---

[52] Words such as *handbreadth* result from the particular treatment of forms adopted here (allowing unhyphenated compounds that are justified by in the OED/WD listing, see Section 4.4).

Nominal suffixes *-ness* and *-ity*. The P_{DE} index for *-ness* (0.318) is found to be substantially larger than that for *-ity* (0.232), as predicted. Nevertheless, *-ity* is itself found to be relatively productive, as compared with *ch#* for which the P_{DE} index (0.139) may well overestimate the value expected to arise from noise sources alone for derivational affixation. Suffix *-ity* falls in the middle of the range for the target suffixes of this study.

Word forms in both *-ness* and *-ity* are sufficiently numerous in corpus segments (V values > 1000, see Table 5-1) to permit illustrations of the use of the P_{DE} measure in closer examinations of specific word formation patterns. Within *-ity*, for example, the word formation pattern *-ability* has often been cited as a particularly productive one (e.g., Aronoff, 1976: 51–52; Matthews, 1991: 74–75). Indeed, among the V_N words of *-ity*, we find many productively formed words with this pattern, such as *buyability*, *hummability*, and *cleanability*. Table 5-2 presents a breakdown of values for P_{DE} and its components, comparing forms in *-ability*, its "rival" *-ibility*, versus all remaining patterns in *-ity* (shown as *-ity*†) taken together.

Table 5-2. Word formation patterns within suffix *-ity*.

Pattern	V	V_N	P_{DE}
-ity	1008.5	234.4	0.232
-ability	232.7	85.6	0.368
-ibility	73.8	13.7	0.186
-ity†	702.0	135.1	0.192

Table 5-2 demonstrates the use of the P_{DE} indices at the level of word formation patterns. The P_{DE} index for *-ability*, markedly higher than that for *-ibility* or *-ity*†, indicates that it is

much more likely that a given *-ability* word is new than is usual for *-ity*. The P_{DE} indices are entirely in accord with the observations of the literature. An alternative way of seeing the remarkable productivity of this pattern comes from the observation that while *-ability* occupies 23% of the V of *-ity* (232.7 of 1008.5), it occupies as much as 37% of the V_N for this suffix (85.6 of 234.3). By contrast, for the *-ibility* pattern, the proportional contribution is small for V_N (6%, 13.7 of 234.4). This is another indication that more new words of *-ability* than *-ibility* are coined.

Another claim in the literature, which can also be explored by deploying P_{DE} at the level of word formation patterns, proposes that native English speakers accept more possible words of *-iveness* than *-ivity* pattern, while they accept more possible words of *-ibility* than *-ibleness* pattern (e.g., Aronoff, 1983; also see Section 1.3). Table 5-3 presents a breakdown of P_{DE} in terms of allowing an assessment of this claim; the table considers *-iveness* and *-ivity* word types, in which the base form has *-ive*, and likewise *-ibleness* and *-ibility* word types, in which the base form has *-ible*.

Table 5-3. Word formation patterns in *-ness*/*-ity*, given base form in *-ive*/*-ible*.

Pattern	V	V_N	P_{DE}
Base *-ive*			
-iveness	118.1	25.8	0.219
-ivity	67.3	16.4	0.244
Base *-ible*			
-ibleness	2.1	1.8	0.880
-ibility	73.8	13.7	0.186

An assessment based on P_{DE} certainly does not rank the word formation patterns in an order matching the predictions made in the literature. P_{DE} indices indicate that *-ivity*

is slightly more productive than *-iveness*; and perhaps that *-ibleness* is more productive than *-ibility*. In the latter case, the ordering reflects an astonishing value of the P_{DE} index for *-ibleness* (0.880). Clearly this demands comment.

The fact that the P_{DE} index for *-ibleness* is 0.880 means that in most cases a given *-ibleness* word type is predicted to be new. The reason that the P_{DE} index for *-ibleness* is astonishing is that *-ibleness* words are rarely used (as exemplified in the extremely low V value, 2.1).[53] Anshen and Aronoff (1989: 200) show that according to the OED, the number of new words following the *-ibleness* pattern has been in decline over the past centuries. Aronoff (1983: 169) observes that new *-ibleness* words are not coined anymore; the extraordinarily low value of V_N for *-ibleness* (1.8) in our current data supports his view. It seems entirely unsatisfactory that a word formation that is no longer productive should be assigned the highest productivity index we have observed.

The type of likelihood that the P_{DE} expresses must be carefully distinguished. Based on word types that have been sampled, P_{DE} tells us what proportion of those word types are new. Here, there may be an affix or word formation pattern that is not often used and that has a relatively small number of word types. Nevertheless, P_{DE} may find such an affix or word formation pattern to be productive if in fact a large proportion of the limited number of word types are new. Thus, the high productivity assigned to *-ibleness* is not strictly an error; rather, the measure is honest, given the data that almost all *-ibleness* word types found are new. A list of *-ibleness* words extracted from

[53] It is of interest that the OED has as many as 149 entries for words in the *-ibleness* pattern. The extremely low V value for *-ibleness* (2.1) provides striking evidence that the number of word types listed in the OED does not always reflect the actual language samples.

the entire BNC—not just WBP used in the study reported here—indeed shows that *-ibleness* words are largely unfamiliar: *contemptibleness*, *horribleness*, *responsibleness*, *producibleness*, and *terribleness*.

An alternative analysis of the data in Table 5-3 looks specifically at V_N. Summing V_N values for *-iveness* and *-ivity*, we have 42.2 new words in total that involve *-ive* bases. Of these, 61% (25.8 of 42.2) have been formed with *-ness*, and 39% (16.4 of 42.2), with *-ity*. In this sense, then, *-ness* does appear to lead to more coinage of new words than *-ity*, given an *-ive* base, and the literature's claim is modestly supported. Following the same procedure, we sum V_N values for *-ibility* and *-ibleness*, and find that of 15.5 new words formed with *-ible* bases, 88% (13.7 of 15.5) have been formed with *-ity*, and 12% (1.8 of 15.5), with *-ness*. Again, the finding matches the claim in the literature.

As shown above, V_N alone can be useful in asking questions of a particular form, for example, "Which of two competing suffixes is more likely to be used, given an X base?"

Nominal suffixes *-er* and *-ee*, and adverbial suffix *-ly*. Intuition suggests that *-er* is more productive than *-ee*. However, the P_{DE} index is larger for *-ee* (0.295) than for *-er* (0.222). The current finding is similar to one in Baayen and Lieber (1991), where *P* indices for *-ee* (0.0016) and *-er* (0.0007) also indicate a higher productivity for *-ee*. Baayen and Lieber (1991: 828) attribute the high productivity of *-ee* to the "vogue" nature of this suffix, following a suggestion made by Marchand (1969: 210).

The low P_{DE} index for *-er* causes us some concern because intuition suggests that this suffix seems to lead to the coinage of many words. A similarly low P_{DE} index is also

obtained for the adverbial suffix *-ly* (0.210), and it does not seem satisfactory to find *-ly* being ranked below suffixes such as *-ity*. The word formation process for *-ly* is highly regular: it attaches to almost any adjective with few restrictions (Aronoff, 1976: 37 fn 4; Bauer, 1992, 2001: 6, 130; Baayen & Renouf, 1996: 82–83), and it could be considered as an inflectional affix (Aronoff & Furhop, 2002: 481–482; Haspelmath, 1996: 49–50; Plag, 1999: 113 fn 2; 2003: 97). Table 5-1 reveals that *-ly* and *-er* have the first and second largest values of V, respectively. Conversely, *-ee* has the second smallest value of V. Based on these observations, we might speculate that the size of V may be correlated with the P_{DE} index. However, a Spearman's test shows no significant correlation between V and P_{DE} in the entire set of suffixes, $r_s = .203, p > .10$

Nominal suffixes *-ion* and *-ment*. Both *-ion* and *-ment* are found among the less productive suffixes, and the distance between these two is the narrowest among the rival pairs of suffixes examined so far: the P_{DE} index for *-ion* (0.162) is only slightly larger than that for *-ment* (0.145). The finding that *-ment* is barely productive (based on a comparison with *ch#*) matches Bauer's (1983: 76; 2001: 8–9) observation that the productivity of *-ment* has been in decline, and his claim that new words are barely coined with *-ment*. Baayen and Lieber (1991) also find a low P index for *-ment* (0.0002, cf. 0.0001 for simplex nouns).

The large V for *-ion* may have resulted from the fact that *-ion* has currently been taken to encompass many different word formation patterns, including *-ation*, *-ization*, and *-ification*. Table 5-4 shows P_{DE} values for *-ion*, analyzed to distinguish among different word formation patterns. In the table, *-ation* includes all forms in *-ation*; these

are further broken down into subcategories *-ization*, *-ification*, and *-ation*†, the latter taking words such as *automation*. The category *-ion*† includes words such as *submission*.

Table 5-4. Word formation patterns within suffix *-ion*.

Pattern	V	V_N	P_{DE}
-ion	2152.9	348.7	0.162
-ation	1497.7	279.4	0.187
-ization	357.8	116.0	0.324
-ification	92.1	16.0	0.174
-ation†	1047.8	147.4	0.141
-ion†	655.2	69.3	0.106

Of interest in Table 5-4 is the contrast between *-ization* and *-ification*. The P_{DE} indices indicate that a given *-ization* word is much more likely to be new than a given *-ification* word, and indeed, any other word in *-ion*. This finding is in line with Bauer's (2001: 184–186) OED-based analysis that there have been an increasing number of new *-ization* words and a decreasing number of new *-ification* words over the last two decades. The finding also matches Aronoff's (1976: fn 8) observation that *-ization* pattern is very productive. The P_{DE} index that is higher for *-ation* (0.187) than for *-ion*† (0.106) also matches Matthews' (1991: 76) observation that only the *-ation* pattern is productive in *-ion* word formation.

Aronoff (1983: 166) notes that *-ization* is much more productive than *-izement*. Our data confirm that this is the case. There is only one instance of *-izement* in the entire BNC, *aggrandizement*, which is not included in V_N in the current data. Using the logic of the V_N-based analysis illustrated earlier for *-ness* and *-ity*, we observe the following. In 100% of the cases, the *-ize* bases received *-ion*.

Verbal suffixes *-ize* and *-ify*. The P_{DE} index for *-ize* (0.262) is larger than that for *-ify* (0.199), as predicted. We may have expected a larger separation between these indices, but should note that new words of *-ify* do include some productive word formations, as in *church* + *-ify* → *churchify*, and *town* + *-ify* → *townify*.

Although Baayen and Lieber (1991) find *P* to be extremely low for *-ize* (0.0000710—cf. 0.0007 for *-ity*), and lower yet for *-ify* (0.0000000, i.e., no hapax), our current data show that *-ize* falls among the productive suffixes, without any consideration of the fact that it is a verbal suffix. There seems to be no real problem with comparing the verbal suffixes *-ize* and *-ify* with suffixes of other lexical categories, with the P_{DE} procedure. In the light of the special treatment (e.g., showing more decimal places of the *P* indices) needed for verbal affixes in Baayen and Lieber (1991), we consider our current data for *-ize* and *-ify* as showing the strength of the P_{DE} measure. As proposed earlier in this section, a comparison of affixes across lexical categories is entirely possible.

Adjectival suffixes *-ish* and *-ous*. Finally, we note that the P_{DE} index for *-ish* (0.347) is substantially larger than that for *-ous* (0.168). The distance between *-ish* and *-ous* in the ranking strongly suggests a markedly higher degree of productivity of *-ish*. As compared with *ch#*, *-ous* appears to have at least some degree of productivity. New words with suffix *-ous* contain many biological or technical terms, such as *acephalous* 'headless' or *apodous* 'footless', that are rarely used in conversational English. The occurrence of these terms are likely due to the fact that we are sampling texts from books and periodicals, and that some of these touch on technical matters.

Summary Remarks. To summarize the preliminary examination of P_{DE} indices, we found that results with the P_{DE} measure as presented in Table 5-1 (and the data in subsequent tables) are overall in line with many claims in the literature, and apparently informative about the productivity of suffixes. We have demonstrated the use of P_{DE} for suffixes, the use of P_{DE} for word formation patterns, and the use of V_N alone. Which method to use to analyze data depends on what question is being asked. We have also noted an important characteristic of the P_{DE} measure, namely that it could assign a high degree of productivity to an affix or a word formation pattern that is used less often as long as a large proportion of its word types are likely to be found new. While there have been no "puzzling" surprises, some results were found to deviate slightly from our expectations, as seen particularly in the relatively low P_{DE} indices for *-er* and *-ly*. We will return to this issue in Section 5.3 where an intuition-based interpretation of the data of the P_{DE} measure will be proposed.

5.1.3 Unlisted Words and New Words

As noted in Section 3.5, although a direct verification of a productivity measure is difficult, a measure can nevertheless be indirectly verified by examining how the measure captures new words. Despite the general difficulty of identifying new words, previous researchers have sought an objective means of identifying new words by turning to dictionary entries (Baayen & Renouf, 1996). On the assumption that words not listed in a comprehensive dictionary are likely new, we could consider words in a corpus that are not listed in a dictionary such as WD (Baayen & Renouf, 1996) as new, and examine where these unlisted new words fall in the critical components of the P_{DE} measure. The

assumption underlying the link between unlisted words and new words, however, is not free of problems. Baayen and Renouf (1996: 77) point out that it is unlikely for a word attested only once in an appropriately large corpus of texts to be a non-neologism simply because it is listed in WD. Since WD is aimed at being comprehensive, new words identified based on their absence in WD are rather limited ones. Another problem pointed out by Baayen and Renouf (1996: 77–78) lies in mixing corpus data and dictionary data:

> From a statistical point of view, combining data from dictionaries and a corpus is somewhat unfortunate. Whether or not a regular morphologically complex word is present in a dictionary is to a large extent arbitrary, there is no common sampling scheme for the dictionary-based and corpus-based counts, and the analysis of the mixed data becomes more complicated, both practically and theoretically.

Despite these problems, the use of WD does provide one objective method of evaluating how new words are captured by a productivity measure.

There are two methods of identifying new words involved in the test that follows. One is the P_{DE} measure, which divides the total number of word types (V) into the number of new word types (V_N) and the number of non-new word types (V_{NN}). The other, as described above, is the WD-based measure that divides V into the number of word types unlisted in WD (henceforth, WD_{UL}) and the number of word types listed in WD (henceforth, WD_L). Figure 5-2 shows a contingency table that depicts the relationship between these two measures.

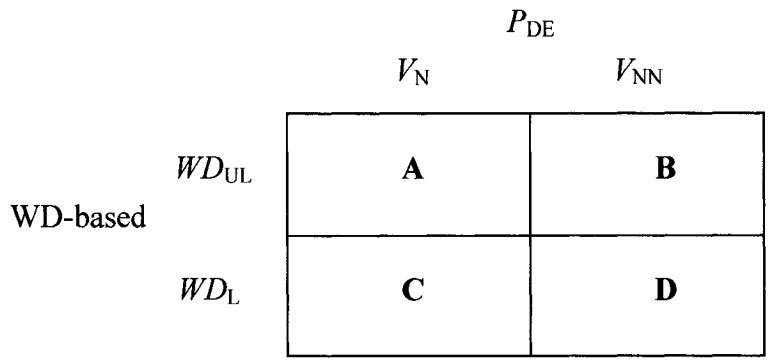

Figure 5-2. Contingency table for the P_{DE} and WD-based measures.

Given the comprehensiveness of WD, what WD_{UL} represents is a rather conservative estimate of the number of new words. What we would wish, then, is for the P_{DE} measure to identify words contributing to WD_{UL} as new words. In other words, the two measures should be associated, rather than independent of each other, in capturing words that fall into cell **A** in Figure 5-2. Table 5-5 shows the number of new and non-new word types identified based on the two measures.

Table 5-5. New versus non-new word types by the P_{DE} and WD-based measures.

Suffix	V_N	V_{NN}	WD_{UL}	WD_L	V_N words in WD_{UL}
-ness	431.2	923.7	150.0	1204.9	102.0
-ity	234.4	774.1	79.4	929.1	62.1
-er	558.6	1959.2	354.1	2163.7	170.6
-ee	26.1	62.5	14.0	74.6	11.7
-ion	348.7	1804.3	155.1	1997.8	88.7
-ment	61.6	362.6	39.1	385.1	18.5
-th	3.5	37.4	2.1	38.8	1.2
-ize	114.5	323.0	36.2	401.4	26.1
-ify	21.1	84.7	9.6	96.2	7.3
-ish	90.6	170.7	53.2	208.1	47.4
-ous	107.1	532.0	35.6	603.5	20.0
-ly	754.3	2830.7	311.0	3274.0	210.9
ch#	29.7	183.9	9.4	204.2	3.4

If the P_{DE} measure and the WD-based measure are associated, the value for cell **A** in Figure 5-2 must be larger than the value expected when the two measures are independent. The last column of Figure 5-2 shows the observed value for cell **A**. We examine this point by contrasting the expected percentage of V_N in WD_{UL} and the observed percentage of V_N in WD_{UL}. Figure 5-3 shows a scatter plot of these two elements.

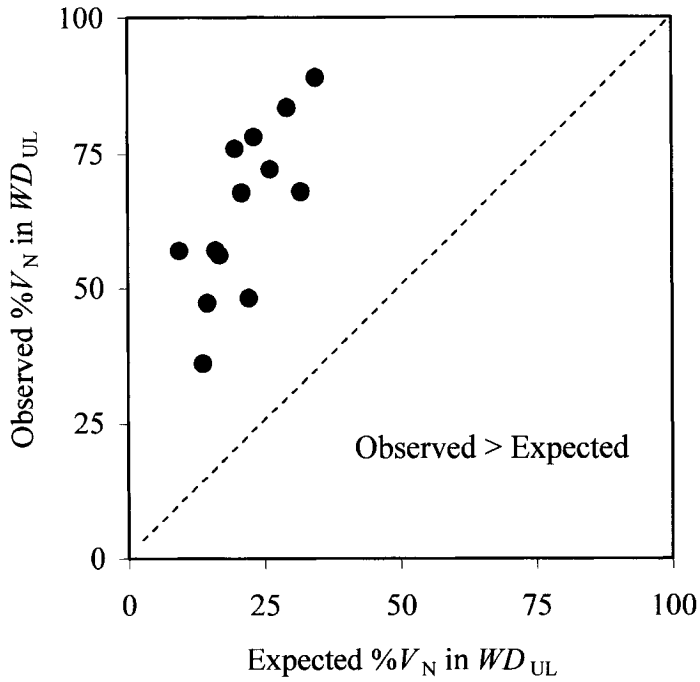

Figure 5-3. Scatter plot for Expected %V_N in WD_{UL} and Observed %V_N in WD_{UL}.

It is evident in Figure 5-3 that the observed percentage of V_N in WD_{UL} is consistently and substantially larger than the expected percentage in all suffixes. These data indicate that, given word types contributing to WD_{UL}, the P_{DE} measure effectively identified a large percentage of them as new words, and this percentage is larger than what would be expected if the two criteria (measures) of identifying new words were independent.

5.1.4 Effects of the Size of Corpus Segments

Having considered results for the P_{DE} measure obtained with 30 million words in each corpus segment, we would now examine the extent to which the behavior of the measure is dependent on the size of corpus segments. In this section, again using WBP,

we look at how P_{DE} changes as segment size increases.[54] Table 5-6 presents P_{DE} values and rank orders for the target suffixes, as a function of segment size which increases in intervals of 10 million words.

Table 5-6. Mean P_{DE} values and suffix rank orders, as a function of segment size.

Suffix	P_{DE}		
	10 million	20 million	30 million
-ish	0.322 [2]	0.332 [2]	0.347 [1]
-ness	0.371 [1]	0.336 [1]	0.318 [2]
-ee	0.313 [3]	0.301 [3]	0.295 [3]
-ize	0.262 [4]	0.260 [4]	0.262 [4]
-ity	0.238 [6]	0.235 [6]	0.232 [5]
-er	0.260 [5]	0.236 [5]	0.222 [6]
-ly	0.226 [7]	0.215 [7]	0.210 [7]
-ify	0.179 [8]	0.191 [8]	0.199 [8]
-ous	0.164 [10.5]	0.165 [9]	0.168 [9]
-ion	0.169 [9]	0.163 [10]	0.162 [10]
-ment	0.153 [12]	0.148 [12]	0.145 [11]
ch#	0.164 [10.5]	0.153 [11]	0.139 [12]
-th	0.078 [13]	0.081 [13]	0.085 [13]

Note: The suffixes are shown in a descending order based on P_{DE} in the 30-million-word case. Each P_{DE} value is a mean over 100 simulation runs.

Overall, P_{DE} index values are remarkably similar among the three segment-size series represented in Table 5-6. A Friedman's test finds no significant difference in P_{DE} among the three series, $\chi^2(2, 13) = 2.627, p > .10$. A Spearman's test shows that the P_{DE}

[54] As is clear from the formulation of the P_{DE} measure, a change in segment size occurs simultaneously to both two segments.

indices are highly positively correlated: for 10 vs. 20 million words, $r_s = .990$, $p < .01$; for 20 vs. 30 million words, $r_s = .984$, $p < .01$; and for 10 vs. 30 million words, $r_s = .971$, $p < .01$.

Based on the above findings, we observe that the P_{DE} measure provides consistent estimates of the productivity of suffixes across corpus segments of different sizes. The ranking of suffixes is stable, and changes in the ranking are by and large restricted to minor re-locations of rank-adjacent suffixes. That is, there are no surprises in the findings over different sample sizes.

5.1.5 Comparison of Spoken and Written Texts

We have thus far exclusively used WBP, but it is also of interest to see how the P_{DE} measure performs on different components of the BNC. Of particular interest is the difference between written and spoken components. Corpus texts of different genres and registers exhibit differences in the type of words used (see, e.g., Biber, 1989); for example, the words used in texts will presumably differ depending on whether the texts are likely to involve concepts that are abstract (as in journals) or concrete (as in conversations). Productivity is also expected to differ among corpus texts of different genres and registers (Plag, Dalton-Puffer, & Baayen, 1999, see below).

Plag, Dalton-Puffer, and Baayen (1999) examined the productivity of 15 English suffixes across three components of the BNC: (a) Written (about 90 million words), (b) Spoken Context Governed (about 6 million words), and (c) Spoken Demographic (about

4 million words).[55] Spoken component (b) represents interviews, speeches, and meetings while spoken component (c) represents spontaneous conversations. Based on word frequency data of the BNC made available by the compilers of the BNC, Plag, Dalton-Puffer, and Baayen statistically estimated how many word types with each suffix would be encountered as words of each component are processed, with the assumption that words occur randomly in each component. Values for V for the three components were estimated for an increasing sampling size of up to about 10 million words (because (b) and (c) have less than 10 million words). Plag, Dalton-Puffer, and Baayen found for all suffixes that a larger number of word types are expected in the written component than in either of the spoken components. Their data showed that as words of the components are processed, the number of sampled word types with a suffix increases at a much higher rate in the written component than in the other components.

Since the P_{DE} measure requires a substantial amount of data for distribution into two segments, we prepare the spoken component of the BNC by combining the "Spoken Demographic" and "Spoken Context-Governed" components of the BNC (see Section 4.2), which will offer some 10 million words in total. The maximal use of this spoken component will be made if the P_{DE} measure is implemented with 5 million words in each segment. We will contrast results from the spoken component against results from the written component that consists only of WBP (similarly sampling 5 million words per segment).

Table 5-7 shows results of a single application of the P_{DE} measure applied to the written and spoken components. Note that only one application of the measure is made

[55] Descriptions of these components are briefly sketched in Table 4-2 in Section 4.2.

because the written component is much larger and repeated application of the measure to a larger text-source will of itself lead to the capture of more word types, quite aside from the text-type differences between the corpora.

Table 5-7. Values for the P_{DE} measure, for the written and spoken components.

Suffix	Written				Spoken			
	V	V_N	P_{DE}		V	V_N	P_{DE}	
-ness	677.5	283.5	0.418	[1]	219.0	109.0	0.498	[1]
-ee	48.0	19.0	0.396	[2]	23.0	8.0	0.348	[3]
-ish	131.0	50.0	0.382	[3]	81.0	38.0	0.469	[2]
-er	1455.5	426.5	0.293	[4]	917.0	278.0	0.303	[5]
-ize	242.0	66.0	0.273	[5]	145.0	47.0	0.324	[4]
-ly	2113.0	513.0	0.243	[6]	941.0	265.0	0.282	[7]
-ity	581.0	133.0	0.229	[7]	284.5	77.5	0.272	[8]
-ify	74.0	15.0	0.203	[8]	48.0	11.0	0.229	[9]
-ous	431.0	77.0	0.179	[9]	219.5	64.5	0.294	[6]
-ion	1427.0	248.0	0.174	[10]	869.5	167.5	0.193	[11]
-ment	295.0	46.0	0.156	[11]	194.5	40.5	0.208	[10]
ch#	148.0	23.0	0.155	[12]	103.5	15.5	0.150	[13]
-th	35.5	3.5	0.099	[13]	25.0	4.0	0.160	[12]

Note: The suffixes are shown in a descending order based on P_{DE} values for the written component.

While the productivity rankings do not differ markedly, a Wilcoxon signed rank test shows a significant difference, $z = -2.621$, $p < .01$. The two rankings, nevertheless, seem to be in approximate agreement with respect to identifying more productive versus less productive suffixes. We also find that the ranking of suffixes for the written

component is somewhat different from one obtained in the earlier section with 30 million words in each segment (see Table 5-6).

We note in Table 5-7 that, matching the observation by Plag, Dalton-Puffer, and Baayen (1999), the written component has larger values of V in all suffixes than the spoken component. V_N is also overall larger for the written component.

Consistently higher values of V in the written component may particularly reflect the fact that our written component consists only of WBP and that the texts of books and periodicals may require a wider range of words to express various concepts. Nevertheless, based on V alone, it is not certain whether more new words or existing words were used in the written component. Figure 5-4 shows a scatter plot comparing written and spoken components: here, the horizontal axis takes the ratio of written-to-spoken for V, that is, the extent to which V in the written component exceeds that in the spoken component, while the vertical axis does the same for V_N, that is, the extent to which V_N in the written component exceeds that in the spoken component.

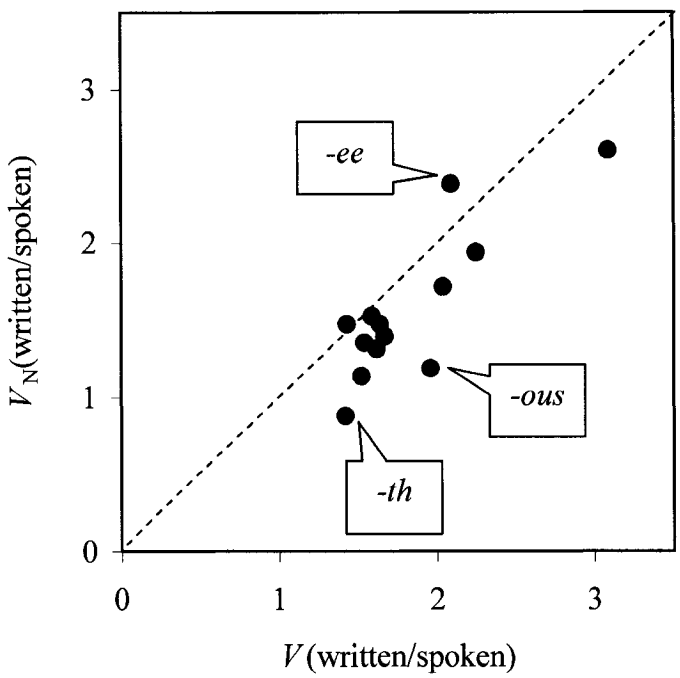

Figure 5-4. Scatter plot comparing V and V_N for written and spoken components.

If the need for more word types in the written component means the need to coin more new word types, the data values in the scatter plot should be located near $x = y$. In Figure 5-4, it seems to be the case that for most suffixes, more word types used in the written component means more new words are coined in that component also. Notable exceptions are *-ee*, *-ous*, and *-th*. In the case of *-ee*, the need for more word types in the written component is met with much more coinage of new words. Conversely, for *-ous* and *-th*, the need for more word types in the written component is not met by more coinage of new words. The implication of the latter case seems to be that for an affix of this type, more word types out of existing words are used in the written component.

The above finding indicates that the written and spoken components of the BNC exhibit some differences in the number of word types used as well as in the productivity

of suffixes. In general, more word types are used in the written component. We also found that there are differences among suffixes as to whether the need for more word types in the written component is met by more coinage of new words. For some suffixes, it seem that more word types in the written component means the use of more existing, rather than new, word types.

5.2 Results for the Usage-Commonality Measure

Having discussed results for the P_{DE} measure, we will now turn to examining results for the UC measure. The purpose of this section is twofold. First, the relationship between the P_{DE} measure and the UC measure will be examined empirically with the BNC data. Clarifying this relationship not only deepens our understanding of the P_{DE} measure but also facilitates the evaluation of data captured by the UC measure. Second, we will examine how the UC measure depicts aspects of productivity that cannot be captured by the P_{DE} measure alone.

5.2.1 Key Concepts and Notations

We extend the mechanism of the P_{DE} measure into the UC measure, a measure of the usage commonality of words across corpus segments. Recall that the P_{DE} measure divides the total number of word types with an affix (V) into the number of new word types with that affix (V_N) and the number of non-new word types with that affix (V_{NN}) based on a cross-comparison of two segments (where unshared words are new and shared words are non-new). In the UC measure, the number of cross-compared segments is

greater: in particular, we examine 4- and 6-segment cases. Given a particular affix and n segments (each of which having size m) in each of those segments, Sr is the number of word types with that affix that are found in r segments ($1 \leq r \leq n$); that is, Sr is the number of word types with that affix that a given segment shares with ($r - 1$) other segments. For example, if we have 6 segments labeled #1 through #6, from the viewpoint of #1, given a particular affix, $S1$ is the number of word types with that affix that are present only in #1, $S2$ is the number of word types with that affix that are present in #1 and one other segment, $S3$ is the number of word types with that affix that are present in #1 and two other segments, and so on.

We let Ssum denote the sum of the Sr data; for instance, in the 6-segment case, $S\text{sum} = S1 + S2 + ... + S6$. In other words, Ssum for an affix is the total number of word types with that affix that are present in a given segment. The data of Sr and Ssum will be obtained individually for each of the n segments, and will be averaged over the n segments to obtain more representative data of Sr and Ssum for a corpus segment of size m. Since the averaging of the Sr and Ssum data is mandatory, we use the notations Sr and Ssum in our data analysis to refer to mean values over n segments.

Similarly to the P_{DE} measure, each corpus segment is interpreted as a group of randomly sampled speakers (see Sections 3.7 and 4.5 for discussion). Sr then expresses how many words are shared by r groups of speakers. The Sr categories form the scale of usage commonality, where $S1$ is the number of word types that are most limited in their usage distribution. We will be examining the ratio of Sr to Ssum, which will be alternatively described as $\%Sr$; in particular, $\%S1$ is of particular interest since it expresses the percentage of words that are most limited in usage commonality.

The major difference between the UC measure and the P_{DE} measure is that the UC measure (via its examination of many segments) makes finer and non-categorical distinctions among new and non-new words based on the usage distribution of words, whereas the P_{DE} measure draws a categorical distinction between new and non-new words to yield a productivity index. However, we could take the view that the P_{DE} measure is merely the 2-segment case of the more general UC measure. As will be shown shortly, the 2-, 4-, and 6-segment cases all share the basic mechanism in capturing new words. The following relationships will be shown to hold between the P_{DE} measure and the 2-segment case of the UC measure: (i) $V_N = S1$; (ii) $V_{NN} = S2$; (iii) $V = S1 + S2 = S\text{sum}$; and (iv) $P_{DE} = V_N/V = S1/S\text{sum} = \%S1$. Comparing the 2-segment case against 4- and 6-segment implementations of the UC measure provides a useful means of examining the relationship between the P_{DE} measure and the UC measure.

5.2.2 Effects of the Number of Corpus Segments

We begin our examination of the relationship between the UC measure and the P_{DE} measure, viewing P_{DE} as the 2-segment case of the UC measure, by seeing how the data of the UC measure differ among the 2-, 4-, and 6-segment cases. Our primary interest lies in how 2- and 6-segment cases differ, in particular, since these are extreme cases in term of the numbers of segments (in an admittedly limited range, albeit).

The examination turns on three elements of interest: *S*sum, *S*1, and *S*right. As discussed in Section 3.8, we let *S*right denote the sum of word types captured in the right half of the usage-commonality scale:

(2) $Sright = \sum_{r=1}^{n/2} Sr$

That is, *S*right for 2 segments is *S*1 only, *S*right for 4 segments is (*S*1 + *S*2), and *S*right for 6 segments is (*S*1 + *S*2 + *S*3).

We examine values for *S*1, *S*sum, and *S*right as a function of the number of segments, with 10 million words sampled per segment (i.e., a total of 20, 40, and 60 million words are used for the 2-, 4-, and 6-segment cases, respectively). Table 5-8 presents values for *S*1, and Table 5-9 for *S*sum and *S*right. All data are means over 100 simulation runs.

Table 5-8. *S*1 among cases differing in number of segments.

Suffix	*S*1		
	2 segments	4 segments	6 segments
-ness	308.4	168.5	121.1
-ity	171.1	91.7	66.2
-er	458.6	224.3	152.7
-ee	18.0	10.1	7.6
-ion	284.1	139.6	95.7
-ment	52.0	24.9	16.9
-th	2.9	1.3	1.0
-ize	79.3	44.0	33.2
-ify	14.5	8.2	6.2
-ish	53.6	33.3	27.4
-ous	82.4	41.5	30.1
-ly	589.6	297.6	209.3
ch#	27.5	168.5	7.5

Table 5-9. *S*sum and *S*right among cases differing in number of segments.

Suffix	*S*sum			*S*right		
	2 seg.	4 seg.	6 seg.	2 seg.	4 seg.	6 seg.
-ness	831.7	831.6	832.4	308.4	306.3	309.2
-ity	718.6	718.9	719.5	171.1	168.5	168.5
-er	1761.0	1761.8	1760.8	458.6	444.7	441.5
-ee	57.6	58.1	58.0	18.0	17.6	17.2
-ion	1681.1	1685.4	1685.5	284.1	272.8	269.3
-ment	339.3	339.7	339.3	52.0	49.5	49.3
-th	36.4	36.2	36.3	2.9	2.6	2.7
-ize	302.2	302.8	303.0	79.3	77.1	76.4
-ify	80.8	81.1	80.9	14.5	14.0	13.8
-ish	166.5	166.8	167.5	53.6	52.0	51.9
-ous	500.6	500.7	501.2	82.4	79.3	78.2
-ly	2610.2	2609.9	2612.5	589.6	567.2	562.1
ch#	167.9	168.3	168.4	27.5	27.6	28.0

In Table 5-8, it is evident that *S*1 values decrease systematically as the number of segments increases. This decrease in *S*1 is understandable given that more segments allow finer distinctions to be made between new and non-new words. Thus, smaller values of *S*1 with more segments could be seen as *S*1 capturing new words more purely; some words that are included in *S*1 in the 2-segment case will fall into *S*2 or *S*3 in the 6-segment case, because their usage commonality is not tightly limited.

We note that *S*sum values in Table 5-9 are remarkably stable over changes in the number of segments. Recall that although the three cases use different numbers of words in total, the unit for which we assess *S*sum is always one segment (of size 10 million words), so it is not unexpected that *S*sum should be similar in this way. Since *S*sum is

stable, it is naturally expected that the differences as a function of number of segments lie in how *S*sum is divided into the *Sr* categories: in the 2-segment case, *S*sum will be distributed over *S*1 and *S*2, whereas in the 6-segment case, *S*sum will be distributed over *S*1 through *S*6.

The values of *S*right in Table 5-9 are also quite similar. A Friedman's test, nevertheless, finds a significant difference in *S*right among the three cases, $\chi^2(2, 13) = 12.745$, $p < .01$. Conducting pair-wise comparisons, a Wilcoxon signed rank test finds a significant difference between 2- vs. 4-segment cases, $z = -3.110$, $p < .01$, and between 2- vs. 6-segment cases, $z = -2.691$, $p < .01$, but no such difference at a conventional level of significance for 4- vs. 6-segment case, $z = -1.766$, $p > .05$. Thus, there seems to be a gap between the 2-segment case on one hand and the 4- and 6-segment cases on the other hand, which may result from the crudeness in identifying new words in the 2-segment case. Such a gap, however, runs against our claim that the 2-segment case is inherently similar to other more fine-grained cases.

Here, we need to consider what we wish to be similar among the different implementations of the UC measure. We certainly do not make use of *S*right by itself in an analysis of productivity, so *S*right alone is not the focus. What we wish to be consistent among the implementations of the UC measure is the description of productivity, namely the ranking of suffixes based on P_{DE}. P_{DE} indices, if we wish, should be obtainable from any implementation of the UC measure, taking *S*sum as V, *S*right as V_N, and *S*right/*S*sum as $V_N/V = P_{DE}$. Table 5-10 shows P_{DE} obtained by *S*right/*S*sum (i.e., P_{DE}) and suffix rank order for different numbers of segments.

Table 5-10. *S*right/*S*sum (P_{DE}) for different numbers of segments.

Suffix	*S*right/*S*sum					
	2 segments		4 segments		6 segments	
-ness	0.371	[1]	0.368	[1]	0.371	[1]
-ish	0.322	[2]	0.312	[2]	0.310	[2]
-ee	0.313	[3]	0.303	[3]	0.297	[3]
-ize	0.262	[4]	0.255	[4]	0.252	[4]
-er	0.260	[5]	0.252	[5]	0.251	[5]
-ity	0.238	[6]	0.234	[6]	0.234	[6]
-ly	0.226	[7]	0.217	[7]	0.215	[7]
-ify	0.179	[8]	0.173	[8]	0.171	[8]
-ion	0.169	[9]	0.162	[9]	0.160	[9]
-ous	0.165	[10]	0.158	[10]	0.156	[10]
-ment	0.153	[11]	0.146	[11]	0.145	[11]
-th	0.080	[12]	0.072	[12]	0.074	[12]
ch#	0.164	[-]	0.164	[-]	0.166	[-]

A careful examination of the P_{DE} values in Table 5-10 reveals that if we exclude *ch#*, the three rankings of suffixes are identical. This is supporting evidence that similar results are expected from the UC measure and the P_{DE} measure.

The relationship between the two measures is as follows. The P_{DE} measure and the multiple-segment cases of the UC measure capture new words in a similar manner, dividing words into new words and non-new words based on their usage commonality. If we wish to have indices of productivity, we make use of the *S*right/*S*sum of the 2-segment case as: *S*right/*S*sum = *S*1/*S*sum = $V_N/V = P_{DE}$. The computational simplicity of cross-comparing only two segments is attractive as a productivity measure, so we assign the 2-segment case a special status as the P_{DE} measure. If we wish for more detailed

analyses of productivity, *S*right is divided into *S*1 and *S*2 in the 4-segment case, and into *S*1, *S*2, and *S*3 in the 6-segment case, and the examination of the ratio *Sr*/*S*sum provides a dissection of a P_{DE} index. How that dissection of the P_{DE} index plays out will be discussed in the next section.

5.2.3 Usage Distribution of Words

In this section, we evaluate data obtained with a 6-segment implementation of the UC measure, where each segment had 10 million words sampled from WBP (i.e., a total of 60 million words was sampled in each simulation run). One simulation run consists of creating 6 segments, and obtaining average values of *S*1 through *S*6, and *S*sum. All data for *Sr* and *S*sum throughout this section are means over 100 simulation runs. To illustrate how the *Sr* data can be evaluated, we will first focus on four nominal suffixes, among which the contrast of *-ness* and *-ity* is the parade case for studies of productivity, and *-ment* and *-th* represent suffixes that are limited in productivity. Figure 5-5 shows the distribution of %*Sr* for *-ness*, *-ity*, *-ment*, and *-th*.

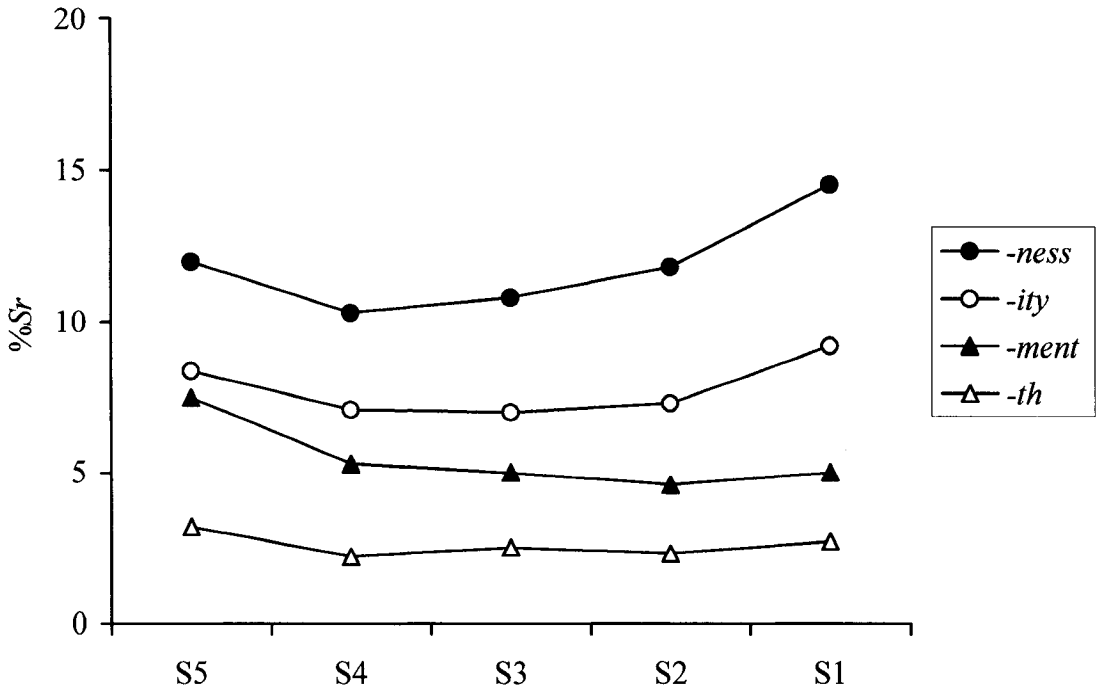

Figure 5-5. %*Sr* for *-ness*, *-ity*, *-ment*, and *-th*.

Note that in constructing Figure 5-5 we systematically omit *S*6, on the grounds that *S*6 fundamentally differs from the other *Sr* categories; the words it represents are known with certainty (by our measures) to not be new. (%*S*6 is, of course, the complement of the sum of %*S*1 through %*S*5.)

What is of particular interest is how %*Sr* changes from *S*3 to *S*1. We claimed in the previous section that the %*S*3, %*S*2, and %*S*1 of the 6-segment case provide a dissection of the %*S*1 (namely P_{DE}) of the 2-segment case. Examining a P_{DE} index for a suffix only gives us a "one-step" analysis of the productivity of that suffix, and it may not always be clear how a given suffix comes to have a particular P_{DE} index. Examining $\%S_3$ through %*S*1 provides us with a "three-step" analysis of the ways a suffix is used in the language, and we consider that this provides valuable insights.

There are two points to note in drawing implications from the distribution of usage of words in Figure 5-5. The first concerns the separation of the curves at different values of *Sr*. For *-ness* and *-ity*, in particular, as we focus more on words more likely to be new by moving toward *S*1 along the *x*-axis, the separation between the two curves increases. A separation of the curves that is greatest at *S*1 is an indication of a higher productivity of *-ness*. In fact, the P_{DE} indices in Section 5.1.2 found *-ness* (0.318) to be more productive than *-ity* (0.232).

The second point to note in Figure 5-5 is that the data for *-ness* and *-ment* show a separation of curves, overall, that is much greater than that for *-ness* and *-ity*. The P_{DE} index for *-ment* (0.145) revealed earlier that *-ment* is barely productive (cf. *ch#* with P_{DE} = 0.139), and that limited productivity is clearly reflected in the *Sr* data. Note that the curve for *-ment* does not have an upward inflection at *S*1, and that height of the curve is low, overall. The height of the curve is even more strikingly low for *-th*. The data for *-th* are in line with our finding based on the P_{DE} indices that this suffix is wholly unproductive (P_{DE} = 0.085).

To explore other *Sr* data, we will examine two further "rival" pairs: *-ee* versus *-er*, and *-ish* versus *-ness*. The *Sr* data for these suffixes are provided in Figure 5-6 and Figure 5-7, respectively.

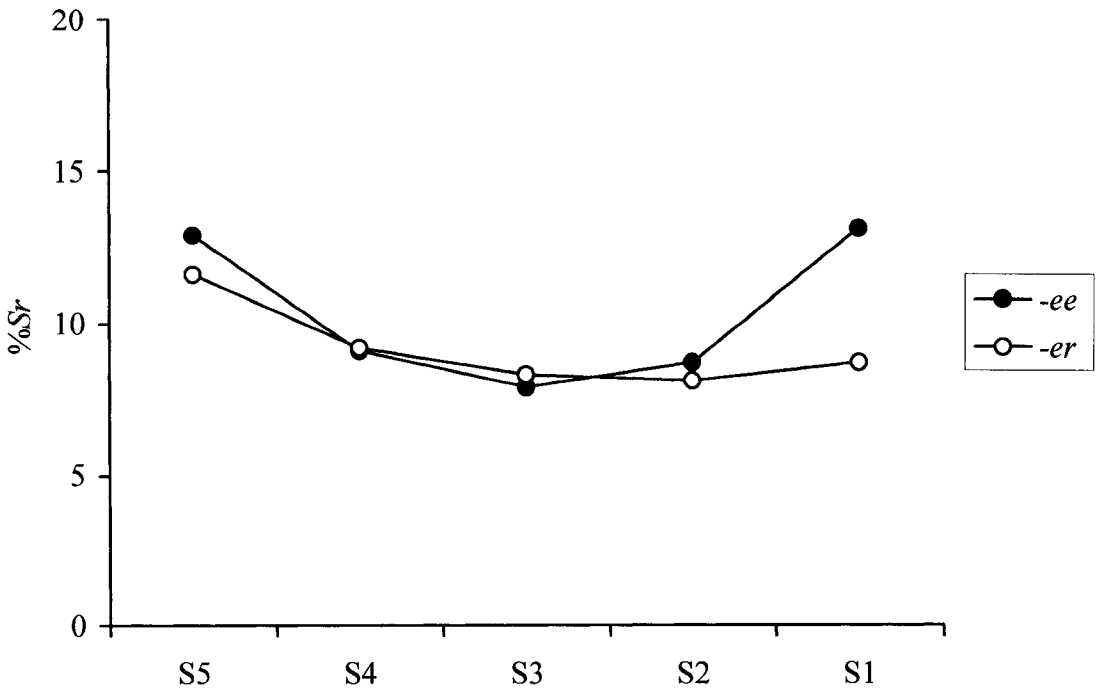

Figure 5-6. %*Sr* for *-ee* versus *-er*.

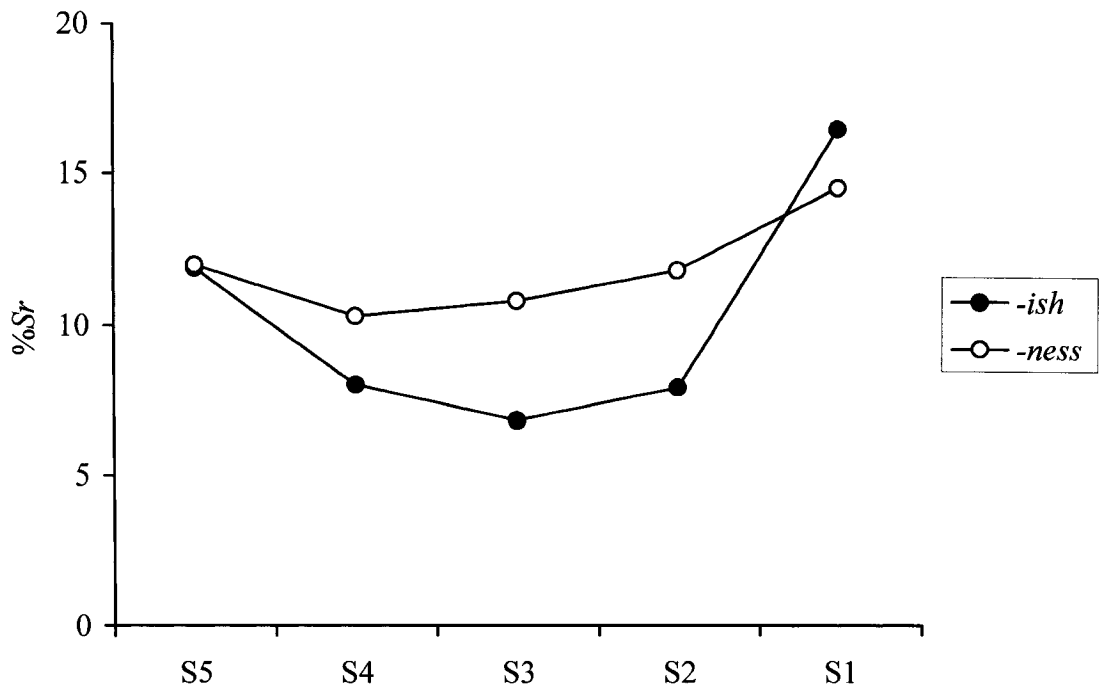

Figure 5-7. %*Sr* for *-ish* versus *-ness*.

In Figure 5-6, we see that the curves for *-ee* and *-er* are quite similar from *S*5 to *S*2, but that the *-ee* curve has a marked upward inflection at the right end. The increase in %*Sr* for *-ee* from *S*2 to *S*1 is key in differentiating its behavior from that of *-er*. We will return to this characteristic of *-ee* shortly, noting here only that *Sr* data provide an explication of the fact that the P_{DE} index has *-ee* (0.295) more productive than *-er* (0.222).

A comparison of the data for *-ish* and *-ness* in Figure 5-7 reveals an interesting characteristic difference between the productivity patterns of the two suffixes: while %*Sr* for *-ness* steadily increases toward *S*1, *-ish* shows a sharp increase in %*Sr* at *S*1, with the result that curves cross over. The P_{DE} index found *-ish* (0.347) to be more productive than *-ness* (0.318). Plausibly, we can interpret the difference between the two suffixes as follows: the word formation process with *-ness* results in word types that are somewhat limited in usage commonality and thus "mildly" new, whereas *-ish* leads to word types that are sharply limited in usage commonality and thus very new. This characteristic difference between *-ish* and *-ness* can be revealed only by the UC measure; the P_{DE} indices simply tell us that *-ish* is more productive on average than *-ness*.

The data for *-ish* and *-ee*, exhibiting sudden increases in %*Sr* at *S*1, point to the importance of *S*1 in analyzing new words: the characterization of these suffixes as notably productive crucially depends on values at the extreme of the usage-commonality scale. Prompted by this observation, Figure 5-8 compares the change in %*Sr* (increase/decrease) from *S*2 to *S*1 in the 6-segment case, for some selected suffixes. Each data point is calculated as %*Sr* minus %*S*2 (and hence, the left-hand side datum is 0 for all suffixes, (%*S*2 − %*S*2) = 0).

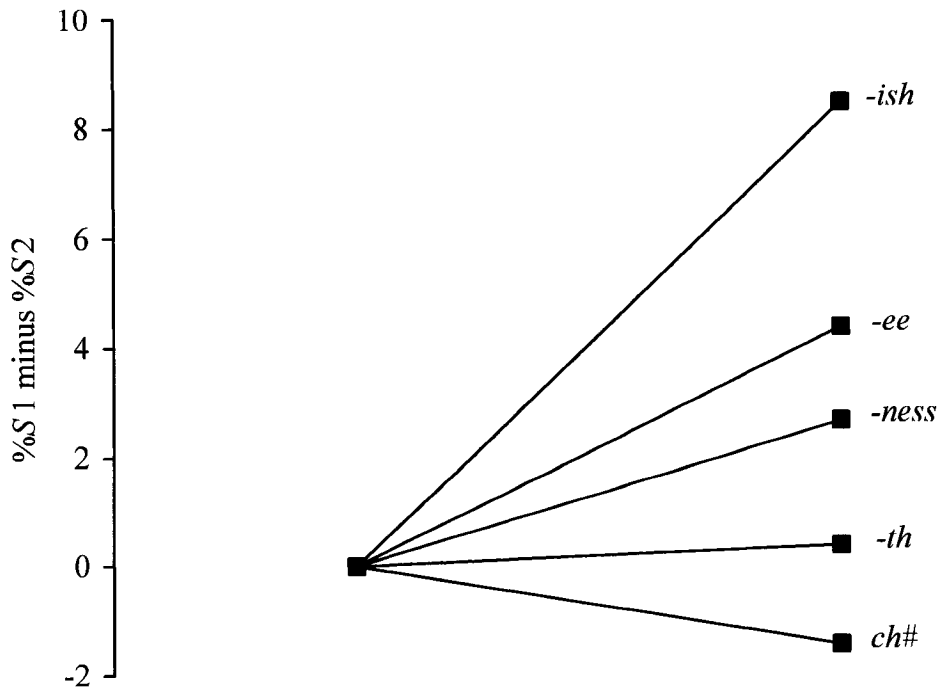

Figure 5-8. Change in %*Sr* from *S*2 to *S*1.

We find that the order of these suffixes in Figure 5-8 (*-ish*, *-ee*, *-ness*, and *-th*) and non-suffix control *ch#* does not correspond in detail to one based on the P_{DE} indices. P_{DE} indices (see Section 5.1.2) rank these as follows:

(3) P_{DE} ranking *-ish* > *-ness* > *-ee* > *ch#* > *-th*

(4) %*S*1 − %*S*2 *-ish* > *-ee* > *-ness* > *-th* > *ch#*

In particular, *-ness* and *ch#* are re-positioned. Why should this be the case?

The order of suffixes provided here is not to be understood as reflecting the degree of productivity *per se*, since we maintain that productivity—construed technically—is specified by P_{DE}. Rather, the data in Figure 5-8 perhaps implicate

something we might think of as a "surprise factor" associated with the use of a new word with a suffix. A suffix with a steeper slope in Figure 5-8 is one whose new words are surprisingly new (so that many fall into *S*1, instead of *S*2 or *S*3), meaning that the use of such new words is likely to draw the attention of listeners. Under this interpretation, new words with *-ish* and *-ee* are expected to be experienced often as notably new, in contrast to those with *-th* and *ch#* which are not likely to occasion surprise, routinely. As for *-ness*, the fact that the suffix leads to a substantial proportion of *mildly* new words (those falling into *S*2 and *S*3) may somehow bleach any surprise associated with the use of new words with *-ness*.

We have shown in this section that there are aspects of productivity expressed by the UC measure that cannot be expressed by the P_{DE} measure. The new words captured by the P_{DE} measure are not equally new, and their usage commonality provides a method of refining their degree of newness. Our understanding of the degree of productivity of suffixes, as specified by the P_{DE}, is enhanced by considering the usage commonality of words as expressed by the UC measure. We have seen in the contrast of *-ish* and *-ness* that productive suffixes can have different ways of being productive: one type, as in *-ness*, consistently leads to a large proportion of generally new words, while the other, as in *-ish*, leads to a large proportion of notably new words.

5.3 Intuition-Based Interpretation

The preceding evaluation of the P_{DE} measure has revealed that some results may deviate from our intuitive expectation: in particular, *-er* and *-ly* are assigned unexpectedly low productivity indices. Although *-er* and *-ly* had values of *V* (number of word types

with an affix) that were among the largest for the suffixes sampled, no significant correlation was found between V and P_{DE}. It is not the case that small P_{DE} is associated with large V. In the light of this, we will attempt in this section to pursue implications for the nature of intuitions about productivity.

When we say that P_{DE} indices are not wholly intuitive, for some suffixes, one possibility is that the data supporting P_{DE} are not in a form to be interpreted by intuition. For example, it is questionable whether speakers have in mind the kind of statistical information that the BNC offers. Speakers presumably cannot tell with any certainty how many *-er* and *-ee* word types there are in the BNC. However, they may be able to predict that there should be many more *-er* word types than *-ee* word types. Given the gap between precise statistical data from the BNC and non-precise speaker expectations, the exact values of V, V_N, and P_{DE} may not have direct relevance to speakers' intuitions about productivity.

There are three points to consider in search of an interpretation of the P_{DE} measure. The first is that, in asking whether corpus statistics match intuitions, we must make an assumption that intuitions are (in part) affected by performance data such as those represented in a corpus. To narrow down the question of what intuitions could be based on, we entertain the possibility that speakers, through their use of (or encounters with) words, form "impressions" about the productivity of word formation processes; when a description of productivity matches these impressions, the description is felt to be intuitive. Our task, then, is to investigate how those impressions could be obtained from performance data.

The second point to consider is the possibility that the type frequency information that speakers have access to may be better represented on a logarithmic scale. Word frequency effects have been extensively studied in psycholinguistics (see Monsell, 1991, for an overview). Howes and Solomon (1951) found that the visual duration threshold (the display time necessary for a subject to identify a visually presented word with some specified reliability) is inversely proportional to the logarithm of the token frequency of that word. It has also been shown (e.g., Bradley, 1978) that reaction time for a word in a lexical decision task is inversely proportional to the logarithm of its token frequency. Although word frequency effects are normally discussed with respect to token frequency, a possibility that we will entertain here is that similarly logarithmic scaling may also be applicable to type frequency. For instance, to the extent that speakers cannot estimate precisely how many word types with *-er* and *-ee* there are in a large language sample like the BNC, the raw values of *V* that we have been examining (e.g., 2517.8 for *-er*; and 88.6 for *-ee*) may not be directly relevant to intuitions. On the other hand, logarithmic transform values (base 10) for *V* (e.g., 3.40 for *-er*; and 1.95 for *-ee*) may better reflect the basis of speakers' intuition that more *-er* word types exist than *-ee* word types.

The third point to consider is Baayen's (1993: 204) view that speakers' intuitive judgments on productivity are ordinal in nature, rather than interval. Speakers may be unable to tell whether *-ness* should be twice or three times more productive than *-ity*, but they may be able to reject any productivity ranking of suffixes in which *-ness* falls at a lower rank than *-ity*. If intuitive judgments on productivity are no more than ordinal, we should focus only on the rank order of suffixes for a productivity measure.

Table 5-11 gives the logarithmic transform (base 10) for the values of V and V_N previously presented in Table 5-1. Values of V_{NN} are also given because they enter crucially into the exploratory formulation that will be introduced in the discussion that follows. Recall that the P_{DE} measure divides the total number of word types with an affix (V) into two sets: new word types with that affix (V_N), and the complement, non-new word types with that affix (V_{NN}).

Table 5-11. Values for V, V_N, and V_{NN}, with their logarithmic transform.

Suffix	V	$\log_{10}V$	V_N	$\log_{10}V_N$	V_{NN}	$\log_{10}V_{NN}$
-ness	1354.9	3.13	431.2	2.63	923.7	2.97
-ity	1008.5	3.00	234.4	2.37	774.1	2.89
-er	2517.8	3.40	558.6	2.75	1959.2	3.29
-ee	88.6	1.95	26.1	1.42	62.5	1.80
-ion	2152.9	3.33	348.7	2.54	1804.3	3.26
-ment	424.2	2.63	61.6	1.79	362.6	2.56
-th	40.9	1.61	3.5	0.54	37.4	1.57
-ize	437.6	2.64	114.5	2.06	323.0	2.51
-ify	105.8	2.02	21.1	1.32	84.7	1.93
-ish	261.3	2.42	90.6	1.96	170.7	2.23
-ous	639.1	2.81	107.1	2.03	532.0	2.73
-ly	3585.0	3.55	754.3	2.88	2830.7	3.45
ch#	213.6	2.33	29.7	1.47	183.9	2.26

There is an inherent problem with simply taking the ratio of $\log_{10}V_N$ to $\log_{10}V$, by analogy to V_N/V in the P_{DE} measure. While the order of suffixes obtained by V_N/V is always the inverse of the order obtained by V_{NN}/V, the same will not hold for log-transformed values. Although not an absolute requirement, we would certainly hope that measures

deriving from V_N and V_{NN} would be in complementary relationship. In particular, an order of suffixes obtained using V_N (however transformed) should have its mirror image in the order of suffixes obtained using V_{NN}.

A means to resolve this dilemma, and one that may be appropriate for exploring the impressions speakers have about word formation processes, is to ask to what extent word types with a suffix are new, as opposed to the extent that word types with the suffix are non-new (cf. what *proportion* of word types with a suffix is new). If we think of $\log_{10}V_N$ as expressing the *extent* to which words are new, and $\log_{10}V_{NN}$ as expressing the *extent* to which words are non-new, these conflicting factors may together influence speakers' overall impressions about the productivity of a given word formation process. As far as the transformed data in Table 5-11 are concerned, V_{NN} is greater than V_N in all suffixes, so the following statement can be made: when the extent to which words are new, $\log_{10}V_N$, approaches the extent to which words are non-new, $\log_{10}V_{NN}$, the suffix may be felt to be productive, with a degree that can be calculated by the ratio of $\log_{10}V_N$ to $\log_{10}V_{NN}$. Conversely, when the extent to which words are non-new, $\log_{10}V_{NN}$, surpasses the extent to which words are new, $\log_{10}V_N$, the suffix may be felt to be not productive, with a degree that can be calculated by the ratio of $\log_{10}V_{NN}$ to $\log_{10}V_N$. Although the latter ratio is currently not of interest, it establishes a complementary relationship between V_N and V_{NN}: the order of suffixes obtained by $\log_{10}V_N/\log_{10}V_{NN}$ is the reverse of the order obtained by $\log_{10}V_{NN}/\log_{10}V_N$.

Concentrating on the extent with which a suffix may be felt to be new, Table 5-12 shows a productivity ranking of suffixes based on the ratio of $\log_{10}V_N$ to $\log_{10}V_{NN}$.

Table 5-12. Productivity ranking based on $\log_{10}V_N$ to $\log_{10}V_{NN}$ ratio.

Suffix	$\log_{10}V_N/\log_{10}V_{NN}$
-ness	0.886
-ish	0.879
-er	0.836
-ly	0.835
-ize	0.821
-ity	0.820
-ee	0.789
-ion	0.779
-ous	0.744
-ment	0.699
-ify	0.684
ch#	0.650
-th	0.344

Taking the view that productivity judgments are strictly ordinal, the precise values seen in the second column of Table 5-12 can be ignored once the order of suffixes is obtained. Interestingly, we seem to gain many improvements in the productivity ranking as compared with that seen earlier in Figure 5-1 based directly on P_{DE}. We especially note the following: (a) *-ness* now counts as the most productive suffix; (b) *-er* and *-ly* move up in the ranking to be close to *-ness* and *-ish*; (c) *-ee* moves down in the ranking, but is still close to *-ity*; and (d) *-ify* is now much lower in the ranking. Perhaps one still unsatisfactory outcome is that *-ly* does not emerge as the most productive suffix. Given the high regularity in the word formation process of *-ly*, adverbial *-ly* seems to merit a higher status among the suffixes examined. We leave this puzzle for future study.

What has been attempted in this section is a speculation stemming from considerations of the kinds of information that could be available to speakers, when they make intuitive judgments about productivity. We have considered, in particular, the way that occurrence frequencies scale, psychologically. The fact that the productivity ranking of suffixes emerging in Table 5-12 is, for the most part, an intuitively appealing one suggests that further investigations along the lines sketched could be rewarding.

5.4 Summary

The P_{DE} measure and the UC measure were put to test in this chapter. Various analyses were given of the data collected by these measures. The productivity ranking of suffixes obtained by the P_{DE} measure was found to satisfy our expectations in many (if not all) respects. The P_{DE} measure was shown to provide productivity rankings that are stable over corpus segments of different sizes. We also examined differences in productivity between spoken and written components of the BNC. The UC measure was found to enhance our understanding of the characteristic differences in the way that productivity manifests itself among affixes. Overall, high correlations were found among implementations of the UC measure using different numbers of segments, and that finding was interpreted as evidence that the P_{DE} measure shares with the multiple-segment UC measure the same basic mechanisms for assessing productivity. In the light of some results that deviate slightly from intuitive expectations, a speculative interpretation of the data gathered under the P_{DE} measure has been offered, based on a psychologically motivated scaling of frequency data.

Chapter 6 Conclusion

The present research tackled the task of assessing the degree of productivity in word formation. Several productivity measures have already been proposed in the literature, but given the complexity of the issue of productivity, descriptions of its many different aspects need to be attempted. We pursued the corpus-based, computational approaches to quantifying productivity that were previously proposed by Baayen (1989, 1992, 2001) and his colleagues. Making use of corpus data because of its advantages over dictionary data in examining productively coined words, we focused particularly on the claim in the literature that the type frequency for an affix in a corpus is unsuited for assessing the degree of productivity.

Borrowing insights from a smoothing technique used in Language Technology for probability estimation, the deleted estimation method (Jelinek & Mercer, 1985), we proposed two methods of assessing productivity, the P_{DE} measure and the UC measure, and their performances have been evaluated based on the data of the BNC. The core concept adopted from the deleted estimation method is that of cross-comparing corpus segments. Both the P_{DE} and UC measures compare corpus segments to identify word types with a given affix that are shared and not shared. Word types not shared by corpus segments are considered new. Based on implementations involving just two corpus segments, the P_{DE} measure offers a simple method of identifying new words in a corpus and yields a productivity index for an affix, while the UC measure identifies the newness of words with an affix in a gradient manner based on how the use of words is distributed

among many corpus segments. Taken together, these measures offer previously unavailable insights about degrees of productivity in word formation. We showed that the proposed measures lead to characterizations of the productivity of affixes that are stable over different sample sizes.

Type frequency in a corpus was shown to be effective in assessing the productivity of affixes. The present data call for a reconsideration of the claim that type frequency is disadvantageous in assessing productivity. While it remains true that the type frequency for an affix in a corpus cannot be equated *per se* with its degree of productivity, we have shown that it is possible after all to assess productivity based on the way an affix's type frequency plays out in the distribution of its word types across corpus segments. In short, degrees of productivity can indeed be investigated without recourse to definitions turning on token frequency.

One of the advantages of a measure that is not based on token frequency lies in what words are identified as new. New words captured by the P_{DE} measure were found to differ from those captured by a productivity measure based on token frequency. The P_{DE} measure does not rule out a new word simply because the word is repeatedly used by the same speaker. Although the majority of new words in a corpus do occur only once, we found that a non-negligible number of new words (on average, one-third) occur multiple times, and the P_{DE} measure has the advantage of capturing those non-hapax words in addition to hapax occurrences.

A key proposal that the present study makes as a consequence of adopting the deleted estimation method concerns how corpus data should be analyzed. We stressed the importance of creating corpus segments so that each corpus segment samples a set of

words used by a group of speakers. This is possible when document boundaries are preserved, allowing corpus segments to be created through randomly sampling at the level of documents (cf. at the level of words). In this approach, crucially, the identification of new words ignores the repetition of any word by the same speaker. Consider the word *causee*. The fact that this word is repeated 9 times in 1 document of WBP makes this word count as non-new under any measure that targets words with low token frequency (e.g., hapaxes), but as new under the P_{DE} measure. However, the intuition is that an assessment of *causee* as new should not be ruled out because the term is used repeatedly by the coining author: it remains the type of word that we would like to consider new, and the P_{DE} measure is specifically formulated to express that intuition.

Since we take each document of the BNC as sampling a speaker, a concept that emerges from the present study is that of document frequency (a count of the number of documents in which a word appears), which plays an important role under the currently proposed measures. Document frequency differs from token frequency in that it represents how many speakers used a given word, rather than how often that word is used by an unknown number of speakers. In the case of hapaxes, we know that a word is used by only one speaker, but for non-hapaxes, token frequency cannot tell us how many speakers have contributed. Even when token frequency is high for a given word, we may still wish to capture that word as new if its multiple occurrences were due to no more than a few speakers. We showed with the BNC data that the situation imagined here is in fact surprisingly common, and thus that it is crucial to not disregard non-hapax forms. In a study of productivity for which the maximal capture of new words is desirable, it seems

more appropriate to consider how many speakers used a given word rather than how many times in total that word has been used.

The present study has shown that there are many types of data to be discovered in a corpus that are useful in assessing productivity. Although assessing frequency information for an entire corpus of texts may be the only approach to analysis given small corpora, assessing frequency information within and across corpus segments becomes viable when the corpus is as large as the BNC: corpus segments are themselves large enough to offer appropriately sized samples of words.

To summarize the general approach developed in the present research, we began by asking about how productivity might best be quantified, and in search of a method with appropriate sensitivity, noted that the identification of new words in a given data set has greater significance than obtaining productivity indices *per se*. The proper delimitation of new words in a corpus is of primary importance, and once new words are properly identified, obtaining productivity indices is a matter of finding the appropriate way to relate information about words so identified to one or other conceptualization of productivity. A productivity measure that does not identify new words properly runs the risk that its depiction of productivity is difficult to evaluate, regardless of how *a priori* plausible the productivity indices may be. In future corpus-based studies of productivity, effort should first be directed to refining the identification of new words, and then one way or other of obtaining productivity indices may be considered. In this sense, P_{DE} is one of many possible ways of expressing productivity. The major obstacle in the study of productivity lies in identifying new words, rather than in obtaining productivity indices.

The present research opens up several paths for future study. First, since we focused on English suffixes whose productivity has been widely discussed in the literature, the findings for productive affixes were relatively easy to assess. However, determining the baseline for *unproductivity* was left in a rather unsatisfactory position. We chose *ch#*, the word ending of a noun, as the baseline condition for determining productivity in affixation, but simply speculated that *ch#* should be unproductive. What is troublesome is that any productivity assessment using *ch#* is expected to carry noise arising from sources other than the processes of affixation, and we do not know to what extent such noise leads to mis-estimation of the (un)productivity of *ch#*. The weakness in the current finding is that we studied only word-ending *ch#* and suffix *-th* as presumably unproductive cases. To gain more insight into unproductivity in a word formation process, we need more examples likely to fall at the lower end of the productivity scale. This may be achieved by examining both a wider variety of word endings and a wider variety of affixes.

Second, a limitation remains on objectively evaluating a productivity measure. As long as the nature of intuitions about productivity, whether expert or naive, remains unexplored, there can be no external standard against which we might evaluate the validity of a proposed productivity measure. Nevertheless, as shown in Section 5.1.3, it is possible to some extent to assess how well or poorly a productivity measure identifies new words. In the present study, we experimented with the relationship between the way words fail to be listed in WD and the way the P_{DE} measure identifies those unlisted words as new. The problem with this verification method is that there is doubtless idiosyncrasy, both major and minor, about which words are not listed in WD. We must therefore first

devise a method for agreeing upon a set of new words and for testing how those new words are treated by a productivity measure. There may exist some objective method, better than dictionary lists, for settling upon a set of new words agreed to be new.

In relation to this same point, another possibility not explored in the present study would verify the graded definition of the newness of words that emerges from the non-categorical identifications of the UC measure, either in relation to newness defined in terms of words unlisted in WD, or in relation to whatever alternative devised to avoid lexicographic idiosyncrasies. We could, for example, examine the percentage of unlisted words that fall into segment frequency categories *S*1, *S*2, *S*3, and so on. The necessary prediction would be that the largest percentage of unlisted words falls into *S*1.

Third, the graded definition of word newness provided by the UC measure could be explored more closely in future studies. While the P_{DE} measure identifies new words categorically for the sake of computational simplicity, the UC measure's graded definition of new words accords with the fact that any list of new words may not be equally new to all speakers. In the present study, the UC measure served to clarify the design of the P_{DE} measure, but the idea of graded newness is of interest by itself, and merits future exploration. From a computational viewpoint, the non-categorical definition of new words seems to be an area that can usefully be pursued. Certainly the availability of sophisticated corpora of larger size opens possibilities for making yet finer distinctions among new words: under the UC measure, the number of corpus segments (which was currently limited to 6, given the size of the BNC corpus) can be increased while maintaining a large number of words in each corpus segment, given a sufficiently large corpus. Non-categorical identifications can be used not only to discover the newness of

words but also the non-newness of words; for example, the UC measure has a potential use in determining how established any given word type is.

Fourth, we have confined our findings to corpus data, but there are some proposals of the present research that could be tested in experiments with human subjects. Intuitions about productivity are of particular interest. Although the source and nature of such intuitions is currently unknown, it may be possible to inquire into the nature of intuitions by examining apparently intuitive and counterintuitive results for different productivity measures. For example, English speakers could be presented with alternative productivity rankings, say for particular affix pairs, and those rankings could be judged as acceptable or not. An experiment of this kind, implemented in a protocol that was appropriate for naive judges, may be revealing about which affix pairs (with what differing degrees of productivity) support stable intuitions in speakers.

In addition, the fact that the data are confined to a corpus of texts that represents the use rather than the perception of words means that it is yet to be shown whether new words identified on the basis of word production correspond to words that are largely unfamiliar to speakers (i.e., not previously encountered). Ultimately, we would like to associate new words with those not previously encountered by the majority of speakers, and one may be able to devise experimental procedures to test the link between production and perception of words in terms of their newness.

Finally, by taking an approach based on type frequency, the psycholinguistic motivation of the measures proposed in the current research has become obscure in terms of their relation to issues such as morphological parsing. One interesting point to note is that the deleted estimation method has its origin in probability estimations turning on

token rather than type frequency. This means that we leave open the possibility that the deleted estimation method could be incorporated into an entirely different kind of productivity measure, one that exploits token frequency. Such a measure would still enjoy the benefits of the cross-comparison of corpus segments in identifying new words so that new words would not necessarily be limited to hapaxes. A comparison between any such measure and the currently proposed measures (exploiting type frequency) would offer further insight into the contribution that the deleted estimation approach makes to the study of productivity.

It is hoped that the methods of assessing productivity that have been proposed in the present research will contribute to the growing literature investigating the issue of productivity in word formation. The methods proposed for capturing new words in corpus data are particular instances of many possible methods that could be devised, given the wealth of information that a corpus of texts provides. The success of the present study shows that Language Technology may suggest innovative techniques for the computational study of productivity. It is also to be hoped that the proposals made (especially those concerning the use of type frequency in assessing productivity) and the examination of corpus-segment data (as opposed to corpus data taken whole) suggest a new direction along which corpus data can be exploited. Given the complexity of the issue of productivity, it seems that a full elucidation of this central—and likely, multi-faceted—aspects of word formation can best be achieved by a synthesis of many different approaches.

References

Academia Sinica Balanced Corpus (Version 3.0) [CD-ROM]. (1998). Taipei, Taiwan: Academia Sinica.

Anderson, S. R. (1982). Where's morphology? *Linguistic Inquiry, 13,* 571–612.

Anshen, F., & Aronoff, M. (1981). Morphological productivity and phonological transparency. *Canadian Journal of Linguistics, 26,* 63–72.

Anshen, F., & Aronoff, M. (1988). Producing morphologically complex words. *Linguistics, 26,* 641–655.

Anshen, F., & Aronoff, M. (1989). Morphological productivity, word frequency and the Oxford English Dictionary. In R. W. Fasold & D. Schiffrin (Eds.), *Language Change and Variation* (pp. 197–202). Amsterdam: John Benjamins.

Archangeli, D., & Langendoen, T. (Eds.). (1997). *Optimality Theory: An Overview.* Oxford, UK: Blackwell.

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

Aronoff, M. (1980). The relevance of productivity in a synchronic description of word formation. In J. Fisiak (Ed.), *Historical Morphology* (pp. 71–82). The Hague: Mouton.

Aronoff, M. (1983). Potential words, actual words, productivity and frequency. *Proceedings of the International Congress of Linguists, 13,* 163–171.

Aronoff, M. (1994). *Morphology by Itself: Stems and Inflectional Classes.* Cambridge, MA: MIT Press.

Aronoff, M., & Anshen, F. (1998). Morphology and the lexicon: Lexicalization and productivity. In A. Spencer & A. M. Zwicky (Eds.), *The Handbook of Morphology* (pp. 237–247). Oxford, UK: Blackwell Publishers.

Aronoff, M., & Fuhrhop, N. (2002). Restricting suffix combinations in German and English: Closing suffix and the monosuffix constraint. *Natural Language & Linguistic Theory, 20* (3), 451–490.

Aronoff, M., & Schvaneveldt, R. (1978). Testing morphological productivity. *Annals of the New York Academy of Sciences, 318,* 106–114.

Ayto, J. (1990). *Dictionary of Word Origins.* New York: Arcade Publishing.

Baayen, R. H. (1989). *A Corpus-Based Study of Morphological Productivity: Statistical Analysis and Psychological Interpretation.* Doctoral dissertation, Free University, Amsterdam.

Baayen, R. H. (1992). Quantitative aspects of morphological productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1991* (pp. 109–149). Dordrecht: Kluwer.

Baayen, R. H. (1993). On frequency, transparency and productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1992* (pp. 181–208). Dordrecht: Kluwer.

Baayen, R. H. (1994). Productivity in language production. *Language and Cognitive Processes, 9* (3), 447–469.

Baayen, R. H. (2001). *Word Frequency Distributions.* Dordrecht: Kluwer.

Baayen, R. H. (2003). Probabilistic approaches to morphology. In R. Bod, J. Hay, & S. Jennedy (Eds.), *Probabilistic Linguistics* (pp. 229–287). Cambridge, MA: MIT Press.

Baayen, R. H., & Lieber, R. (1991). Productivity and English word-formation: A corpus-based study. *Linguistics, 29,* 801–843.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database* [CD-ROM]. Philadelphia, PA: LDC, University of Pennsylvania.

Baayen, R. H., & Renouf, A. (1996). Chronicling the Times: Productive lexical innovations in an English newspaper. *Language, 72,* 69–96.

Barker, C. (1998). Episodic *-ee* in English: A thematic role constraint on new word formation. *Language, 74,* 695–727.

Bauer, L. (1983). *English Word-Formation.* Cambridge, UK: Cambridge University Press.

Bauer, L. (1988). *Introducing Linguistic Morphology.* New York: Columbia University Press.

Bauer, L. (1992). Scalar productivity and *-lily* adverbs. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1991* (pp. 185–191). Dordrecht: Kluwer.

Bauer, L. (1998a). Is there a class of neoclassical compounds and is it productive? *Linguistics, 36,* 403–422.

Bauer, L. (1998b). When is a sequence of two nouns a compound in English? *English Language and Linguistics, 2,* 65–86.

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

Baxter, W. H., & Sagart, L. (1998). Word formation in Old Chinese. In J. L. Packard (Ed.), *New Approaches to Chinese Word Formation: Morphology, Phonology and Lexicon in Modern and Ancient Chinese* (pp. 35–76). Berlin: Mouton de Gruyter.

Bertram, R., Schreuder, R., & Baayen, R. H. (2000). The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology, 26,* 489–511.

Biber, D. (1989). A typology of English texts. *Linguistics, 27,* 3–43.

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics, 19* (2), 219–241.

Bolozky, S. (1999). *Measuring Productivity in Word Formation: The Case of Israeli Hebrew.* Leiden: Brill.

Booij, G. E. (1977). *Dutch Morphology: A Study of Word Formation in Generative Grammar.* Amsterdam: Foris.

Bradley, D. C. (1978). *Computational Distinctions of Vocabulary Type.* Doctoral dissertation, MIT.

British National Corpus (World Edition) [CD-ROM]. (2000). Oxford, UK: Oxford University Computing Services.

Burnard, L. (2000). *Reference Guide for the British National Corpus (World Edition).* Oxford, UK: Oxford University Computing Services

Bybee, J. L. (1998). The emergent lexicon. *Chicago Linguistic Society, 34,* 421–435.

Carstairs-McCarthy, A. (2002). *An Introduction to English Morphology.* Edinburgh, UK: Edinburgh University Press.

Chantrell, G. (Ed.). (2002). *The Oxford Dictionary of Word Histories.* Oxford, UK: Oxford University Press.

Chen, P. (1999). *Modern Chinese: History and Sociolinguistics.* Cambridge, UK: Cambridge University Press.

Chen, S. F., & Goodman, J. (1998). *An Empirical Study of Smoothing Techniques for Language Modeling* (Tech. Rep. No. 10-98). Cambridge, MA: Center for Research in Computing Technology, Harvard University.

Chomsky, N. (1970). Remarks on nominalization. In J. A. Roderick & P. S. Rosenbaum (Eds.), *Readings in English Transformational Grammar* (pp. 184–221). Waltham, MA: Ginn.

Church, K. W. (2000). Empirical estimates of adaptation: The chance of two Noriegas is closer to $p/2$ than p^2. *Proceedings of the 18th International Conference on Computational Linguistics* (pp. 173–179). Saarbrücken, Germany.

Church, K. W., & Gale, W. A. (1991). A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language, 5,* 19–54.

Concise Oxford Dictionary of Current English (7th ed.). (1982). Oxford, UK: Oxford University Press.

Cutler, A. (1980). Productivity in word formation. *Chicago Linguistic Society, 16,* 45–51.

Cutler, A. (1981). Degrees of transparency in word formation. *Canadian Journal of Linguistics, 26,* 73–77.

Dict.org. (2003). The DICT Development Group. <http://www.dict.org/>.

Evert, S., & Lüdeling, A. (2001). Measuring morphological productivity: Is automatic preprocessing sufficient? *Proceedings of the Corpus Linguistics 2001 Conference* (pp. 167–175). Lancaster, UK.

Fabb, N. (1998). Compounding. In A. Spencer & A. M. Zwicky (Eds.), *The Handbook of Morphology* (pp. 66–83). Oxford, UK: Blackwell Publishers.

Frauenfelder, U. H., & Schreuder, R. (1992). Constraining psycholinguistic models of morphological processing and representation: The role of productivity. In G. E. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1991* (pp. 165–183). Dordrecht: Kluwer.

Gale, W. A. (1994). *Good-Turing Smoothing Without Tears.* (Statistical Research Reports No. 94.5). Florham Park, NJ: AT&T Labs Research.

Garside, R., Leech, G., & McEnery, T. (Eds.). (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London: Longman.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika, 40,* 237–264.

Google. (2004). Google, Inc. <http://www.google.com/>.

Guo, J. (1993). PH: A Chinese corpus. *Communications of COLIPS, 3* (1), 45–48.

Haspelmath, M. (1996). Word-class-changing inflection and morphological theory. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1995* (pp. 43–66). Dordrecht: Kluwer.

Haspelmath, M. (2002). *Understanding Morphology.* London: Arnold.

Hay, J., & Baayen, R. H. (2002). Parsing and productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 2001* (pp. 203–235). Dordrecht: Kluwer.

Hockenmaier, J., & Brew, C. (1998). Error-driven learning of Chinese word segmentation. In J. Guo, K. T. Lua, & J. Xu (Eds.), *12th Pacific Conference on Language and Information* (pp. 218–229). Singapore: Chinese and Oriental Languages Processing Society.

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology, 41,* 401–410.

Jelinek, F., & Mercer, R. (1985). Probability distribution estimation for sparse data. *IBM Technical Disclosure Bulletin, 28,* 2591–2594.

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River, NJ: Prentice Hall.

Katz, S. M. (1987). Estimates of probabilities from sparse data for the language model component of a speech recogniser. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 35* (3), 400–401.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to special issue on the Web as corpus. *Computational Linguistics, 29* (3), 333–347.

Kiparsky, P. (1982). Lexical morphology and phonology. In I. S. Yang (Ed.), *Linguistics in the Morning Calm* (pp. 3–91). Hanshin, Seoul.

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

Leech, G., & Smith, N. (2000). *Manual to Accompany the British National Corpus (Version 2) with Improved Word-Class Tagging.* Lancaster, UK: UCREL, Lancaster University.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus.* Harlow: Longman.

Li, C., & Thompson, S. A. (1981). *Mandarin Chinese: A Functional Reference Grammar.* Berkeley, CA: University of California Press.

Lin, H. (2001). *A Grammar of Modern Chinese.* München: Lincom Europa.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition, 40,* 121–157.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press.

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

Matthews, P. (1991). *Morphology* (2nd ed.). Cambridge, UK: Cambridge University Press.

Meys, W. J. (1975). *Compound Adjectives in English and the Ideal Speaker-Listener.* Amsterdam: North Holland.

Monsell, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner & G. W. Humphreys (Eds.), *Basic Processes in Reading: Visual Word Recognition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Muthmann, G. (1999). *Reverse English Dictionary: Based on Phonological and Morphological Principles.* Berlin: Mouton de Gruyter.

Nishimoto, E. (2003). Measuring and comparing the productivity of Mandarin Chinese suffixes. *Journal of Computational Linguistics and Chinese Language Processing, 8* (1), 49–76.

Norman, J. (1988). *Chinese.* Cambridge, UK: Cambridge University Press.

Oxford English Dictionary (2nd ed., Version 3.0) [CD-ROM]. (2002). Oxford, UK: Oxford University Press.

Packard, J. L. (2000). *The Morphology of Chinese: A Linguistic and Cognitive Approach.* Cambridge, UK: Cambridge University Press.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28,* 73–193.

Plag, I. (1998). The polysemy of *-ize* derivatives: The role of semantics in word formation. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1997* (219–242). Dordrecht: Kluwer.

Plag, I. (1999). *Morphological Productivity: Structural Constraints in English Derivation.* Berlin: Mouton de Gruyter.

Plag, I. (2003). *Word-Formation in English.* Cambridge, UK: Cambridge University Press.

Plag, I., Dalton-Puffer, C., & Baayen, R. H. (1999). Morphological productivity across speech and writing. *English Language and Linguistics, 3* (2), 209–228.

Ramsey, R. S. (1987). *The Languages of China.* Princeton, NJ: Princeton University Press.

Renouf, A. (1987). Corpus development. In J. Sinclair (Ed.), *Looking Up: An Account of the Cobuild Project in Lexical Computing* (pp. 1–40). London: Collins ELT.

Renouf, A. (1993). A word in time: First findings from the investigation of dynamic text. In J. Aparts, P. de Haan, & N. Oostdijk (Eds.), *English Language Corpora: Design, Analysis and Exploitation* (pp. 279–288). Amsterdam: Rodopi.

Ryder, M. E. (1999). Bankers and blue-chippers: An account of *-er* forms in present-day English. *English Language and Linguistics, 3,* 269–297.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tense of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 2: Psychological and Biological Models* (pp. 216–271). MIT Press.

Scalise, S. (1984). *Generative Morphology.* Dordrecht: Foris.

Schreuder, R., & Baayen, R. H. (1995). Modeling morphological processing. In L. B. Feldman (Ed.), *Morphological Aspects of Language Processing* (pp. 131–154). Hillsdale, NJ: Lawrence Erlbaum.

Schultink, H. (1961). Produktiviteit als morfologisch fenomeen. *Forum der Letteren, 2,* 110–125.

Shorter Oxford English Dictionary (Version 2.0) [CD-ROM]. (2002). Oxford, UK: Oxford University Press.

Siegel, D. (1979). *Topics in English Morphology.* New York: Garland.

Spencer, A. (1988). Bracketing paradoxes and the English lexicon. *Language, 64,* 63–82.

Spencer, A. (1991). *Morphological Theory: An Introduction to Word Structure in Generative Grammar.* Cambridge, UK: Cambridge University Press.

Sproat, R. (1988). Bracketing paradoxes, cliticization and other topics: The mapping between syntactic and phonological structure. In M. Ereraert, A. Evers, R. Huybregts, & M. Trommelen (Eds.), *Morphology and Modularity* (pp. 339–360). Dordrecht: Foris.

Sproat, R., & Shih, C. (1996). A corpus-based analysis of Mandarin nominal root compound. *Journal of East Asian Linguistics, 5,* 49–71.

Sproat, R., Shih, C., Gale, W., & Chang, N. (1996). A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics, 22* (3), 66–73.

Van Marle, J. (1985). *On the Paradigmatic Dimension of Morphological Productivity.* Dordrecht: Foris.

Van Marle, J. (1992). The relationship between morphological productivity and frequency: A comment on Baayen's performance-oriented conception of morphological productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1991* (pp. 151–163). Dordrecht: Kluwer.

Walker, J. (1936). *Walker's Rhyming Dictionary.* New York: Dutton.

Webster's Third New International Dictionary, Unabridged (Version 2.5) [CD-ROM]. (2000). Springfield, MA: Merriam-Webster.

Webster's Third New International Dictionary, Unabridged (the 1913 ed.) [Electronic Data File]. (1999). Oxford, UK: Project Gutenberg.

Wheeler, C. J., & Schumsky, D. A. (1980). The morpheme boundaries of some English derivational suffixes. *Glossa, 14,* 3–34.

Witten, I. H., & Bell, T. C. (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory, 37,* 1085–1094.

WordNet (Version 1.7.1) [Software]. (2003). Princeton, NJ: Princeton University.