Relative Use of Phonaesthemes in the Constitution and Development of Genres

James Christopher Harbeck

A thesis submitted to the Faculty of Graduate Studies in partial fulfillment of the requirements for the degree of

Master of Arts

Graduate Program in **Linguistics**,

York University, Toronto, Ontario, March 2016

Abstract

My research question is “Does the presence of phonaesthemes in words play a role in the constitution and evolution of genres?” A phonaestheme is a phonemic grouping that correlates well above chance with a particular semantic quality in etymologically unrelated words; phonaesthematic words are generally seen as vivid, expressive, and involved. I explore the nature of phonaesthemes and genres and the role of features such as phonaesthemes in the constitution of genres. I select a set of phonaesthemes to evaluate and choose a representative set of lemmas and matching non-phonaesthematic lemmas. I survey these in six genres over three time periods in the US and the UK. I analyze the results and their implications for phonaesthemes and for genre constitution, finding, among other things, that phonaesthemes are important in the social positioning of genres. The summary answer to my research question is thus found to be “Yes, it does.”

Acknowledgements

I would like to thank the members of my committee, most notably my adviser, James Walker; I would also like to thank Peter Avery for supervising an earlier research effort that led to this thesis, and for giving me guidance and a warm welcome in the linguistics department at York; I would like to thank the several members of Drinkquistic Inquiry, led by Max Baru, for keeping me sharp; I would like to thank Jennie Worden for giving me the idea of taking courses in linguistics; I would like to thank various other friends as well, and I will, in person, with drinks; I would like to thank my employer for accommodating my academic schedule throughout my academic career in linguistics; and most of all I would like to thank my wife, who has had to put up with all this.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction and background

My research question is **"Does the presence of phonaesthemes in words play a role in the constitution and evolution of genres?"** The first order of business in this is therefore to establish what phonaesthemes are (and aren't), and to define and consider the nature of genres and their constitution and evolution. This is the task of Chapter 1: to define the terms, give important background information, and establish the validity of the research project. In Chapter 2, I will describe the means and materials I used to address the question; this will also give useful supporting information. In Chapter 3, I will present and discuss the results of the research project. Chapter 4 will give a brief high-line summation. I am also including appendices with full data sets.

1.1 Phonaesthetics and phonaesthemes

A basic standard view in linguistics is that words are composed entirely of morphemes, which are the smallest meaning-bearing units of language, and that the association between sound and meaning is arbitrary. Onomatopoeia is seen as a circumscribed exception. In opposition to this view is a current that studies iconic and indexical relationships between sound and meaning and discerns variously reliable relationships between certain sounds or sets of sounds and the meanings of the words they are found in. These sound patterns exist on a stratum overlapping the morphological, and may be only sometimes present – and sometimes ambiguously so. Some authors have sought to discern a meaning value in every single sound; others are more reserved, picking out clusters that have gained a certain attractive force and leaving the rest to be assumed as arbitrary. Recent overview articles such as Perniss et al. (2010) and Dingemanse et al. (2015) have made strong cases for iconicity being a foundational part of language, though it is

not equally present in all aspects or all lexical items. We draw from these studies the clear understanding that speakers tend to associate at least some kinds or clusters of sounds with certain kinds of meaning regardless of etymology, at least in appropriate contexts. As well, many authors discern in these usages a particular expressivity – a greater concreteness, a more vivid imagery, even in those cases that could not be called imitative.

This should not seem so unreasonable. We are not machines, after all, and we're not especially logical. We learn word parts by correlating, observing patterns, and abducing. As Bolinger (1950) pointed out, etymology is of no special relevance because ordinary naïve speakers' judgements are crucial in determining what word parts are morphemes and what they mean. As Rhodes and Lawler (1981, 22) put it, "etymology is a weak reed to lean on in semantics" – especially when the etymology of many common words is uncertain not only to the average user but to lexicographers as well. In the ordinary speaker's perception, a morpheme is best characterized in the view of the usage-based model, as described by Bergen (2004, 306): "a morpheme is at its core a recurrent sound-meaning correlation across the lexicon" (a definition that, with only slight changes, could be used for a phonaestheme, as we will see). This can be influenced by unrelated but resemblant words; for example, Bolinger (1950) observed that *ambush* tends to make people think of someone hiding in the bushes, and *hierarchy* brings to mind *higher*. This kind of association is the basis for the commonly observed process of folk etymology. Resemblance shapes our expectations regarding words and can affect their forms; Bolinger (1968, 24–25) notes that the spelling *miniscule* is overtaking *minuscule* by analogy with *mini*.

Rhodes and Lawler (1981) present a detailed theory of athematic metaphorical effect by association through common onsets and rimes – effectively a thoroughgoing theory of phonaesthemes, though not a universal overriding one:

> We do **not** claim that everyone in the speech community always uses these words in the way we suggest. We **do** claim that these associative meanings

> are available as guides for interpretation (to supplement textual convergence) of words encountered by a speaker (particularly a child) for the first time. The fact that a language learner may supplant initial hypotheses with more sophisticated understandings as his competence increases in no way implies that his initial understandings (and the general strategies that produce them) thereby become inaccessible. We wish to suggest that the assonance-rime interpretation strategies are far more common (and far more productive) than they have generally been given credit for. (Rhodes and Lawler 1981, 22; emphasis in the original)

Magnus (2001, 6) articulates a principle of "phonosemantic association": "When semantic domain S is associated disproportionately frequently with phoneme X, then people will be inclined to associate semantic domain S with phoneme X productively." As an example, "If a fundamental word like 'house' in a given language begins with an /h/, then Phonosemantic Association will cause words with similar sound and meaning to cluster to it" (7). This is a bit of a strong statement, but if we were to put *may* in place of *will* it would not be so unreasonable. More reasonable is the specific and qualified result Magnus gets from a survey of the lexicon: "Monosyllabic words in English which contain a given consonant fall within much narrower semantic domains than one would expect if the relationship between phonology and semantics were arbitrary" (2001, 76). Add to this what Nuckolls (1999, 226) observes: linguistic sounds "express our emotional states, aesthetic apperceptions, and the alignments and interrelations we have with other members of our social world, none of which can be neatly separated from denotational reference." Not just denotations but attitudes, emotions, and performative qualities may tend to associate with sounds.

The objection may be raised (and has been, for example by Bolinger (1968)) that if sound symbolism were an important force in language, we would be able to guess the meanings of words in foreign languages. To this we can make two immediate responses: first, as we will see, in some instances – particularly but not exclusively

of "expressive" words – the rate of accurate guesses, while not 100%, is well above chance; second, Chinese characters are all based on iconic representations, and in many cases those representations are still perceptible if you know what you're looking for, but none or almost none of them are perspicuous to the naïve reader. A clear and certain photographic iconicity is not in the question here. The associations we are looking at are not tidy, Boolean, and compositional, but they are statistically significant. We are dealing here not with certainty but with probability. The associations are clearly guided by cultural preferences and existing forms – consider the particular stylings of Chinese calligraphy, or the existing onomatopoeia and other words that make *pan* a reasonable imitation of a gunshot in French but not so much in English – but are also guided by the means available (calligraphy brushes and paper; the phonemes and syllabification rules of a language) and, of course, by features of what is being represented (for example, *moo* is not likely a plausible imitation of a gunshot in any language).

On the other hand, we cannot simply erase compositionality and arbitrariness from language processes; they are important, even if not the whole story. Householder (1946) posited a tripartite division of English vocabulary: items that have an entirely arbitrary relation to their sounds; items made up partially or entirely of phonaesthemes; and items with arbitrary relations but with their meanings affected to some extent by association with phonaesthemes. If we leave out the first division (to the extent that it truly exists – I will assume it does, although we will see that quite a few authors believe it does not), we can consider several overlapping categories covering the range from direct imitation with minimal involvement in linguistic processes to highly standardized and conventionalized, and more morphosyntactically integrated, but still expressive word parts; we can call these categories sound symbolism, onomatopoeia, ideophones, and phonaesthemes. I will give a brief overview of each of the first three kinds before proceeding to a fuller review of literature on phonaesthemes.

1.1.1 Sound symbolism

In Peircean terms, sound symbolism is the use of sounds in language for their iconic value – not necessarily as direct imitations of other sounds, but at least with some direct connection by resemblance rather than through purely conventional association (symbols) or as evidence of what they refer to (indexes). "Sound symbolism" is quite a broad term and in some senses can be seen as covering onomatopoeia, ideophones, and phonaesthemes; more narrowly, it refers to using linguistic sounds for expressivity. As Nuckolls (1999, 228) says, "The term sound symbolism is used when a sound unit such as a phoneme, syllable, feature, or tone is said to go beyond its linguistic function as a contrastive, non-meaning-bearing unit, to directly express some kind of meaning."

The most abundant evidence, as Nuckolls (1999, 230) tells us, is for diminutive symbolism associated with high front vowels; researchers from Jespersen (1922) and Sapir (1911; 1929) to the present have found associations with smallness, brightness, lightness, quickness, height, nearness, and intimacy. This is not a universal; a few striking exceptions have been identified – for example, Diffloth (1994) looks at Bahnar, a Mon-Khmer language in which central and high vowels are associated with largeness and low vowels are associated with smallness. But, exceptions notwithstanding, the association is attested widely across languages. Ohala (1994, 343) refers to the *frequency code*, something documented across not only languages but even species (among mammals):

> high F_0 signifies (broadly) smallness, non-threatening attitude, desire for the goodwill of the receiver, etc., and low F_0 conveys largeness, threat, self-confidence, and self-sufficiency.

While the frequency code as Ohala identifies it is based on fundamental frequency (F_0), vowel frontness (second formant, F_2) has been found to correlate to F_0 in sound symbolism. For example, Geenberg (2010) found that in tasks where adults were to simulate "baby talk" with a stuffed animal doll, positive "cute baby talk" had

significantly higher F_2 and often higher F_0 and lower F_1, while sad or consoling "poor baby talk" had the opposite.

Other contrasts in sound symbolism involve such things as consonant qualities (notably resonance, sustain, and abruptness). For example, Usnadze (1924) and Köhler (1930) did experiments presenting subjects with two non-representative line drawings, one curvy and one angular, and two invented words, *maluma* and *takete*. The subjects were asked to match the words to the drawings; by a clear margin, *maluma* was matched to the rounded figure and *takete* to the angular one. The result has been reproduced in the intervening decades in several languages for various age groups from early childhood to adult by various researchers using the same or similar stimuli, most recently including Ramachandran and Hubbard (2001), Maurer, Pathman, and Mondloch (2006), and Nielsen and Rendall (2011). This does not translate into a clear and overriding pattern in the full natural language, however: Monaghan et al. (2012) found that, *ceteris paribus*, the only significant differences in sound patterns in English between words with round and angular referents was a greater tendency for words with angular referents to have velar sounds and voiceless sounds.

With the different sound contrasts come different semantic contrasts, not all of which are – or even can be – represented with direct sonic iconicity. Abelin (1999, 60) identifies a limited but fairly sizeable set of semantic categories that are represented by sound symbolism in Swedish: "Sound, Movement, Light, Surface structure, Consistency (plasticity), Wetness, Dryness, Attitude, Slang, Jocular, Pejorative, Mental feeling, Bodily feeling, Separation, Putting together (convergence), Diminutive, Augmentative, Form, Iterative." These are not all directly related to sensory inputs, which indicates that sound symbolism depends not only on learning from one's senses but on "innate capacities for metaphor" (Abelin 1999, 64). Kaufman (1994) finds that Huastec sound-symbolic roots refer especially to light and to moving forms.

But how, exactly, *do* we connect the sensory inputs with the sound-symbolic associations? What makes high and front equal small for most languages? There are various ideas. It has long been tempting for many to associate it directly with synaesthesia, the sensory crossover experienced literally by some people. The problem with this, as Cytowic (1989) has shown, is that for actual synaesthetes the associations are highly idiosyncratic and individual. If we are to posit some quasi-universal sound-sense connection – or to explain one that has been observed – we cannot appeal to simple inherent neurological association or crosstalk. We need to find a proper causal chain.

As mentioned, it is common to connect the F_2 of high front vowels with the general significance of F_0, which is also associated with smallness and lightness – for example, in some African languages, the same ideophone (see below) will mean a small or pleasant version of something when said on a high tone, but a large or unpleasant version when said on a low tone (Ameka 2001, 30). As Nuckolls (1999, 229) notes, "The increase in tension used for higher pitch and the decrease in tension used for falling pitch have a universal tendency to be associated, metaphorically, with the contrasting ideas of incompleteness (high pitch) and completeness (falling or low pitch)." This may have a connection to human sexual dimorphism; Ohala (1994, 337) points out that the size difference between men and women would by itself predict a much smaller difference in size and pitch between male and female vocal apparatuses. We can also without difficulty connect it to the difference in physical size and vocal pitch between children and adults. The significant lowering of the male voice occurs at puberty, at the same time as another display feature, facial hair; on this basis, Ohala posits (342) that "the enlargement of the vocal apparatus occurs to enhance the acoustic component of aggressive displays." At the very least, it is seen that higher F_2 is associated with smaller and "nicer" whereas lower F_2 is associated with larger and "meaner" or more negative, possibly in connection with the concurrent changes in body size and personal behaviour – self-presentation and outlook – in adolescence (see Eckert 2010). In a study of two early adolescents, Eckert found backing of /o/ and /aɪ/ "to show a

broad range of negative feelings including fear, sadness, annoyance, victimization and so on" (2010, 79).

Ohala (1999) connects the frequency code with confrontations among animals, wherein the submissive typically emits high-pitched yelps and the dominant emits low growls, and even with features of the human smile: "I propose . . . that the smile – and its opposite, the 'o-face' – originally served as a component of the acoustic element of these [threat] displays. In the smile the mouth corners are drawn back to effectively shorten the vocal tract and to give rise to higher resonant frequencies." (101) Ohala's explanation of the smile on the basis of its acoustic effects is novel, however. Others such as Abelin (1999, 39) are more inclined to see the smile as primary and the association of higher F_2 with the smile as an effect of that. Some such as Bolinger (1968) connect the size of the mouth cavity more directly with the meaning, at least for size symbolism. Others such as Perniss et al. (2010, 6) and Ramachandran and Hubbard (2001) propose similar "cross-modal mapping" such as mapping round mouth to rounded objects. Traunmüller (1996) found support for proprioceptive effects in first- and second-person pronouns cross-linguistically in 37 etymologically unrelated cases: even accounting for areal effects, a statistically significant set had "oral closure and sustained voicing" in the first-person pronoun and "oral pressure build-up and explosion" in the second-person pronoun (148), consistent with a hypothesis that the first-person pronouns tend to prefer sounds that seem to stay within the head, whereas the second-person pronouns tend to prefer sounds that seem to be projected away from the head. In a similar vein, Jespersen (1922, 396) cites the Roman Publius Nigidius Figulus (98–45 BCE) as saying that *vos* ('you') puts the lips forward towards the other person, while *nos* ('we') does not. This is of course a naïve and facile post-facto explanation, but it does remind us that there may be gestural components – deictic as well as iconic – that could be worth studying. The articulatory gesture is an essential part of the act. As Austerlitz (1994, 255) says,

> too little attention seems to be paid to the obvious fact that very little children not only hear what is being said to them but also watch and see what facial gestures the speaker is making while speaking. The child – even the very small child – then imitates and thus associates muscular events with acoustic-articulatory ones.

There is also feedback from the visual to the auditory, as is demonstrated by the well-known McGurk effect (McGurk and MacDonald 1976).

The explanation of sound-symbolic associations is therefore still quite a fertile field. We will also see that the waters can be muddied somewhat further by conventionalizing and abstracting forces as are commonly present in language.

1.1.2 Onomatopoeia

Onomatopoeia is the linguistic rendering of non-speech (and usually non-human) sounds. It is therefore imitative. But it is well known that onomatopoeic representations of sounds (e.g., animal noises) vary somewhat from language to language. As Marchand (1959, 151–152) points out,

> The imitative principle is often misunderstood or misrepresented. It is commonly thought that an onomatopoeia should be the exact rendering of the corresponding noise. The explanations as to the difference between languages is that "our speech organs are not capable of giving a perfect imitation of all 'unarticulate' sounds" and that therefore "the choice of speech sounds is to a certain extent accidental." This is, of course, right, but only partly. It overlooks the fact that an onomatopoeia is not a mere imitation of a sound.

I have addressed this above in my introductory remarks as well, in response to Bolinger's (1968) objection.

It may be useful to follow Rhodes (1994, 279), who posits a spectrum of onomatopoeia ranging from "wild" to "tame." "At the extreme wild end," he says,

"the possibilities of the human vocal tract are utilized to their fullest to imitate sounds of other than human origin. At the tame end the imitated sound is simply approximated by an acoustically close phoneme or phonemic combination." Along with adhering to established phonemes, onomatopoeia can also often adhere to established patterns and combinations: *clink, clank, clunk, plink, plunk, ping, bing, bang*, et cetera give one example of this; *patter, splatter, mutter, rattle, prattle,* et cetera give another. Thus, onomatopoeia makes use of sound symbolism but tends towards conventionalization. It can be thought of as a kind of conventionalized sound symbolism, but, as we have seen, sound symbolism also includes referents that do not make sounds, and it is also in its realizations in a given language conditioned at least by the available phonemes and perhaps also at least to some degree by expectation. We may more precisely say that onomatopoeia is a directly performative and ostensibly representative usage of the sound-symbolic function using the means available in a language. It is important to note, however, that onomatopoeic representation is not a one-way street. While, as Abelin (1999, 14) says, "An onomatopoeic word is constrained by the sound it imitates," it is surely also the case that our perception of natural sounds is conditioned by expectation: existing onomatopoeias, known words, and the overall phonological patterns of our language will condition us to expect some patterns much more than others. Categorical perception can apply to more than just human speech, and memory is subject to the influence of what is known and believed.

Notwithstanding this, onomatopoeia can be seen to form a thin hot current in the cool river of our language: as Abelin (1999, 14) notes, "there seems to be a general agreement that onomatopoeic (and also sound symbolic) words usually do not undergo the same phonological changes as other words, e.g. according to Grimm's law, as long as they still have a sound imitative meaning." However many woofs a dog may utter, they will never be wooves. Even tame onomatopoeia, and sound symbolism more broadly, has something of the "wild" in it, arising outside the usual system and appearing across systems. This is seen to be so in other sound-symbolic forms as well, such as the more ideophonic forms in Huastec, as the sound changes

"would destroy established sound–meaning correlations" (Kaufman 1994, 71). As Traunmüller (1996, 147) says, "in cases of universal sound symbolism we should, then, not only expect similar forms to arise without any linguistic connection, but also that forms motivated in such a way should survive for a longer time and more easily be diffused by borrowing."

But should an imitative word lose its directly representative aspect, it enters into the general current. Such is the case, for instance, with some imitative names of animals and insects. For example, *cicada*, originally from Latin, was pronounced /ki ka da/, intended as an imitation of the insect's bombination, but the modern English /sɪ keɪ də/ manifests *k*-frication and the Great Vowel Shift, perhaps along with a lesser tendency for Anglophones to perceive buzzes as made of individual clicks. Thus onomatopoeia is tamed to the point where it is no longer even recognized as such, and it has become a fully conventional and apparently arbitrary word without frankly imitative function.

Another effect of the more persistently imitative nature of onomatopoeia and sound symbolism can be a greater association with less common – and newer – phonemes and other features. As Matisoff (1994, 121) says, "There is much evidence to suggest that it is the lexically rarer tones in a language which are typically exploited for special jobs: in morphophonemic processes, in incompletely assimilated loanwords, or for affective/symbolic purposes." Palatalization in Japanese, which in most combinations is not found in natively Japanese words, is also used in their "mimetic words" – a kind of ideophone (Hamano 1994). In Yir-Yiront, Alpher (1994, 162) finds that "The statistical distribution of initial consonants . . . in ideophones differs dramatically from that in ordinary words." In Greek, Joseph (1994) finds that /ts/ and /dz/ are heavily represented in expressive "allolanguage" and underrepresented elsewhere (223), and in fact "are the primary exponents of phonic expressivity in general in Greek" (230). The /ø/ of Finnish, a comparatively new phoneme in the language, has a stronger association with sound symbolism, as noted by Austerlitz (1994). We may find a parallel in English /z/, likewise

comparatively new as a separate phoneme, and notable for its use in imitative and expressive word forms.

The causes of such associations can be several. Imitation may have played a role in the establishment of a given phoneme in a language, although it is not likely the main source. More to the point, the phoneme, being newer, is more available – with fewer conventionalized forms to confuse the communication – and potentially more ostentatious, and less subject to the kinds of conventional alterations that might reduce its imitative potential. As Matisoff (1994, 121–122) writes of use of rarer tones, this ensures that they "will not overburden the system by creating large numbers of new homophones – and the salience afforded by their very rarity makes them appropriate for grammatical or symbolic duty." They gain a greater association with expressivity because there are fewer associations with less-expressive words, and they are more attention-getting because of their uncommonness as well. Onomatopoeia is a performance, and cues such as less-common sounds more associated with performativity will help set the stage.

1.1.3 Ideophones

Ideophones are a type of word that adds a particular performativity to a sentence. They are especially remarked in many African languages, but have also been observed in various forms in other languages, from Japanese (Hamano 1994) to Huastec (Kaufman 1994), and even arguably in English. Voeltz and Kilian-Hatz (2001, 3) give a good account of their nature: "Ideophones simulate an event, an emotion, a perception through language." A good English illustration of the nature of ideophones can be gotten from a lurid account in David Mamet's play *Sexual Perversity in Chicago* (Mamet 1974): "by accident it catches her a good one on the ass, and *thwack*, a big red mark" (14); "Zip, zip, zip, and she gets into the flak suit" (15); "and WHOOSH, the whole room is in flames" (17). The words filling the roles of ideophones would be *thwack; zip, zip, zip;* and *WHOOSH*.

This example may, however, give the misleading impression that ideophones are simply a version of onomatopoeia. In fact, "Most ideophones do not imitate sounds

in nature simply because many make no reference to sound" (Childs 2003, 120). For example, in Ewe, *lilililili* on a high tone means "nice good sweet smell" and on a low tone means "very bad smell" (Ameka 2001, 30). In Yir-Yoront, *chawarrq* refers to "picking up and carrying off," *poth* refers to "smoke puffing up from a fire," and *pillii* refers to "woman's sexual arousal" (Alpher 1994, 172–173). Childs (1994, 182–183) gives examples of ideophones from Gbaya and Ijo referring to various kinds of light, from Hausa and Yoruba referring to surface or shape quality, and from Kisi referring to balance and temperature. Ideophones do have some things in common with some onomatopoeia, as mentioned above: for one thing, they are often immune to phonological rules (Childs 2003, 122) and tend to have distinctive phonological patterns, such as "patterns of sound symbolism, reduplicative structures, or distinct patterns of tones" (Matthews 1997, 169); for another, they tend to have very little affixation or other morphology (Bodomo 2006, 204), which is consistent with treatment of sound symbolism and onomatopoeia. On this basis, one may expect them to be used mainly as interjections, as we see in the Mamet example above, or as appended adverbials; however, they can occupy any syntactic position, although within a given language their syntactic roles are often circumscribed. For example, they can be in any class in Ewe (Ameka 2001, 32); they can only be clause-final in Kisi, and must be introduced by a dummy verb (Childs 1996, 85); they can only be adverbials in Gbeya (Samarin 1991, 53). But they are in any event not nonce-words or fanciful ejaculations; they are not like "wild" onomatopoeia. They are established forms, fully lexicalized, and always – regardless of their performativity – filling a specific grammatical slot (Childs 1994, 180).

Though ideophones are not always directly iconic, there can be some degree of iconicity. There may be what Dingemanse et al. (2015, 606) call *relative iconicity*: "relations between multiple forms resemble analogical relations between meanings." There is also *gestalt iconicity*, where the structure of a word is patterned on the structure of an event, for example syllabic repetition to indicate real-world iterativity (Perniss, Thompson, and Vigliocco 2010, 3). Magnus (2001, 77) discerns what we may call *syllabic iconicity*: "The position that a consonant occupies in a

syllable also affects its meaning. Consonants that appear before the vowel form the backdrop for the action of the word, and consonants that appear after the vowel express the result of the action implicit in the word." Nygaard et al. (2009) have also found that speakers use prosody to help process word meaning. This is a sort of morpholexical analogue of syntactic iconicity, in which syntactic structures tend to mirror real-world relations such as "sequence, contiguity, repetition, quantity, complexity, and cohesion" (Perniss, Thompson, and Vigliocco 2010, 2).

One other thing ideophones have in common with onomatopoeia is performativity, making them what we may call vocal gestures (as in Voeltz and Kilian-Hatz 2001, 3). In Japanese they are called "mimetic words," even when they aren't truly imitative, as with *kyoro-kyoro* 'to look around curiously without focusing on one thing' (Hamano 1994, 149). They may even be accompanied by physical gestures (Voeltz and Kilian-Hatz 2001, 3; Childs 1996, 84); in some cases, the gesture is an integral part of the ideophone: "In Igbo, the ideophone /kpáṃ-kpáṃ/ is always accompanied by two claps of the hands in time with the pronunciation" (Welmers 1973, 463). We may thus want to think of ideophones as standardized irruptions of performativity into speech. But we should not think of them as rigidly, ritually codified; they often have variants, and choice of ideophone is still reasonably free – even family members will not always be in perfect agreement (Childs 1994, 198).

1.1.4 Phonaesthemes

What phonaesthemes are

A workable definition of *phonaestheme* is **a phonemic grouping that, within a language, correlates well above chance with a particular semantic quality in etymologically unrelated words.** This is more stringent than some presentations, which consider the phonaesthemic value of context- and topic-specific phonemes or clusters even when the great majority of occurrences of that phoneme or group do not appear to have the target semantic value, such as Bergen (2004, 293), who sees them as "form-meaning pairings that crucially are better attested in the lexicon of a language than would be predicted, all other things being equal," and in practice

looks at combinations of phonemes, especially onsets. Others such as Rhodes (1994) expand their purview to single phonemes, considering what things they tend to express when they are used expressively – in words that are patently sound-symbolic. For instance, he discerns a significance of "anchored" in /p/ and /b/ onsets, giving as examples "push, pop, bump, bounce, etc." (276). Even broader are those views that see phonaesthetic value in all phonemes, such as Magnus (2001, 4), who presents her "Phonosemantic Hypothesis":

> In every language of the world, every word containing a given phoneme has some specific element of meaning which is lacking in words not containing that phoneme. In this sense, we can say that every phoneme is meaning-bearing. The meaning that the phoneme bears is rooted in its articulation.

This is not a novel position with Magnus, but it is a premise rather than a conclusion, and it is a premise that is far from universally accepted. Magnus's experiments give evidence for phonaesthemes, but do not sufficiently support an unquestioning adoption of her phonosemantic hypothesis, and since it is not necessary to accept it in order to research phonaesthemes, I will not lay that particular stone in my foundation.

A more qualified approach is that of Rhodes and Lawler (1981), who propose that since the strategies that lead to the production of phonaesthemes – analyzing words by resemblance and decomposing syllables by onset and rime – are always operating at some level, we can always bring phonaesthetics (not the term they use, but what they are referring to) into meaning analysis at some level, even if faint. This is a very appealing proposal and one I have no interest in arguing against, but in much of the language the semantic effects of phonaesthetics are too faint to disentangle from other effects, so for the purpose of this study I will limit myself – and my definition – to what is clearly defensible. To have an acceptably rigorous study, I must limit myself to phonaesthemes that have a statistical defensibility, and I will refer otherwise to *purported* phonaesthemes. A still more stringent test would specify a phonaestheme's usability in neologisms, but I believe we will find that this

is part and parcel of what they are for the same cognitive reasons that bring them to being in the first place. I wish to propose that they also have an inevitably heightened expressive or performative quality, but this is not a feature of their definition – rather, it is something that I will have to demonstrate.

To some extent, we may say that phonaesthemes are like ideophones but on the sub-word level. This analogy carries only so far, however. Ideophones are clearly lexicalized, recognized consciously as word forms with significance; phonaesthemes are expressive and arguably performative, but they are not quite so lexicalized per se; indeed, speakers may make full use of them without being fully conscious of them as word parts. But, as Bloomfield (1933, 156) says, "to the speaker it seems as if the sounds were especially suited to the meaning." While they may seem to have a certain inevitability to speakers of a language, however, they do not form a reliable pattern inter-linguistically: languages "show some agreement, but probably more disagreement" (Bloomfield 1933, 156) in what phonetic forms suit what objects.

Phonetic composition, place in words, and distribution in language

Perhaps the best known examples of phonaesthemes in English are the onset clusters *gl*–, associated with light and vision, and *sn*–, associated with the nose and mouth. Phonaesthemes are not limited to onsets, however, although onsets seem strongly represented among them; rimes and codas (including syllabic ones such as the /zl̩/ in *fizzle, sizzle, dazzle*) may also be phonaesthemes. As Bergen (2004, 292) points out, phonaesthemes are like bound morphs and "cranberry morphs" (such as the *cran* in *cranberry*) in that they can't stand on their own; like the former and unlike the latter, they appear in numerous words (indeed, since we identify them on the basis of their repeat occurrence, they are by definition never one-offs); but they are unique in that they "occur in words in which the remainder is often less morpheme-like than the phonaesthemes are themselves" (Bergen 2004, 292).

As well, phonaesthemes overlap with morphemes; a phonaestheme in any given word will be a part (rarely the whole) of a morpheme (and occasionally will cross morpheme borders). Inasmuch as they are typically smaller than morphemes but

still carry meaning in some way, some such as Abelin (1999, 5) even use them to question whether morphemes are truly the minimal meaning carrying units. But phonaesthemes' role is not semantically exhaustive; as Abelin says, although you can use phonaesthemes to create nonce onomatopoeia in some instances, with the total sense of the word being the sum of the effects of its parts, in a lexicalized sound-symbolic word the meaning is *more* than the sum of its parts.

What kinds of words contain phonaesthemes? According to Bergen (2004, 290), "in general, phonaesthemes seem to appear in content words over function words, and in more specific (or subordinate level) rather than more general (or basic level) words." They are also relatively stable over time (Abelin 1999, 49–50) – again, this is a diagnostic feature of them: if they were not stable we would not have identified them. But they can even be common features of related languages, suggesting deep historical roots: "in a given set of related languages, phonaesthemes that appear in some languages will also appear in other languages but in words that are not cognates" (Bergen 2004, 290). Phonaesthemes may also have variants, analogous to allophones. For instance, Abelin (1999, 7) notes that in Swedish *pj–, bj–,* and *fj–* are all pejorative and may perhaps be seen as versions of the same underlying phonaestheme, although she does not propose rules governing which appears where.

Performative nature

We have established the imitative nature of onomatopoeia and sound symbolism, which makes an utterance a de facto performance of sorts, and we have seen that a central characteristic of ideophones is their performative nature, even as they are in many cases arbitrarily associated. As Bloomfield (1933, 156) says, "Symbolic forms have a connotation of somehow illustrating the meaning more immediately than do ordinary speech-forms." Malkiel (1994) goes so far as to expressly equate "phonosymbolism" with expressivity. What we will tend to find with phonaesthemes is that they also have a certain performativity – they lend an air of demonstration and are perceived as more vivid and involved and less detached.

Indeed, most of the "symbolic forms" Bloomfield uses for illustration are phonaesthematic: he lists several *fl–, gl–,* and *sn–* words. We can see that some phonaesthemes have evidently onomatopoeic aspects: the *spl–* and *–ash* in *splash* both carry some sound imitation. But performativity need not be strictly imitative of things directly available to the senses; anything that permits of figurative reference – metaphoric, metonymic, synecdochic – can be expressed in vivid, performative terms: *crunching numbers, making a splash, sparkling wit*, and so on. We have also seen with ideophones that any quality, however abstract, can lend itself to expressive emphasis if it can be endued with a conceptual vividness. Abelin (1999, 90–92) surveyed the most common semantic features associated with onset phonemic clusters in Swedish and found that the five most frequent – by a fair margin – were 'pejorative', 'sound', 'long, thin form', 'quick or strong movement', and 'wetness'. Following behind were 'talking', 'light', and 'diminutive'.

In sum, phonaesthemes seem available for any context where a particular vividness – illustration, demonstration, involvement, expression of emotional attitude – is desired, even when discussing abstract topics. A phonaestheme participates in a particular tone. This performativity can thus condition the contexts and genres for which phonaesthematic words are seen as more or less appropriate. Some writers (such as Joseph 1994) make a distinction between conventional "microlanguage" and what they call "allolanguage," which includes "non-human communication systems, child language, interjections, language play, and the like, and is expressive, affective, connotative, colorful, and iconic, while microlanguage has none of these properties" (222). The data others have found regarding phonaesthemes and the data I will add show that this distinction is weak, porous, perhaps even nonexistent: performativity and expressivity can be present in "microlanguage" through phonaesthemes, among other means.

Systematicity

As a language changes over time and expands its vocabulary, we might at first expect that it will increase in arbitrariness – the initial expressive bases of word

forms will gradually be etiolated, bleached, worn threadbare. However, if clusters emerge and the vocabulary grows under the influence of these clusters, then the process over time may be something more the opposite: just as the even distribution of the matter of the universe immediately after the big bang came to cluster and clump and form planets and stars and galaxies, so too may these lexical attractors gain an increasing gravitational force and increasing influence. This is the effect I discussed above, what Magnus (2001, 6) calls *phonosemantic association*: "When semantic domain S is associated disproportionately frequently with phoneme X, then people will be inclined to associate semantic domain S with phoneme X productively." Bolinger (1968, 242) describes the process:

> Given a particular word for a particular thing, if other words for similar things come to resemble that word in sound, then, no matter how arbitrary the relationship between sound and sense was to begin with, the sense is now obviously tied to the sound. The relationship between sound and sense is still arbitrary, as far as the outside world is concerned (and would appear that way absolutely to a foreigner), but within the system it is no longer so.

This kind of gravitational clumping is in the line of what Dingemanse et al. (2015) call *systematicity*: statistical regularities in association of form with function. Such regularities have been attested by various studies. Monaghan et al. (2014) used multiple measures of phonetic and semantic distance to calculate correlation and found a small (>0.03) but significant (p<0.0001) overall correlation between the two in English, even among etymologically unrelated monomorphemic words. Specific examples abound. For example, Reilly et al. (2012) found that in English, shorter words were associated with concreteness and longer ones with abstractness, except for one- and two-phoneme words, which were thought of as less concrete likely due to an association with function words. The origin of this association is plausibly ascribable to our use of classically derived words for abstract concepts, but the experiment – involving constructed "nonwords" – showed that there is now a tendency to expect the association regardless. Similarly,

Monaghan et al. (2007) noted phonological cues that tend to be predictive of word classes in several languages: in English, syllable length and proportion of vowel sound help identify nouns, and approximants in the first syllable help identify verbs; in Japanese, fricatives and rounded vowels are associated with nouns, while coronals are associated with verbs; in French, bilabials in the first syllable are associated with nouns, while proportion of vowels helps identify verbs. Some of these associations are for fairly clear morphological reasons, but that still leaves the question of why the inflectional and derivational morphemes happened to have those characteristic tendencies in the first place. Other instances are aspects of which few speakers are conscious, and yet speakers in general form automatic expectations (with varying degrees of anticipated probability) on the basis of them.

Systematicity is useful for children in learning language – and indeed the patterns learned in language acquisition are prone to generalization that further reinforces the same pattern. The inclination to systematic learning is what allows children to learn the inflectional and derivational morphology of a language, but the same faculty seems also to be put to use more broadly. Systematic correspondences between word sounds and grammatical categories can help children learn language, as Cassidy and Kelly (1991), Monaghan, Christiansen, and Chater (2007), Fitneva et al. (2009), and Monaghan et al. (2012) have found for several languages. On the basis of their findings, Monaghan, Christiansen, and Chater have formulated a Phonological-Distributional Coherence Hypothesis, "which predicts that there will be correspondence between phonological properties of words and their grammatical category" (2007, 266). Words learned earlier tend to show less arbitrariness and more iconicity, according to Dingemanse et al. (2015, 609) and Monaghan et al. (2014). This can give a sense of a more basic, vivid, elementary concreteness to words that more plainly exhibit systematic cues.

Systematicity is also useful for adults in quickly processing language. As Reilly et al. (2012) found, adult speakers exploit phonological regularities to facilitate lexical access. Tendencies of word classes allow "syntactic bootstrapping," whereby we are

able to quickly tentatively slot words syntactically on the basis of their phonological characteristics as a first or early step in processing. This is not the only factor involved, and in some cases it is not even a reliable one – many words exist in identical form in multiple lexical classes, including several of the ones I am studying in this thesis. But it is a factor.

Systematicity can involve form-meaning associations that are otherwise arbitrary or happenstance (neither fricatives nor bilabials are intrinsically more "nouny"), but they can also involve form-meaning associations that are imitative or performative, and they may even help spread the association with a performative or "vivid" orientation to words that partake in the association. Dingemanse et al. give ideophones as an illustration of interplay of systematicity, iconicity, and arbitrariness: "ideophones are built from language-specific phonological inventories (introducing a degree of arbitrariness), they show various cross-linguistically recurring correspondences between form and meaning (iconicity) and they can be recognized as a word class by language-specific phonological cues (systematicity)" (2015, 604). We will find that these statements are true of phonaesthemes as well.

A fully developed language cannot sustain a pervasive iconicity or even an absolutely consistent systematicity; there are too many things to name, and many of them are not susceptible to iconic representation. But its foundation – its most basic vocabulary – is, we see, prone to greater iconicity and systematicity. As Reilly et al. (2012, 1) tell us, concrete words tend to be shorter than abstract words, and young children acquire concrete words well before they acquire abstract ones. It follows from this that iconicity and systematicity are more likely to be associated with shorter words and with earlier learning. As well, in a language such as English, where the core vocabulary is heavily Germanic while much of the more learnèd vocabulary comes from Latin and Greek, the characteristics of the old Germanic core words will be associated with the early-learned characteristics, and the characteristics of French and Classical words will be associated with such things as one learns later in one's education, such as abstraction and formality. Reilly and

Kean (2007) also point out that in English, concrete nouns tend to be Germanic in origin, while abstract ones are most commonly Latinate. As we will see, though, this can present a challenge – and something to control for – in a comparative evaluation such as the one that is the meat of this thesis. Systematicity is, after all, the reason phonaesthemes beyond the plainly iconic can come about and be effective. We can expect that it may also contribute to the perceived tone, level, and genre-appropriateness of phonaesthemes and words containing them. We will need to try to separate the length effect (and early-learning Anglo-Saxon effect) from such phonaestheme effect as there may be.

Probabilistic nature

Systematicity involves probability and probabilistic learning. Phonaesthemes are not tidy, they are impressionistic; this is not phonomathematics, after all. Or perhaps it is: Bergen (2004, 302) credits the effect of phonaesthemes on development of form and meaning in words to "statistical over-representation of a particular pairing between form and meaning in the lexicon." When we talk about morphology, we do not need to talk about how likely a given set of phonemes is to be that morpheme, or to what extent it is that morpheme. But when we talk about phonaesthemes, statistical evaluation enters into the discussion early and often. As Bergen (2004, 303) notes, "Research on a number of seemingly unrelated topics indicates that language users integrate and make use of statistical correlations between sound and meaning, even when these relations do not play a productive role in the linguistic system." For example, Cassidy, Kelly, and Sharoni (1999) found that in a corpus of 490 English-language personal names, 80% of the consonant-final ones are male and 72% of the vowel-final ones are female, while 94% of iambic disyllabic English personal names are female.

So, too, with phonaesthemes: Bergen (2004, 293) shows that 39% of word types and 60% of word tokens starting with *gl* relate to light; 28% of types and 19% of tokens starting with *sn* relate to the mouth, as do 25% and 27% (respectively) starting with *sm*. These are well above chance, but they are not strict or solidly rule-governed as

we would expect with morphemes or phonemes. Other phonaesthemes have also been identified. Hutchins (1997) proposed 46 phonaesthemes; Otis and Sagi (2008) found that 27 of them were statistically significant.

Abelin (1999, 57–58) presents two key points that should be considered in any model for phonaesthemes: "1. Some sounds/sound combinations are (judged to be) better suited for some (types of) meanings, within a given language or for many languages"; "2. Some meanings are better suited for being expressed with some of these sounds/sound combinations." What we see from this is again that phonaesthemes are probabilistic, tending to occur with a given sense but not reliable occurring all the time. Beyond this is the question of what even counts as a phonaestheme. Do we use some version of statistical significance? How close does an association need to be, how strong does an influence need to be? When we look at a work such as Marchand (1959), we see quite a lot of possible or purported phonaesthemes listed, including quite ordinary sounds that seem to have just a little correlation with a particular sense. Is it reasonable to call something a phonaestheme if only a small number of its instances are in expressive words, but it has a consistent reference among those that are? Indeed, can we even say that phonaesthemes are psychologically real as opposed to artifacts of analysis?

Psychological reality of phonaesthemes: experiments

Experiments give evidence that phonaesthemes have some psychological reality. Bergen (2004) conducted a priming experiment that showed that "phonaesthemes, despite being noncompositional in nature, displayed priming effects much like those that have been reported for compositional morphemes" (290). Subjects were presented with sequential pairs of words that shared either (a) a phonaestheme, (b) just a phonological connection, (c) just a semantic connection, (d) a connection both phonological and semantic, but not well represented in the lexicon (e.g., *crook* and *crony*), or (d) no connection. A shared phonaestheme between words produced a facilitatory priming effect (decreased processing speed for the second word) that was greater than just the additive effects of form priming and semantic priming.

Notably, “phonaesthemic priming is not observed between simply any two words that by chance share both some phonological form and some meaning – it surfaces only when the form-meaning pairing is well attested in the lexicon” (Bergen 2004, 291). These effects emerge even when the subject is under time pressure and thus processing the stimuli unconsciously, like natural language (301).

Magnus (2001) went in the other direction, asking subjects to invent novel words for meanings that happened to be associated with phonaesthemes, and found that the associated phonaesthemes were used at greater-than-chance frequencies; as well, when subjects were presented with invented words that included phonaesthemes and asked to invent meanings for them, their invented meanings matched with the phonaestheme’s usual meaning with greater-than-chance frequency. However, “greater-than-chance frequency” in these cases can mean 25% of the time, which is likely different from what one would get when putting together words using known morphemes. Magnus also surveyed the lexicon for specific associations between semantic categories and phonemes and found such associations did exist – for example, words in the categories “Bulges, Mountains, Humps and Peaks,” “Fountains and Blowing,” “Foundations,” “Beginnings,” and “Pairs, Names, Pictures, Symbols” were found to favour labial consonants. Given the arbitrariness of the semantic classes and the broadness of the phonemic set, this result is not a strong pillar in the edifice of phonaesthetics, but it may at least be a wall joist.

Abelin (1999) did several experiments with Swedish speakers matching meaning to form and vice versa, both free and forced-choice. In one experiment, 14 subjects were shown 38 questions where a meaning was presented and they were asked to choose the most likely of three possible constructed forms to match the meaning; 28 of these received a majority of expected answers, and one had 14/14 matches (223). In another experiment, 15 subjects were shown 38 constructed words and were given a choice of three possible meanings; 29 of these received a majority of expected answers (i.e., consistent with the sense of the identified phonaestheme),

and 4 had 15/15 matches (226). This is, admittedly, a "soft" result, but it is consistent with what has been found elsewhere. In a third experiment, 15 subjects were asked to invent meanings for 6 constructed words. With one exception, each word had between 3 and 10 constructed definitions matching the semantic category for an identified phonaesthemes. In the last experiment, 14 subjects were each asked to invent a word for each of 6 general meanings associated with phonaesthemes. Most but not all of the items were given a greater-than-chance portion of constructed words that contained known phonaesthemes identified for those meanings, in a frequency similar to that seen in the general vocabulary. More interestingly, Abelin found that when asked to create new words to express specific meanings, subjects tended to "encode the semantic features in initial clusters rather than in final clusters" (Abelin 1999, abstract). However, since Swedish verbs have inflectional suffixation on all forms, finality is less final in verb roots and so may be less salient.

Universality

We have seen that there is, if not universality, at least a broad commonality in some aspects of sound symbolism – in particular relations between vowel location and such qualities as size, weight, and possibly shape. Phonaesthemes appear to play a similar role in a language: attaching expression of certain semantic values to particular sounds or sets of sounds. To what extent is there carry-over between languages? An important point of phonaesthemes is that they are *not* simply etymologically based – a sound-meaning correlation in a set of words that all have a common root already has an explanation and needs no new fancy polysyllabic term to posit another. So it does not necessarily follow automatically that languages that split from their common ancestor so long ago as to be mutually unintelligible would have the same or closely related phonaesthemes. The Swedish examples above (pejorative *pj–*, *bj–*, *fj–*) are not broadly productive in English, for instance. Still, it seems plausible that there would at be at least some carry-over, and in fact there are points of commonality between Swedish and English; for example, Abelin (1999, 35) notes that there are *fl–* words for unsteady movement in both languages (*flicker*,

flutter; *fladdra*, *flaxa*). Similar points of resemblance may be present in other Germanic languages as well; for example, German has *flimmern* ('shimmer, flicker') and *flink* ('nimble'), as well as various *fl*– words relating to flight and flames, obviously cognate with their English counterparts.

But to what extent do phonaesthemes carry over between languages, and how? Would they be subject to the same sound changes as have prevailed generally? If so, this would put them at odds with a principle of onomatopoeic words observed above: that they are generally exceptions to regular sound changes due to their imitative character. Or are there multiple currents? Abelin (1999, 22) suggests that there could be both imitative (sound-symbolic) inputs and sound-clustering (what I might call gravitational) inputs. She adds, "If the semantic-phonetic relationships of motivated words could be analytically treated one by one, my assumption is that the existence of universality in phonesthemes on the phonetic side (i. e. that e.g. imitation of 'wet sounds' is done with the same speech sounds in different languages) is most likely at a level of (combinations of) distinctive features, e.g. voiceless, fricative, etc." (1999, 22) So, for instance, while English phonaesthemes for light include *fl*– and *gl*–, Swedish phonaesthemes for light include *bl*– and *gn*– (1999, 35).

There is also the question of commonality between languages that are entirely unrelated and even without contact. While onomatopoeic sounds may be expected to have some features in common when they are imitating the same originals, it would be striking to find matching sounds in phonaesthemes representing things that don't make sounds. But to what extent should we expect onomatopoeia to imitate the same originals (even beyond variations in local fauna), and to what extent should we expect phonaesthemes to focus on the same non-acoustic properties? Are some things so basic – various kinds of intensity or motion, for instance; visual or tactile extremes or types – that they can be expected to show up in unrelated, geographically disparate languages? We already have something of an answer to this in certain quasi-universals of sound symbolism, such as the

kiki/bouba types of distinction, and we can see that ideophones also often focus on what we might think of as cardinal qualities. To the extent that physical properties are mapped consistently by different languages onto non-physical properties, we can expect similar kinds of topics for phonaesthematic expression.

Phonaesthematic attraction

What does the existence of phonaesthemes suggest about the nature of language and its use? For Bergen (2004, 290), "the results support a view of the lexicon in which shared form and meaning across words is a key factor in their relatedness, and in which morphological composition is not required for internal word structure to play a role in language processing." Compositionality is part of the picture but not all of it. This view is supported by findings such as those of Cassidy, Kelly, and Sharoni (1999), who found that in English male names tended to have word-initial stress and to end in consonants, while female names tended to have word-final stress and to end in vowels, a pattern supported but not fully accounted for by the morphology of Latin and some other related languages; of Kelly, Springer, and Keil (1990), who found that adults and children who speak English have and use an internalized correlation between the number of syllables in a word and the complexity of what it names; of Cassidy and Kelly (1991), who found that English verbs tend to be shorter than English nouns (possibly for reasons relating to their syntactic positions), and that both adults and children are more likely to assume pseudowords are verbs if shorter and nouns if longer; and of Sereno (1994), who found that frequent English verbs have more front vowels than back vowels, while the reverse is true for frequent nouns, but no such pattern is found in the less frequent lexical items, which may suggest that words tend to be shaped by more frequent usage to conform with more expected pattern correlations – something that phonaesthemes seem also to be. (Sereno does not propose a clear explanation for the origin of the phenomenon; one might speculate about Germanic ablaut and umlaut morphology, but that would require support and would in turn leave us wondering about its origins.)

Bolinger (1968, 219) makes the point well about the relationship between expectation and sound and meaning:

> Children sense the associative possibilities and coin words with them: ***If the house is as old as that it's raggy, shaggy, and daggy***, remarked one seven-year-old. . . . The makers of multiple-choice tests find phonesthemes useful as distractors for their questions; if ***twisted*** is offered as an equivalent for ***knurled***, it is on the assumption that persons not fully acquainted with ***knurl*** will assume that it is related to ***twirl, whirl, birl, tirl, furl,*** and ***gnarl.*** Shifts of meaning often go in the direction of a family of words having phonesthematic ties. The word ***bolster*** no longer suggests a padded and comparatively soft support but rather a stiff and rigid one, because of the attraction of ***brace, bolt, buttress.*** (Of seventeen persons tested on this point, thirteen voted for 'rigid.')

Bergen (2004, 304) observes that both network models and connectionist models "predict that statistical recurrences across words, like phonaesthemes, will automatically rise to the status of organizing structures in a language." Bolinger (1950) argued that similar forms in words in similar semantic areas would tend to exercise a sort of attraction on each other; Hock and Joseph (1996, 293) give one example of this, where English *sacke* became *sag* by analogy with *drag, flag,* and *lag*, which have in common a sense of "slow, tiring, tedious motion." They call this effect *phonesthematic attraction*, although they could at least as well have called it *phonosemantic attraction*, since similar effects can operate on words that are not phonaesthematic. For example, Malkiel (1994) documents how French *clore* 'close' from Latin *claudere* gave way over time to *fermer* under the influence of *firmare* 'make firm' and *ferrum* 'iron', and Spanish *pechar* 'to bolt' shifted to *fechar* under the same *ferrum* influence. From this we can see that phonaesthematic attraction, such as it may be, is really just another instance of systematicity – word formation and adaptation by analogy. The formation of blended words through use of pseudomorphemes – common examples include *–copter*, *–gate*, *–aholic*, *–palooza*,

and –*mageddon* – functions quite similarly. The difference here is only that one of the qualities involved is the expressivity and performativity of the phonaestheme, and this may carry with it a particular tone and level of use and of self-presentation of the speaker or author.

Along with this, we see that some words seem to become more expressive over time – to shift meaning towards a more expressive sense. Jespersen (1922) gives the example of *patter*, which came from *paternoster* and at first referred just to repeating that particular prayer, but has come to refer to rapid speech that may be suggested by the sound of the word "patter." In short, the tendency to systematicity already discussed appears to manifest itself as phonaesthemics as well. This would be consistent with the view of Jespersen (1922): that languages over time grow richer in sound-symbolic, expressive words. It is not likely, however, that this is the dominant factor in language development. Sound shifts would tend to be suppressed by the influence of sound symbolism, which would be at risk of losing its imitative quality.

This can lead us to further explorations of the role of phonaesthetics in language change: to what extent these attractive effects have shaped the form of words (through shifts in form as well as through neologism) and the choice of one word over another for a given meaning. But we need also to address the extent to which phonaesthemes, such as they are, truly relate to aesthetic and performative aspects of words, as opposed to being simple statistical correlations. Are they genuinely sound symbolic or related to ideophones? How do we prove this? To what extent are words that use them seen as more "vivid" and vice-versa, and to what extent are they used in contexts that are genuinely more performative or "expressive"? One way to come at these questions is to examine the interaction of phonaesthemes with genre.

1.2 Genre

1.2.1 The nature of genre

Communication with language involves lexis, phonology, and morphosyntax, but it goes beyond that. "From a sociolinguistic standpoint," Halliday (1978, 61) says, "a text is meaningful not so much because the hearer does *not* know what the speaker is going to say, as in a mathematical model of communication, but because he *does* know. He has an abundance of evidence, both from his knowledge of the general (including statistical) properties of the linguistic system and from his sensibility to the particular cultural, situational, and verbal context; and this enables him to make informed guesses about the meanings that are coming his way."

This is a question not just of which possible value to give to a word such as *snipe* – ornithological, military, figurative? – but of the structure of the text, not just syntax but events described and flow of reasoning, and of the emotive attitude towards the text and what it describes – the author's attitude and the attitude expected of the reader, which may not be the same thing. It is a matter even of the attitude towards the act of communication, and the expected behaviour. What is the author or speaker doing, and how? And the reader or listener? The text of a play constructs these roles more distinctly than most texts, but even with a newspaper or novel there is an expectation of the situation of the writer, and that of the reader. Our full understanding of a text is contingent on our understanding of its genre. As Halliday (1978, 137): says, "To say that a text has meaning as literature is to relate it specifically to a literary universe of discourse as distinct from others, and thus to interpret it in terms of literary norms and assumptions about the nature of meaning." To those who say that literary criticism can proceed quite well simply by evaluating a work on its own merits without reference to the genre of which it is a part, Genette (1992, 81) points out that it inevitably resorts to generic conceptions and expectations without being aware of it – for example even the existence of such a thing as a *novel*, and central facts about its nature. Rosmarin (1985, 14) cites Gombrich's image in *Art and Illusion* of any work of art (thus including literature)

being like a snowman: "all art, even that which strives to conceal this fact, begins with a schema. Thus when we make a snowman we 'work the snow and balance the shapes till we recognize a man. . . .'" Any created, structured communication must similarly work from and with a schema (or multiple schemata). Of course linguists know well that there has to be a pre-existing understanding of syntax and other features of a language; we need only extend that understanding to the larger levels, beyond the sentence. And as with creation, so too with comprehension and explication (criticism). Can there be text without genre? As we will see, this question is along the same lines as "Can there be communication without pragmatics?" Bawarshi (2000, 338), looking just at written texts, proposes – in response to Foucault's "author function" – a "genre function," "which constitutes all discourses' and all writers' modes of existence, circulation, and functioning within a society, whether the writer is William Shakespeare or a student in a first-year writing course, and whether the text is a sonnet or a first-year student theme." In this perspective, a text could no more be free of genre than it could be free of grammar.

Much genre theory focuses on literary genres. For Genette (1992, 64), genres are literary categories (or rather aesthetic ones, since other arts also have genres), while modes are categories that belong to linguistics, and in particular to pragmatics. Genette (1992, 82) posits an *architext*, which is a sort of archetype of genre: each text relates to the architext or architexts of the genres to which it belongs or relates; this relationship is *architextuality* (83). There will never be a perfect match, of course; as Rosmarin (1985) says, genre is "a finite schema capable of potentially infinite suggestion," (44) and "genres can never be perfectly coincident with texts unless we posit as many genres as texts" (45) – which would be the limit case as genres become more and more specific.

Genre extends beyond literature, however. A language in a society does not exist without a system of genres, and that system is all-encompassing. We do not think twice about it in much of life; as Todorov (1990, 10) says, "everyone knows that one must not send a personal letter in the place of an official report, and that the two are

not written in the same way." He continues, "Any verbal property, optional at the level of language, may be made obligatory in discourse; the choice a society makes among all the possible codifications of discourse determines what is called its *system of genres*." Biber and Conrad (2009, 23) take the same view: "register/genre variation is a fundamental aspect of human language. All cultures and languages have an array of registers/genres, and all humans control a range of registers/genres." (For *register* versus *genre*, see below.) Every context and function dictates its own linguistic exigencies, and we develop expectations for them – expectations that organize by association. Each genre has its expectations, some topic-specific, some overtly set (novels have titles; news articles have headlines; Tweets have neither and are limited to 140 characters), some matters of learned convention or current trend, and some statistically learned. "Like any other institution," Todorov (1990, 19) says, "genres bring to light the constitutive features of the societies to which they belong." Overall, the impressionistic, resemblance-based, effect-directed way we initially constitute genre is similar to how we constitute phonaesthemes; formal recognition and codification is the next step, and commonly happens with genres, producing a feedback effect as I will discuss further below. A similar formal recognition and codification does not typically take place with phonaesthemes, although it could.

I have been speaking of *genre* here without reference to the related term *register* or other terms for varieties of a language divided other ways. It may be tempting to subsume one under the other. As Halliday (1978, 185) says, "A dialect is 'what you speak' (habitually); this is determined by 'who you are', your regional and/or social place of origin and/or adoption. A register is 'what you are speaking' (at the given time), determined by 'what you are doing', the nature of the ongoing social activity." Analysis of register, for Biber (2009, 2) "combines an analysis of linguistic characteristics that are common in a text variety with analysis of the situation of use of the variety." But while *genre* is often used in an overlapping way with *register*, it may carry a sense of a larger scope, or a specifically literary medium, or even something above and beyond such concerns as syntax and lexis. The most common

definition usable in linguistics (as opposed to the critical-institutional version bandied about by some scholars of literature and scorned by others – see Miller 1984, 151) involves the structure of the text. Ferguson (1994) makes the distinction clear by articulating the working assumptions involved in studying the respective forms. He gives the working assumption of studies of register variation as follows:

> A communication situation that occurs regularly in a society (in terms of participants, setting, communicative functions, and so forth) will tend over time to develop identifying markers of language structure and language use, different from the language of other communication situations. (Ferguson 1994, 20)

For studies of genre, it is this:

> A message type that recurs regularly in a community (in terms of semantic content, participants, occasions of use, and so on) will tend over time to develop an identifying internal structure, differentiated from other message types in the repertoire of the community. (Ferguson 1994, 21)

Analysis of genre, for Biber and Conrad (2009, 2) "is similar to the register perspective in that it includes description of the purposes and situational context of a text variety, but its linguistic analysis contrasts with the register perspective by focusing on the conventional structures used to construct a complete text within the variety, for example, the conventional way in which a letter begins and ends." Biber's and Conrad's view thus limits the scope of genre to a subset of all communication, that set with clear textual structure; for them, a predominant concern in genre analysis is "rhetorical organization" (17) and those structures that occur in specific locations in the text and are specialized to the function (16); the analysis of overall lexicogrammatical features they leave to register and stylistic analysis. They make a distinction between genre as they construe it in linguistic analysis and "literary genre," "varieties of literature that employ different textual conventions," such as poetry, drama, and fictional prose (19). Similarly, for Halliday

(1978, 134), "The generic structure is outside the linguistic system; it is language as the projection of a higher-level semantic structure. It is not simply a feature of literary genres; there is a generic structure in all discourse, including the most informal spontaneous conversation." In this view, a genre carries an expected structure of events or arguments, and a type of things described, but the specific choice of words and grammar is a matter of what register is preferred for that genre.

For the purposes of this thesis, I will be operating with a definition of *genre* that distinguishes it from *register* by this question of structure and the larger construction of roles implied with it, but I will not be leaving out the syntax and lexis that register makes use of; rather, I will be subsuming the various registers used by a genre into that genre, as it is not possible to define a genre without specifying the registers appropriate for use in it (often different registers for different parts of a text). I will not be engaging in what Biber speaks of as specifically genre analysis – I am not examining the structure of the genres – but I will be doing what he calls a feature of register analysis: assessing details of the lexis. When surveying a genre that may have multiple registers in it – for example, newspapers – I will for the most part be analyzing the genre as a whole rather than breaking out the individual registers. The reason for this is partly practical – in many cases it would be far too time-consuming and would produce unusably small numbers per register per genre – but also because at each order of magnitude there are constitutive characteristics; because performative expressivity is at least as much a feature of genre as of register, and is a key feature signalled by phonaesthemes; and because a genre that may use many different registers will still have one audience reading or listening to all of it, and different genres are aimed at different audiences or at the very least position themselves in different social and intellectual statuses relative to their audiences.

I speak of orders of magnitude because genre is not a single level. Genre may be thought of as broadly analogous to syntax's XP: a genre may have more specific genres within it and may in turn be contained within an even less specific genre.

Fiction, for instance, contains genres such as science fiction, which in turn contains genres such as steampunk sci-fi, speculative sci-fi, near-future fiction, dystopian YA sci-fi (which is also within the young adult genre, showing that there can be overlap), and so on; and fiction in its turn is within a larger genre of published narrative literature, and so on. Newspapers are a genre; news articles in newspapers are a genre; headlines of news articles in newspapers are a genre; each of these has its register or registers, and in cases such as headlines there is little difference between genre and register. Genette (1992, 65) uses a genus-species analogy:

> a "genre" like the novel or comedy may also be subdivided into more specific "species" – tale of chivalry, picaresque novel, etc.; comedy of humours, farce, vaudeville, etc. – with no limit set a priori to this series of inclusions. We all know . . . that with a little ingenuity one can always multiply the positions between the species and the individual, and that no one can set a limit on this proliferation of species. . . . In short, any genre can always contain several genres.

This may make genre seem merely an artifact of analysis, an arbitrary delimitation that produces its own object. But the properties of genres at each level are real; they are discernible, as we will see.

If the limit of the genre as we make it more specific is the individual text, is genre at the other limit something that simply fades into generality as the set of all linguistic expression? How is genre motivated? Todorov analyzes genre as speech act:

> is there any difference at all between (literary) genres and other speech acts? Praying is a speech act; prayer is a genre (which may be literary or not): the difference is minimal. But to take another example, telling is a speech act, and the novel is a genre in which something is definitely being told; however, the distance between the two is considerable. Finally, there is a third case: the sonnet is surely a literary genre, but there is no verbal activity such as

> "sonneting"; thus genres exist that do not derive from a simpler speech act. (Todorov 1990, 20–21)

There is a feedback effect, though, as with all socially instituted speech acts. Genre also to some extent creates its occasion, as Miller (1984, 162) observes:

> At the level of the locution or speech act, idiosyncratic motives (or what I earlier called intentions) predominate. . . . But at the level of the genre, motive becomes a conventionalized social purpose, or exigence, within the recurrent situation. In constructing discourse, we deal with purposes at several levels, not just one. We learn to adopt social motives as ways of satisfying private intentions through rhetorical action. This is how recurring situations seem to "invite" discourse of a particular type.

Genres and registers also feed back on usage, and can influence usages in other genres too. Biber (1995) cites Reder's (1981) analysis of Vai: "He found that there are systematic differences between speech and writing in Vai (e.g., certain medial consonants are deleted more frequently in speech, and indefinite noun phrases occur more frequently in writing), and that in their speech, literate adults use the forms associated with writing more frequently than non-literate adults" (Biber 1995, 281).

In short, genre can be viewed as a handy kind of script. Social interaction through texts of whatever sort is not an ongoing free-form improvisation; it always involves a choice of scripts to play out:

> what we learn when we learn a genre is not just a pattern of forms or even a method of achieving our own ends. We learn, more importantly, what ends we may have: we learn that we may eulogize, apologize, recommend one person to another, instruct customers on behalf of a manufacturer, take on an official role, account for progress in achieving goals. We learn to understand better the situations in which we find ourselves and the potentials for failure and success in acting together. As a recurrent, significant action, a genre

> embodies an aspect of cultural rationality. For the critic, genres can serve both as an index to cultural patterns and as tools for exploring the achievements of particular speakers and writers; for the student, genres serve as keys to understanding how to participate in the actions of a community. (Miller 1984, 165)

This leads us – and Bawarshi, writing 16 years later – to the idea that our social reality requires genre for its constitution:

> What about identifying genres not only as analogical to social institutions but as actual social institutions, constituting not just literary activity but social activity, not just literary textual relations but all textual relations, so that genres do not just constitute the literary sites in which literary actors (writers, readers, characters) and their texts function, but also constitute the social reality in which the activities of all social participants are implicated? In other words, to what extent is the university as an institution and the roles enacted within it . . . constituted by its genres: research articles, grants, assignment prompts, lectures, critical essays, course evaluations, memos, oral exams, committee minutes, to name just a few? (Bawarshi 2000, 347)

"Genres, in short," Bawarshi says later, "constitute the very exigencies to which their users in turn rhetorically respond, so that the genre function does not simply precede independently of us but is rather something we reproduce as we function within it" (355). When we consider questions of performance, of imitative or quasi-imitative expressivity as with phonaesthemes, genre sets the script and context, licenses these usages, and specifies where and how they may be used, and it is also in return specified by them: a genre that over time increases or decreases in its use of such forms may be expected to increase and decrease generally in its concreteness and performative expressivity. Whether this turns out to be true is one of the questions I will be answering with my research in this thesis.

1.2.2 Defining genres

Given that genre exists recursively at multiple levels, how *do* we define and delimit genres? What divisions are useful?

We should be careful not to simply impose a top-down tree and expect differentiation to increase at each lower level. A common-sense distinction between "literary" and "non-literary" genres, for instance, could lead us astray, as Todorov (1990, 11) says:

> If one opts for a structural viewpoint, each type of discourse usually labeled literary has nonliterary "relatives" that are closer to it than are any other types of "literary" discourse. For example, certain instances of lyric poetry and prayer have more rules in common than that same poetry and the historical novel of the *War and Peace* variety.

Ure (1982, 18) gives the example of letters dictated by illiterate Romanian soldiers during World War I: they "would often follow the formulae of oral poetry, itself influenced in certain respects by the existence of literate skills in members of the wider community." This also connects gives a window on the ways in which genres arise: the soldiers did not have an established "letter" genre to work within, so they used a genre that seems appropriate to the occasion.

Genre inevitably begets genre, and is begotten by genre: "A new genre is always the transformation of an earlier one, or of several: by inversion, by displacement, by combination" (Todorov 1990, 15). We must start with a schema, after all. Faced with a new situation of utterance, a person must adapt an existing structure – an existing genre – and which one is chosen will help to construct and define the new situation. Bawarshi (2000, 340) cites Kathleen Jamieson's (1975, 411) description of how George Washington's first report to Congress drew on the British tradition of the speech from the throne, and the effect that had on both the structure of the utterance and the tone of the occasion – including the form and tone of the responses from Congress. This construction of genre as a pattern that is developed

on the basis of similarity to precedent has some resemblance to the emergence and spread of phonaesthemes through systematicity – unsurprisingly, given that they're being developed by the same organ.

But how are genres identified? By analysis, but by analysis of real existing features; and, once that analysis is made, there is a further feedback loop.

> Genres are . . . entities that can be described from two different viewpoints, that of empirical observation and that of abstract analysis. In a given society, the recurrence of certain discursive properties is institutionalized, and individual texts are produced and perceived in relation to the norm constituted by that codification. A genre, whether literary or not, is nothing other than the codification of discursive properties. (Todorov 1990, 17–18)

It is conceptually most tidy, though perhaps not altogether realistic, to imagine each more specific distinction of genre dividing from others with which it is grouped in the next higher level on the basis of one particular attribute, or, at the very least, a clear set of attributes which members of one genre at that level possess and members of another do not, as Steen describes:

> Thus, the genre of an advertisement is to be contrasted with that of a sermon, a recipe, a poem, and so on. These genres differ from each other on a whole range of attributes. . . . The subordinates of the genre of the advertisement are less distinct from each other. The press advertisement, the radio commercial, the television commercial, the Internet advertisement, and so on, are mainly distinguished by one feature: their medium. The superordinate of the genre of the ad, advertising, is also systematically distinct from the other superordinates by means of only one principal attribute, the one of domain: It is "business" for advertising, but it exhibits the respective values of "religious," "domestic," and "artistic" for the other examples. (Steen 1999, 112)

If we start to carefully examine the members of the various genres Steen names, however, we are likely to find that there is no quality that all members of a genre have that no non-members have. Moreover, while we can, for instance, distinguish sub-genres of advertising by medium, we can also distinguish them by style, subject, target market, industry, or what have you, across mediums: smart-ass ads on subway posters for internet service providers, for instance, surely have more in common with smart-ass animated web ads for internet service providers than they have with text-heavy subway poster ads for social-service-providing religious organizations. Exhaustive, tidy, tree-based taxonomies are inevitably deliberately naïve.

Miller (1984, 151) presents the diversity of distinguishing criteria as her opening problem: "For example, rhetorical genres have been defined by similarities in strategies or forms in the discourses, by similarities in audience, by similarities in modes of thinking, by similarities in rhetorical situations." The result is that genre criticism is seen as simplistic, top-down, reductivist, prescriptive, and prone to creating "tiresome and useless taxonomies" (she quotes Thomas Conley). For Miller, "a rhetorically sound definition of genre must be centered not on the substance or the form of discourse but on the action it is used to accomplish" (151) – that is, "a particular effect in a given situation" (153). If we see words as like ingredients, morphosyntax as like kitchen implements, and the effect we wish to produce as like the final dish, genre is like the recipe – not just in the order and technique in which the parts are assembled, but even as far as the difference between popping something in the oven or flambéing it at the table. And when the diner orders crêpes Suzette, the latter – not the former – will be expected. Likewise, genre can dictate not just information content and structure but the style of its expression. Play scripts are written to be performed, of course (and read, too), but other genres also demand some level of performance, and while we may not buy tickets for seats to listen to a friend's narration of a wild weekend, we will probably be disappointed if it is dry and abstractly technical ("Eye contact was achieved at 8:14 pm and there

was a total elapsed time of 2 hours 43 minutes between first contact and coitus"). No one needs to instruct us in this; we learn it from experience.

Genre is constituted not just by a growing statistical expectation by association between various texts used for more or less the same purpose, however, or by the official identification of a genre as such by some textual hierophant such as a critic or scholar. It can be led by specific texts, exemplars; in particular, exceptions to the usual rules themselves help to solidify the existing rules and to set new ones:

> in order to be an exception, the work necessarily presupposes a rule; [and] no sooner is it recognized in its exceptional status than the work becomes a rule in turn, because of its commercial success and the critical attention it received. Prose poems may have been exceptional in the days of Aloysius Bertrand and Baudelaire; today, who would dare write a poem in alexandrines, in rhymed verses – except perhaps as a new transgression of a new norm? Have not Joyce's exceptional word plays become the rule for a certain modern literature? Does not the novel, however "new" it may be, continue to exert its pressures on the work being written today? (Todorov 1990, 15)

Genre division will vary from culture to culture – a genre in one cultural-linguistic group may not have an exact or even approximate analogue in another, and the ways in which similar sets of genres are distinguished may also vary. Moreover, genres that may seem uniform to external observers may be seen as distinct within the culture, with separate names. Biber and Conrad (2009, 34–35) give examples from Samoan, Apache, and Arabic-speaking Islamic cultures of genres of speech and narrative that are distinguished by subject matter or place within an occasion.

All aspects of language may be relevant in register and genre. This includes morphosyntactic features and choice of lexis, naturally, but it also includes aspects that may seem peripheral and are often harder to analyze. For instance, the genre "reading a bedtime story aloud" has well-known intonation patterns that are

different from ordinary conversation or even from reading many other materials aloud. The genre "heavy metal band name" requires ostentatious typographical styling where possible.

Studies that have been done have focused mainly on morphosyntax, and to a lesser degree on lexis. Various characteristic features have been documented in specific genres. Bednarek (2006) found that while broadsheet newspapers tend to use forms for mitigation and negation, tabloids tend towards statements of emotions and evaluations, with a leaning towards surprise (*dramatically, strikingly*). Hyland (1998) found that science articles make much use of hedges; Hyland (1999) found that while textbooks and research articles both use metadiscourse, they don't use it in entirely the same way: textbooks tend to use more textual metadiscourse (such as logical connectives, frame markers, and evidentials) while research articles tend to use more interpersonal metadiscourse (such as hedges, emphatics, and attitude markers); Hyland and Tse (2005) found that abstracts from articles and theses make much use of *that*-clauses to allow an epistemic stance while referring to the writer's findings. Stotesbury (2003) found that abstracts in different fields use different kinds of stance expressions. For example, abstracts in the humanities tend to use evaluative expressions such as adjectives, nouns, and adjuncts, while abstracts in the natural sciences tend to use modal verbs. MacDonald (2005) found, conversely, that articles on science in popular publications avoid hedges and instead use concrete nouns as sentence subjects and make much use of human narrative. Vilha (1999) found that in popular articles and guidebooks on medical topics, expressions of possibility are much more common than expressions of necessity. Charles (2006) examined the different use of reporting clauses, such as the matrix clause of this sentence, in theses in different disciplines. Bruthiaux (1996) examined newspaper classified ads for different kinds of item, and found different levels of syntactic elaboration: auto and apartment ads have little in the way of syntactic structure, and auto ads are the most collocationally rigid, while personal ads have more creativity in compounding and job ads are more syntactically elaborated. Ferguson (1983) identified several characteristics of sports announcer talk (SAT),

including sentence-initial and copular deletions, inversions, heavy modifiers, and resultatives. Reaser (2003) identified differences between TV and radio commentators on sports, most notably including a much greater frequency of subject deletions in radio description of live action. These studies and many more show us that every variable aspect of language use can be important in the constitution of genre.

The analysis of genre from a literary criticism perspective is clearly top-down, but so to some extent is the analysis of genre and register on the basis of situation or specific isolated features. While institutionalized genres are sensibly divided by institutionalized boundaries (allowing for arguments about the exact location of the boundaries – is Margaret Atwood a science-fiction author or not?), colloquial or functionally emergent genres may be better analyzed from the bottom up. Moreover, even for recognized genres, the actual nature of the genre is not necessarily best analyzed on the basis of a taxonomically expectable set of properties. A thing that is learned by habit and association and intuition – that is to say, a thing that is learned statistically – may perhaps be best analyzed statistically. We may do well to examine a large number of variables and do a multi-factorial regression and, on the basis of that, discern apparent factors that "explain" significant parts of the variation – and then analyze the factors to give them real-world names and explanatory hypotheses. This is what Douglas Biber and his colleagues have done. Biber and Conrad (2009, 56) explain that whereas what they call "register markers" (distinctive usages that are strongly associated with a particular register) and "genre markers" (often present in a specific place in a genre, such as "Dear [name]" at the start of a letter or "Amen" at the end of a prayer) are rare, "register features" ("features that are pervasive and frequent in a register") are more reliable but must be looked at in bulk and statistically, to compare relative frequency between registers. A register feature "might occur to some extent in most (maybe all) registers, but it will be notably frequent in only some registers and comparatively rare in other registers."

In order to discern axes on which registers may truly be distinguished, Biber and his colleagues have run computer analyses of dozens of features – grammatical features such as verb inflections and types of relative clauses, lexical features such as frequency of nouns versus pronouns, semantic features such as types of nouns (abstract, human, etc.) and verbs (mental, activity, etc.) – in hundreds of sample texts in multiple registers (bear in mind their more restricted use of *genre* and broader use of *register* – as Biber (1995, 1) puts it, as "a cover term for any variety associated with particular contexts or purposes"), and run multidimensional statistical analyses on them to discern explanatory factors, i.e., patterns of correlation. "One important point to keep in mind is that the researcher does not decide which features to group together; rather, the statistical analysis identifies the groupings that actually co-occur in texts" (Biber and Conrad 2009, 227). Moreover, "no single parameter or dimension is adequate in itself to capture the full range of variation among registers in a language. Rather, different dimensions are realized by different sets of co-occurring linguistic features, reflecting different functional underpinnings (e.g., interactiveness, planning, informational focus and explicitness)" (Biber 1995, 36).

So, for instance, Biber (1995) surveyed four languages – English, Korean, Somali, and Tuvaluan – looking at multiple genres ("registers") in each. In English, Biber surveyed 9 genres and another 21 sub-genres from a synchronic corpus of 960,000 words as well as diachronic corpora sampling the 17th, 18th, 19th, and 20th centuries, analyzing 67 features in 16 categories (tense and aspect markers; place and time adverbials; pronouns and pro-verbs; questions; nominal forms; passives; stative forms; subordination features; prepositional phrases, adjectives, and adverbs; lexical specificity; lexical classes; modals; specialized verb classes; reduced forms and dispreferred structures; co-ordination; and negation). From the results for these – per feature per genre – a multi-dimensional analysis identified 11 dimensions that accounted for decreasing amounts of the shared variance. The first factor accounted for 26.8% of the shared variance; the second, 8.1%; the third, 5.2%; the fifth, 2.9%; and so on down to the tenth and eleventh, which each

accounted for 1.9% of shared variance (Biber 1995, 120). Biber named this first, most significant dimension "Involved versus Informational Production" and found the following factor loadings (Biber 1995, 142):

Table 1.1: Co-occurring linguistic features in "Involved versus Informational Production" (Biber 1995, 142)

Table 6.1 *Co-occurring linguistic features on English dimension 1: 'Involved versus Informational Production.' (Features in parentheses have lower weights and are not used in the computation of dimension scores)*

Dimension 1	
'Involved Production'	
Positive features:	
Private verbs	0.96
THAT deletion	0.91
Contractions	0.90
Present tense verbs	0.86
Second person pronouns	0.86
DO as pro-verb	0.82
Analytic negation	0.78
Demonstrative pronouns	0.76
General emphatics	0.74
First-person pronouns	0.74
Pronoun IT	0.71
BE as main verb	0.71
Causative subordination	0.66
Discourse particles	0.66
Indefinite pronouns	0.62
General hedges	0.58
Amplifiers	0.56
Sentence relatives	0.55
WH questions	0.52
Possibility modals	0.50
Non-phrasal co-ordination	0.48
WH clauses	0.47
Final prepositions	0.43
(Adverbs)	0.42
'Informational Production'	
Negative features:	
Nouns	−0.80
Word length	−0.58
Prepositions	−0.54
Type–token ratio	−0.54
Attributive adjectives	−0.47
(Place adverbials	−0.42)
(Agentless passives	−0.39)
(Past participial postnominal clauses	−0.38)

He charted the mean scores on this dimension for 23 registers (Biber 1995, 146):

Figure 1.1: "Involved versus Informational Production" across genres (Biber 1995, 146)

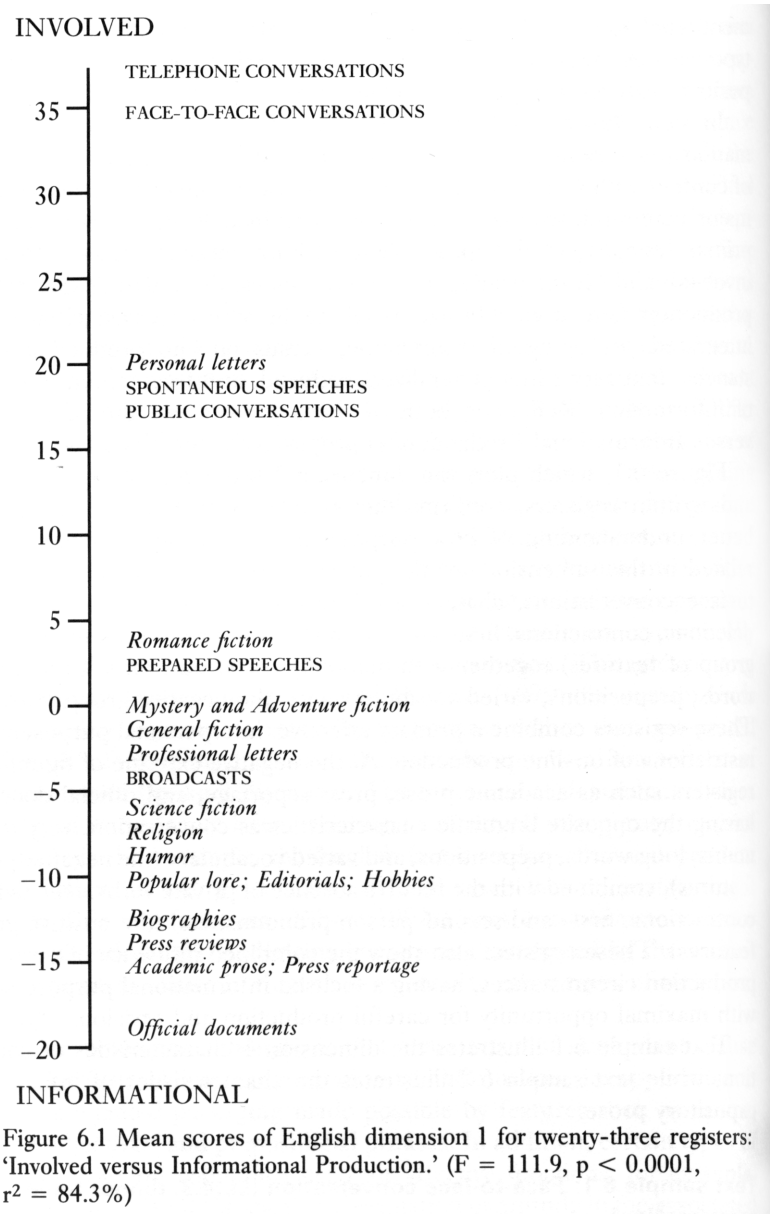


Figure 6.1 Mean scores of English dimension 1 for twenty-three registers: 'Involved versus Informational Production.' ($F = 111.9$, $p < 0.0001$, $r^2 = 84.3\%$)

Another dimension that is likely to be of interest to us is Biber's fifth dimension, "Non-Abstract versus Abstract Style."

Table 1.2: Co-occurring linguistic features in "Non-abstract versus Abstract Style" (Biber 1995, 163)

Table 6.5 *Co-occurring linguistic features on English dimension 5: 'Non-abstract versus Abstract Style.' (Polarity reversed – see note 3)*

Dimension 5	
[No positive features]	
Negative features:	
Conjuncts	−0.48
Agentless passives	−0.43
Past participial adverbial clauses	−0.42
BY-passives	−0.41
Past participial postnominal clauses	−0.40
Other adverbial subordinators	−0.39

Figure 1.2: "Non-abstract versus Abstract Style" across genres (Biber 1995, 165)

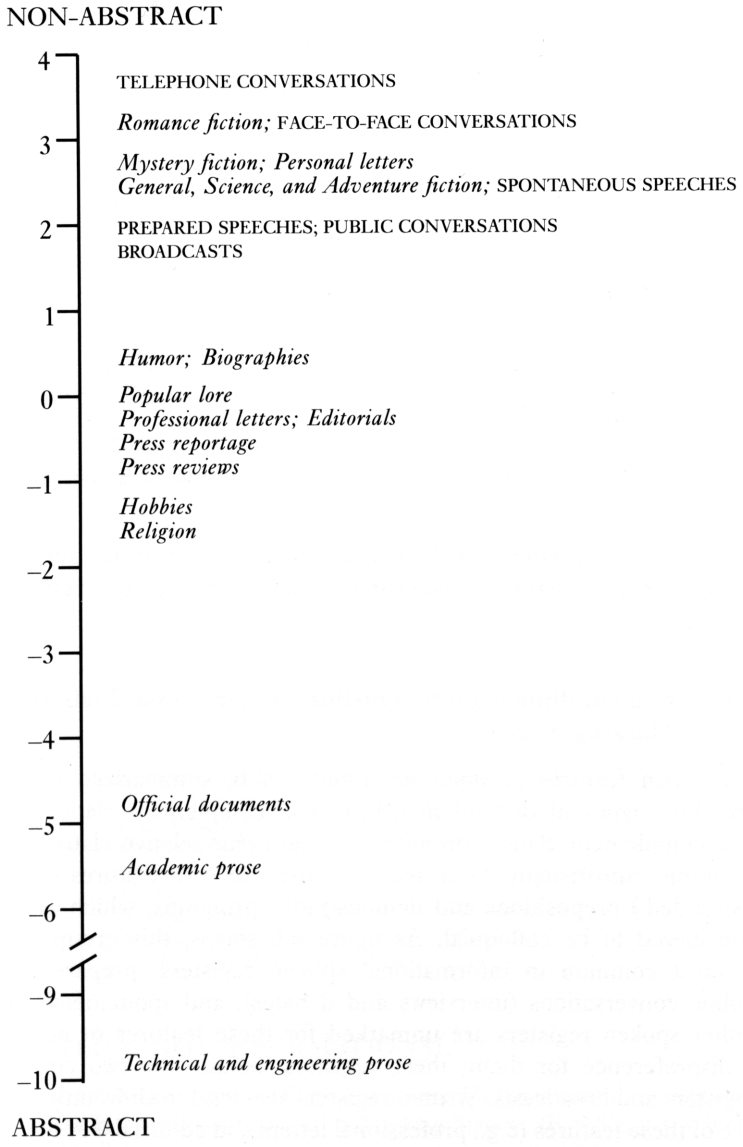


Figure 6.5 Mean scores of English dimension 5 for twenty-four registers: 'Non-abstract versus Abstract style' (Polarity reversed). ($F = 28.8$, $p < 0.0001$, $r^2 = 58.0\%$)

This dimension accounts for much less of the inter-register variation, but it seems quite reasonable that phonaesthemes would be more characteristic of a non-abstract style, so I will be comparing my results with Biber's rankings on this dimension as well as the first dimension.

Biber found some consistencies and some differences between his study languages with regard to the determining dimensions for registers:

> all four languages have dimensions relating to interaction, production circumstances, informational focus, personal stance, and narration. The functional priorities of the languages differ, however. English shows the greatest allocation of resources to distinguishing among various kinds of informational focus, with dimensions relating to production circumstances and argumentation/persuasion being less important but notable. (Biber 1995, 270)

A most important conclusion is that "even when registers are defined at a high level of generality (e.g., conversation, editorials, personal letters), and even when comparisons are across markedly different language families and cultures, parallel registers are indeed more similar cross-linguistically than are disparate registers within a single language" (Biber 1995, 279).

Biber (1986) explores the nature of the differences between spoken and written genres and examines why different studies of these differences have apparently contradictory findings. His examination of various studies shows that researchers use different – and sometimes unclear – definitions of key objects of analysis, such as what constitutes a sentence, and that they often give undue weight to a few factors, a few specific texts, or a few text types (386). His own detailed multi-dimensional analysis shows that there are three key dimensions that account for the lion's share of variation between speech and text: "Interactive vs Edited Text," "Abstract vs Situated Content," and "Reported vs Immediate Style" (410). These dimensions may also be of interest with regard to the use of phonaesthemes in genres.

One question that arises from some analyses is: Is something that exists statistically and intuitively but is not consciously or formally recognized as such a genre? Of course, as soon as we analyze it, we *have* recognized it consciously, but the users

may still not have. Does a genre require official or quasi-official recognition? Is an unrecognized genre in any sense a genre? We have seen that a genre is a grouping of arbitrary size of texts with common features, notably common purposes and means of achieving those purposes. We can also safely say that not just any arbitrary grouping of texts is a genre; some notion of an architext (Genette 1992, 82) prototype (Steen 1999, 111) is surely an essential. But since our linguistic expressions are, as we have seen, guided in good measure by statistical expectations that may not be consciously acknowledged, even architexts or prototypes or similar such schemata may exist without any scholarly hierophant's imprimatur. Should we make a distinction between officialized genres and ones that have not been given an official stamp of existence – sports commentary by veteran broadcasters on the last game before their retirement, perhaps, or personal voice-mail messages beginning with "We have to talk"? One thing to remember in this is that genres, like registers, are pragmatic constructs, existing as they do to serve a specific communicative purpose in a specific context. As Biber (1995, 313) writes, "even though register distinctions have strong linguistic correlates, they are defined on the basis of situational characteristics such as the relations among participants, the production circumstances, and the major purposes and goals of communication." As such, if we find that texts from two genres resemble each other as much, and in the same ways, as they resemble texts from within their genres, we cannot from this necessarily say that they can be constituted as a single genre. We *can*, however, ask why this similarity would be. Are there similar functional demands? Was one of the genres developed in imitation of the other, or were they both developed in imitation of a third?

1.2.3 Historical development of genres

The seminal research in historical development of genres and registers has been done by Douglas Biber and his collaborators. The history of a language's registers and genres is a view on the history of that language: as Biber (1995, 13) says, "linguistic change interacts in complex ways with changing patterns of register variation." We should not just say that it is trivially true and thus uninformative that

register and genre change with language; the existence, emergence, and disappearance of specific registers and genres and the processes involved in their development and change can play important roles beyond those of their components. For example, Biber (1995, 22) notes similarities between languages in their "patterns of evolution following the introduction of written registers." Moreover, he posits that there may be universals of register variation.

Biber and Finegan (1989) consider the diachronic changes in three factors – "Informational versus Involved Production," "Elaborated versus Situation-Dependent Reference," and "Abstract versus Nonabstract Style" – in fiction, essays, and letters from the 17th century to modern times. These factors are emergent from multidimensional analysis rather than imposed a priori. Although the different texts had clearly different values for each factor, they had generally parallel development on each dimension: "17th-century texts are relatively oral; 18th-century texts become more literate in style; and later texts then gradually shift to more oral styles. By the modern period, the three genres are usually considerably more oral than their 17th-century counterparts" (Biber and Finegan 1989, 498–499). This means that "across the four centuries all genres have tended towards more involved, more situated, and less abstract styles" (507). We may well wonder whether this means more phonaesthemes – and indeed we will see the answer in my research results. The reasons for these developments are surely many; Biber and Finegan speculate that "a general preference for rationalism over emotionalism" (512) marking the 17th century was a factor, as was an increase in the use of English rather than Latin for scholarly articles; in the 18th century, "the rise of a popular, middle-class literacy" (513) will have been a likely factor; increasingly democratic tendencies – along with the nationalism of such as Noah Webster – seem likely factors in the 19th century (515).

Another important factor is the relation of written genres to spoken ones. We have already seen that a new genre in a language – whether developed in response to a new circumstance or converted from a register formerly written in another

language (e.g., Latin) – will model itself on existing genres. There is still the question of what factors take precedence in choosing which genre to use as a model. Biber (1995, 288) finds that in English, written genres model on existing written genres rather than spoken ones:

> the early written prose registers – seventeenth-century letters, fiction, essays, and science prose, plus eighteenth-century medical prose and legal opinions – were already quite different from conversational registers shortly after their introduction into English. That is, these written registers did not simply adopt spoken linguistic conventions when they entered English; rather, from the earliest periods these registers developed distinctive linguistic characteristics in response to their differing communicative purposes and production circumstances.

Moreover, "written registers developed to become even more clearly distinguished from spoken registers over the first 100–200 years of their history, although subsequent developments are more complex" (Biber 1995, 288). This would tend to suggest that during that middle phase of their development – for the genres and registers in question, the 1700s and 1800s – they should be less performative or expressive in orientation, less geared towards emulation of overt physical gestures. Biber's diachronic chart of scores on the dimension "Involved versus Informational Production" illustrates this:

Figure 1.3: "Involved versus Informational Production" over time for several genres (Biber 1995, 289)

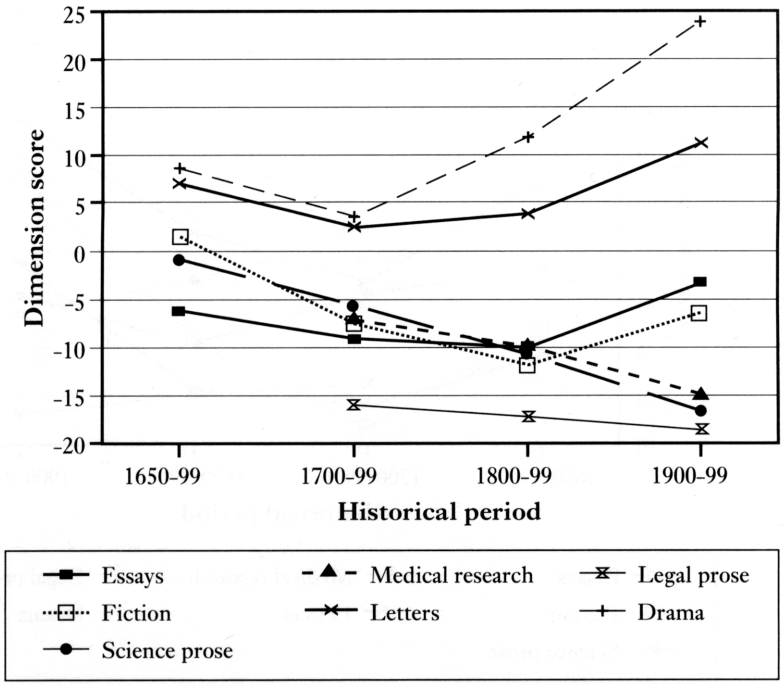


Figure 8.4 Historical change among seven written and speech-based registers along English dimension 1: 'Involved versus Informational Production'

We see that science, medical, and legal prose continue the trend away from involved production, while essays, fiction, letters, and drama swing back towards the involved. On the other hand, the variation is less clear cut for "Non-Abstract versus Abstract Style" – fiction shows a dip, but the others do not:

Figure 1.4: "Non-abstract versus Abstract Style" over time for several genres (Biber 1995, 291)

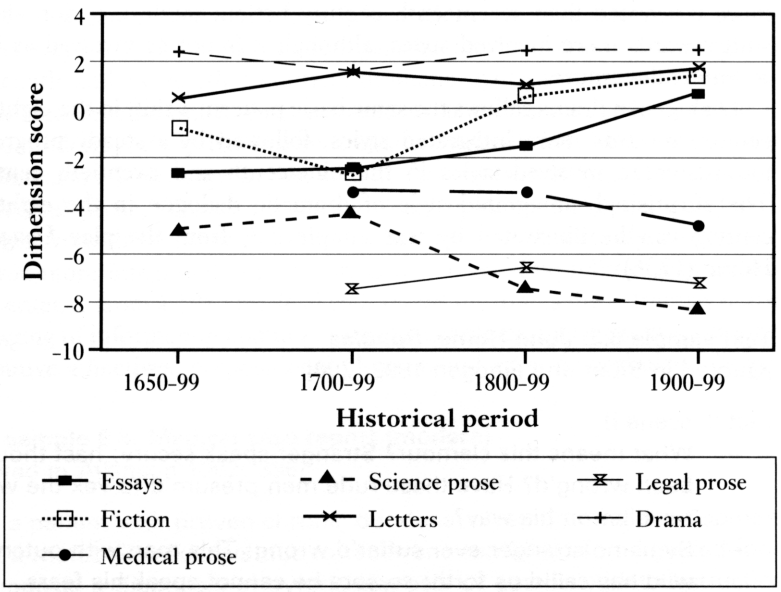


Figure 8.6 Historical change among seven written and speech-based registers along English dimension 5: 'Non-abstract versus Abstract Style'

Various others such as Vande Kopple (1998), Atkinson (1999), and Gross, Harmon, and Reidy (2002) have also found what Biber's research suggests and the casual reader may have noticed: that science writing has changed over the past 3 centuries from personal and involved narratives, with author-centred pronouns and complex embedded clauses, to objective statements making much use of abstracts, passives, and complex noun phrases.

Along with becoming more distinguished from spoken genres, written genres have become more distinct from one another over the centuries. Biber and Conrad (2009) observe that there was very little distinction in noun phrase complexity between written genres in the 17th and 18th centuries. "In fact, it is only in the twentieth century that the specialist informational registers in writing develop highly distinctive non-clausal discourse patterns, with extremely dense use of noun premodifiers and prepositional postmodifiers" (Biber and Conrad 2009, 263). Internal (structural) as well as external (cultural context) factors may play roles:

> This historical change can be attributed to two influences: (1) an increasing need for written prose with dense informational content, associated with the "informational explosion" of recent centuries, and (2) an increasing awareness among writers of the production possibilities of the written mode, permitting extreme manipulation of the text. (Biber and Conrad 2009, 263)

This shift in genre definitions will affect the lexical usage in a given genre, and may affect the word forms used as well – business writing is anecdotally noted for its preference for certain kinds of nominalizations, for example. But genre may also have an effect that spreads to the language as a whole, contributing terms that start within its limited sphere (such as *exit* from drama and *retarded* from clinical literature) and spread throughout the language, often with broadening or shifting of reference. We have noted above the potential feedback effect between genre and choice of word forms, as with the example of *dwarfs* → *dwarves* given by Hock and Joseph (1992, 162). This is an example of systematicity that has had the needs and effects of a particular genre as its attractor or centre of gravity.

In studies such as those looking at sports announcer talk (Ferguson 1983 and Reaser 2003), we have seen that circumstances such as the live and fast-occurring activity encourage certain syntactic choices, such as deletions, and that these choices consequently have the effect of conveying immediacy and live action and so may be used even more where there are fewer other details (such as visible action) to convey the immediacy. We have also seen that choice of lexis feeds into the desired effect (for example in conveying a dramatic tone in tabloid newspapers – Bednarek 2006 – and in distinguishing formal adult-learned language from informal child-learned language – Reilly et al. 2012), and we can anticipate that the association of certain kinds of words with certain registers and genres will have a feedback effect. Thus, a genre that tends to have more ostentatiously performative word forms can be expected, by its existence and form, to invite and even require a more ostentatiously performative, expressive attitude and approach from the speakers and writers, and this will include not only shorter words but more

expressive and performative words. An important task will be to separate the effect of performativity – as in phonaesthematic words – from that of word length, frequency, and age of learning.

To what extent does the performative effect of phonaesthemes bear on genre as opposed to register, as I have chosen to define them for the purposes of this study? Naturally, each genre has its registers, as I have said – within the genre of "newspaper" there are more specific genres such as "tabloid" and within that more specific still such as "advertisement," "sports reportage," "political column," and "breaking news," all of which can also be called registers except that when viewed as genres they allow us to include the structure of the text. We will reasonably want to know the aggregate effect of these various registers and small genres on the greater genre as a whole, but it is truly the aspect of genre that scripts a text. The interpersonal and circumstantial considerations that we examine when we look at register matter, but these are not the same from instant to instant, and they may be structured towards or away from moments of heightened expressivity. Fictional narratives are expected to have climaxes, for instance, and to have different characters who speak differently and exist in different circumstances – and their conversations will have beginnings, middles, and ends. We should not assume that expressivity and "vividness" will be the same in all parts; indeed, even in the narrative sections of a work of fiction, it is reasonable to imagine that word choices may be more oriented to the expressive in parts that typically are meant to be more involving, such as climaxes or opening pages of novels. Genres such as nonfiction magazine articles that may be otherwise similar but lack such peaks of action could be expected to have less use for phonaesthemes; genres that are structured to ostensibly preclude emotional involvement, such as science articles, could be expected to have even less use; on the other hand, genres that give explicit direction for performance, such as play and movie scripts, should be expected to have more use for phonaesthemes, and genres that maintain peak expressivity, such as certain kinds of poetry or song lyrics, could also be expected to have more use for phonaesthemes. Beyond all this, as I have said above, it is genre that determines

audience; all of the registers used within a given genre are used for the same audience and work together to position themselves in relation to the audience, and to position the audience – and the speakers or authors – socially and intellectually.

With all of these factors in mind, it is time we turned to the meat of the matter: the choice of phonaesthemes, of words using those phonaesthemes and words with which to compare them, and of corpora in which to survey usage of these words. This is the subject of Chapter 2, following which we will see the results of the effort and explore their implications in Chapter 3.

Chapter 2: Selection of research materials

2.1 The project

My research question is **"Does the presence of phonaesthemes in words play a role in the constitution and evolution of genres?"** I am approaching this question by comparing rates of usage of phonaesthematic and roughly fungible non-phonaesthematic words in different genres and different time periods. The first steps in this process involved the selection of the study materials: which phonaesthemes to focus on, which lemmas to use as representative of those phonaesthemes, which genres and time periods to compare, and which corpora to use for the study. In this chapter I will detail the processes by which I made those choices, and in the next chapter I will present and analyze the results of the corpus research.

For each step of the material selection, there were important factors to take into consideration. To begin with, I could not survey phonaesthematic words without a clear set of viable phonaesthemes. Indeed, the phonaestheme evaluation and selection step was a necessary gatekeeping step to demonstrate the viability of phonaesthemes as an object of analysis; I can only produce acceptable results if I have phonaesthemes that have some demonstrated reality as such. I aimed to select six phonaesthemes – three onset phonaesthemes and three rime phonaesthemes. With approximately 15 study lemmas per phonaestheme (five each for six sets; see below), this would give 90 lemmas, which would be sufficient for usable results without needing to draw on phonaesthemes that had less firmly demonstrated reality as such.

It would not be possible to survey all words containing all phonaesthemes, nor even necessarily all words containing the target phonaesthemes. In order to have a

suitable basis for comparison, I need to assemble a suitable set of words for comparison for each target phonaestheme: words containing the phonaestheme (I will hereinafter call this set **P**); words similar in sense and usage not containing the phonaestheme (set **S**); and words containing the phonemic cluster but not having similar sense and usage (set **C**). As just mentioned, five lemmas for each phonaestheme and set were expected to be sufficient, and as we will see, more than five for each set can be difficult to come up with for some phonaesthemes.

As I will discuss below, genres presented several factors to consider. The selected genres need to be different enough to give a usable comparison, but genres that are too topic-specific risk giving skewed results. As well, a given corpus must yield enough data to produce statistically significant results, while still being specific enough to give meaningfully distinct results. The selection process was a balancing effort that was also constrained by the corpora available. Once I had made the initial selections and begun the corpus research, I found that further modifications to the selections were necessary in response to the results.

The final results are intended to allow synchronic and diachronic comparisons of comparative rates of usage of phonaesthematic words in different genres. This will give us a view to the interaction between such words and the different genres, helping us to understand not only the natures of the genres but also the nature and function of phonaesthemes.

2.2 Phonaesthemes

An initial challenge in identifying phonaesthemes is that different authors have different criteria – and differently stringent ones at that. No author has put together an exhaustive list of phonaesthemes, but some have assembled fairly good lists of what they consider phonaesthemes. With an eye to my own research purposes and the definition of phonaestheme I am working with – **a phonemic grouping that, within a language, correlates well above chance with a particular semantic quality in etymologically unrelated words** – I assembled an initial list for

consideration from sets mentioned by several authors. I found that the concept of phonaesthemes and the approach to phonaesthematic words underwent some refinement and elaboration over time. Bloomfield (1933, 156) listed words with sounds that, to the speaker, "seem especially well suited to the meaning"; his examples include words beginning with *fl–, gl–,* and *sn–* and ending with *–ump* and *–ack*. Bolinger (1950) gave a number of examples of phonaesthemes, including not just *fl–, gl–, kl–, –ash,* and *–ump* but also sets such as *sp—t, str—p,* and *st—nt* and multi-syllable endings such as *–amble, –usty,* and *–utter*, which expand the definition into full-fledged pseudomorphemes as used in portmanteau words. Subsequently, Bolinger (1968, 219) focused on examples such as *–ump* words, which suggest "heaviness and bluntness," as well as the *–rl* words, which suggest twisting or spiraling, the *–s(t)le* set including *hassle, tussle, bustle,* and *wrestle*, and other looser examples of analogy.

Rhodes and Lawler (1981) presented an in-depth work on the subject in which they detail a theory of what they call *athematic metaphors*, wherein they decompose many words into onset and rime pairs: *ring* into *r–* "non-abrupt onset (of sounds)" and *–ing* "BE/MAKE a sound with an extended envelope" (7); *hump* into *h–* "larger" and *–ump* "3D" (16); *snatch* into *sn–* "quickly" and *–atch* "come to hold" (16); *flap* into *fl–* "2D non-extended" and *–ap* "BE/USE a surface" (16); and several more. For each purported sense they identify several other words with the same phonemic cluster and analogous sense. The presentation of data is engaging but obviously very selective and impressionistic; for example, *sn–* is well known as a phonaestheme relating to the nose, but they also present it as having the sense 'quick'; while their examples of *snap, snag,* and *snip* establish an analogy, some rigorous testing for productiveness would be needed to establish that this is more than coincidence or massaging of the data. However, they provide a usable list of more than two dozen possible phonaesthemes that can be tested for statistical defensibility. Rhodes (1994) expands on Rhodes and Lawler (1981) and adds a few more examples.

Magnus (2001) is a dissertation that focuses on a few phonemic clusters and looks at the various "phonosemantic classifications" possible for them. Her aim is to be exhaustive, and so she gives us a catalogue such as this one for /gl/ (40; formatting hers):

> ***Reflected or Indirect Light*** -- glare, gleam, glim, glimmer, glint, glisten, glister, glitter, gloaming, glow
>
> ***Indirect Use of the Eyes*** -- glance, glaze/d, glimpse, glint
>
> ***Reflecting Surfaces*** -- glacé, glacier, glair, glare, glass, glaze, gloss
>
> ***Other Light or Sight*** -- globe, glower
>
> ***Understanding*** -- glean, glib, glimmer, glimpse
>
> ***Symbols*** -- gloss, glyph
>
> ***Ease*** -- glib, glide, glitter, gloss
>
> ***Slip*** -- glide, glissade
>
> ***Quantities*** -- glob, globe, glut
>
> ***Acquisition/Stickiness*** -- glean, glimmer, glue, gluten, glutton
>
> ***Strike*** -- glance
>
> ***Containers*** -- gland, glove
>
> ***Joy*** -- glad, glee, gloat, glory, glow
>
> ***Unhappiness*** -- gloom, glower, glum
>
> ***Natural Feature*** -- glade, glen

It barely needs mentioning that such a cataloguing is an entirely post-facto exercise and has little if anything to say about the productive potential of these form-sense pairings. But it does provide initial material that can be subjected to testing and analysis.

From the above sources I selected a list of possible phonaesthemes to study. I excluded any that clearly would not produce strong statistical associations – any single-sound onsets, for instance, which show up in far too many different words. I used the definitions given by the authors as my basis for the semantic set that a given phonaestheme associates with. Where multiple definitions for a phonaestheme were given, I included multiple phonaestheme entries. I found, as I proceeded in my work, that in some cases other definitions better covered common characteristics of the set, and I added or substituted those (I indicate this in Table 2.1 with "Harbeck"). I classified the phonaesthemes by phoneme, not by spelling; all possible spellings of a given phonemic set were surveyed. In only one case did I observe difference between a specific spelling and all spellings of the phonemic cluster that was important enough to make a note of; that phonaestheme was not used in my final study set, however.

The purported phonaesthemes that I selected to survey, along with their definitions, the sources for those definitions, and an example word for each, are in Table 2.1.

Table 2.1: Purported phonaesthemes for evaluation

Phonaestheme	Example	Source	Semantic set
bl–	blare	Rhodes 1994	loud, air-induced sound
fl–	flat	Rhodes 1994; Rhodes and Lawler 1981	2-dimensional extended
fl–	flutter	Harbeck	loose motion
fr–	fringe	Harbeck	chaos; excrescence
gl–	glow	Magnus 2001; Bolinger 1950	light
kl–	clang	Rhodes 1994	abrupt onset
kl–	cling	Rhodes and Lawler 1981	together
kr–	crick	Bolinger 1950	bent
kr–	crash	Bloomfield 1933	noisy impact
kr–	crimp	Harbeck	clenching or restriction
pl–	plop	Rhodes 1994	abrupt onset
pl–	plank	Rhodes and Lawler 1981	1 dimensional thick
skr–	scratch	Rhodes 1994	complex onset with white noise component
skr–	scrape	Bloomfield 1933	grating impact or sound
skr–	scrimp	Harbeck	clenching or restriction
skw–	squish	Rhodes and Lawler 1981	compressed
sl–	slick	Bloomfield 1933	smoothly wet
sn–	snout	Rhodes and Lawler 1981	nose
spl–	splash	Rhodes 1994	complex onset with white noise component
spl–	splash	Harbeck	wet and messy
spl–	split	Harbeck	division
spr–	spray	Rhodes 1994	complex onset with white noise component
spr–	sprawl	Harbeck	disarray
str–	string	Rhodes 1994	1-dimensional, flexible
str–	strain	Harbeck	effort or constraint
tw–	twirl	Bolinger 1950; Rhodes and Lawler 1981	twisting motion; rotatory
–ərl	twirl	Bolinger 1968; Harbeck	(Bolinger: spinning or spiralling) Harbeck: circular or curved shape or motion
–æp	flap	Rhodes and Lawler 1981	surface
–æp	clap	Harbeck	sharp sound
–æʃ	crash	Bolinger 1950	hit, fragments
–ætʃ	catch	Rhodes and Lawler 1981	hold, come to hold
–up	loop	Rhodes and Lawler 1981	curve
–ɑp	stop	Rhodes and Lawler 1981; Harbeck	(Rhodes and Lawler: cessation of motion) Harbeck: motion ending abruptly
–sl̩	bustle	Bolinger 1968	frenzied or chaotic action
–ʌmp	lump	Bolinger 1968	"heaviness and bluntness"
–ʌmp	lump	Rhodes and Lawler 1981	3 dimensional solid
–ʌst	dust	Bolinger 1950	surface formation
–ʌst	thrust	Harbeck	force

My approach, starting with purported phonaesthemes, is not the only theoretically possible approach to identifying phonaesthemes statistically, and it has the potential weakness of starting top-down with sets that have been pre-identified on the basis of anecdote and impression (although such native speaker impressions are a common good starting point for linguistic analyses). An approach that started with semantic commonalities and identified phonemic clusters associated with them could in theory also produce results, and might identify further phonemic clusters worth considering. However, an effectively infinite number of semantic sets is possible, and inclusion of exclusion of words in sets can be difficult and sometimes quite arbitrary. A more delimited approach – which would again start top-down with native speaker impressions – would be to take a pre-existing taxonomy and survey according to that; for a previous research paper on this subject, for instance (Harbeck 2014), I used a few specific sets from *Roget's International Thesaurus* (Chapman 1977). An exhaustive survey of all the sets in *Roget* could produce interesting results, and could be done with a well-designed computer program, but such a global fishing expedition would be beyond the scope of this study, although it could make a considerable separate project of its own. Similarly, a survey of all possible phonemic clusters is theoretically possible but would be far surplus to requirements and would be unlikely to produce more usable results than my present approach.

It may be noticed that the definition of *phonaestheme* I am using makes no mention of expressivity or performativity. However, when we look at the phonaesthemes listed for examination, we can see that they all have notable expressive potential. We have already seen (with reference to Reilly et al. 2012, among others) that the systematicity that gives rise to phonaesthemes operates most strongly in shorter and more concrete words. We have also seen that phonaesthemes operate in the same realm as sound symbolism, onomatopoeia, and ideophones, which is to say they have a sense of iconic expressivity. Thus we will expect phonaesthematic words to be used more in genres that are more expressive, demonstrative, involved, and concrete, and less in genres that are more abstract and detached. As well, given

their generally shorter length and greater concreteness, and the unrestrained tone that comes with more direct iconicity or demonstrativeness, we may also expect that they will be more associated with less formal or élite genres and registers.

For each purported phonaestheme, I needed to survey all etymologically unrelated roots using that phonaestheme. I determined that proper nouns and words not in current use should be excluded. Various viable ways of determining a cutoff for what is in current use are available. I chose to include all those words that were to be found in a specific dictionary, although in a few instances I included a word not found in the reference dictionary that I knew to be in current use. Given that I was aiming to choose phonaesthemes that had very clear statistical basis and had already been identified as such by others, I felt that this approach was well justified and optimally efficient. The dictionary I chose was the *Oxford Concise Dictionary of English Etymology* (Hoad 1996). This had the advantage of allowing me to identify etymologically related roots at the same time as I constructed the word set. As the dictionary is a print edition, I used the search function on the electronic *Oxford English Dictionary* (2015) to find all words ending in the rime phonaesthemes; the set of words in Hoad 1996 is a subset of the set in the *Oxford English Dictionary*, so I needed only to confirm the presence of each word from the search results in the etymological dictionary and I could be assured that none were left out. In the case of words that were etymologically related, I chose the word that I deemed most basic or representative and removed the others from consideration.

As already mentioned, inclusion of a word in a particular semantic common set can sometimes be a judgement call. Thus, rather than simply counting words as "in" or "out," I scored words as 1 if they were definitely in the semantic set, 0.5 if they were more loosely related to it, and 0 if they were definitely not in it. In this way I produced an absolute score as the total score for each set of words beginning or ending with an identified possibly phonaesthematic cluster, and I calculated a relative score by dividing the absolute score by the total number of words scored.

The full sets of words and scores are included as Appendix A. The relative scores are listed in Table 2.2.

A major issue that we can see with phonaesthemes is that the semantic commonality they purportedly express is sometimes quite loosely defined. An objection may quite fairly be made that for any random set of words a semantic classification could be loosely made that would capture a substantial portion of them. I therefore included in my survey two sets of 50 words chosen at random and discerned in each set a semantic commonality on the level of those given for phonaesthemes, and used this as a basis for determining which phonaesthemes could defensibly be said to exist at a level greater than chance or tendentious analysis. To choose these sets of words, I used the online random number generator random.org to generate two mutually exclusive sets of 50 numbers from the set of all page numbers in Hoad 1996 and I used the first word on each page thus chosen.

Table 2.2: Results of phonaestheme survey and scoring

Phonaestheme	Relative score	Semantic set
bl–	0.08	loud, air-induced sound
fl–	0.15	2-dimensional extended
fl–	0.17	loose motion
fr–	0.25	chaos; excrescence
gl–	0.27	light
kl–	0.19	abrupt onset
kl–	0.22	together
kr–	0.08	bent
kr–	0.06	noisy impact
kr–	0.09	clenching or restriction
pl–	0.05	abrupt onset
pl–	0.02	1 dimensional thick
skr–	0.18	complex onset with white noise component
skr–	0.19	grating impact or sound
skr–	0.31	clenching or restriction
skw–	0.20	compressed
sl–	0.23	smoothly wet
sn–	0.36	nose
spl–	0.38	complex onset with white noise component
spl–	0.46	wet and messy
spl–	0.42	division
spr–	0.22	complex onset with white noise component
spr–	0.44	disarray
str–	0.26	1 dimensional, flexible
str–	0.27	effort or constraint
tw–	0.21	twisting motion; rotatory
–ərl	0.5	circular or curved shape or motion
–æp	0.33	surface
–æp	0.24	sharp sound
–æʃ	0.41	hit, fragments
–ætʃ	0.24	hold, come to hold
–up	0.2	curve
–ɑp	0.36	motion ending abruptly
–sl̩	0.17	frenzied or chaotic action
–ʌmp	0.46	"heaviness and bluntness"
–ʌmp	0.31	3 dimensional solid
–ʌst	0.21	surface formation
–ʌst	0.21	force
random 1	0.23	type of person
random 2	0.12	resembling white fabric

We can see that the random control groups set a bar too high for many of the purported phonaesthemes to clear. By this I am not saying that there is no

psychological reality to phonaesthemes that failed to score higher than the control groups; as the intersection of two sets – a phonemic set and a semantic set – they may well have sufficient presence to have a systematic effect and to be used productively. Experiments by such as Bergen (2004), Magnus (2001), and Abelin (1999) have indicated as much. But for an effort such as the present one, it is important that the results be clearly defensible as more than just an artifact of analysis. It is also worth noting that the semantic set for the control group with the higher score – "type of person" – is among the broadest, loosest, and least expressive of all the sets. I intentionally made the broadest reasonable set I could in order to set the bar high, so as to make the choice of phonaesthemes as defensibly stringent as possible.

Choice of which phonaesthemes to focus on was conditioned not only by which had the highest relative score. In some cases the absolute score (reflecting the number of current words having the phonaestheme) was so high that even though the relative score (portion of all words containing the phonemic cluster) was only slightly above the control group it was still worth considering the phonaestheme. Another important criterion was the availability of specific words (lemmas) to study as representative of each phonaestheme. I could only select phonaesthemes that would give five usable words containing the phonaestheme, the same number of roughly equivalent (semantically similar and syntactically substitutable) words not containing the phonaestheme, and the same or a similar number of words containing the phonemic cluster that did not have the semantic commonality. This last group proved the most difficult and, as we will see, ultimately the least valuable in the final results. The final choice of phonaesthemes was thus done partly in conjunction with the choice of lemmas. I will go into further detail about the choice of lemmas in the next section, but here, to complete this section, is the list of phonaesthemes that were finally chosen:

gl–

sn–

spl–/spr–

–æʃ

–ərl

–ʌmp

You will notice that the *spl–* and *spr–* onsets are treated as one group. This is just for the sake of having enough lemmas in total to study; it is not an assertion that they are in fact one group – the senses are similar but distinct.

2.3 Lemmas

The phoneme selection exercise provided full lists of words that included the phonaesthemes, so the initial work for the second phase was accomplished in the first phase. In order to properly compare the phonaesthematic words with non-phonaesthematic words, however, suitable non-phonaesthematic words needed to be selected. These words would be of similar sense and part of speech (noun, verb, adjective) and ideally of similar frequency in use. I identified potential words with the aid of various thesauruses (including Chapman 1977), and then I consulted the Corpus of Contemporary American English (COCA) (Davies 2008) for total counts of each word in order to have a sense of the words' frequency in use.

I also consulted the *Oxford English Dictionary* to find the date of first citation of each word under consideration, so that I could be reasonably assured of having words that would be usable in multiple time periods. This did not guarantee equal rates of use over all time periods, of course, but means of estimating total use over time (such as Google ngrams) are still of imperfect reliability, and, given that usage of "involved" and "non-abstract" language in various genres is known to have dipped in the 1700s to 1800s (see Biber 1995), we may reasonably expect usage of phonaesthematic terms in general to have been less, so applying a compensation factor could obscure one of the very effects I am hoping to discern.

The list of possible phonaesthematic words includes many that can function as multiple parts of speech. As my research project here is focused on a phonological-semantic effect rather than a syntactic one, I did not see any reason to focus on one part of speech or to do separate counts for different parts of speech. (Indeed, as we will see, in order to have numbers at a reasonable level of statistical significance, and to compensate for subject-specific effects, I did not break out data for individual lemmas in the final analysis either.) This also made it possible to use databases other than fully parsed corpora, thereby broadening my available sets of data. I did find it necessary in some cases to use multiple derived forms of the same root in the **S** (semantically related non-phonaesthematic) set in order to match all parts of speech for a corresponding phonaesthematic word.

The **C** set, words that contained the target phonemic cluster but did not have the semantic commonality, was the most difficult set to assemble. Since I chose phonaesthemes with high relative scores – as high as 0.50 in one case – and since some words had individual scores of 0.5, meaning the total number of at least partly related words was greater than the absolute score for the phonaestheme, for some of the phonaesthemes the clear majority of the words available were at least weakly related to the semantic commonality, leaving a rather smaller set of control words. As well, some phonaesthemes have multiple semantic valences: *spl*– can have a phonaesthemic sense of 'division' as well as having the target sense of 'wet and messy', for instance, and while a word such as *split* would not be part of the study phonaestheme, it would be arguably phonaesthematic and so not usable as a control word. The set of possibly usable words in the end included a number that would not be as likely to be seen in the same target texts, and a few that turned out to be drastically overrepresented in certain genres. I am including the **C** set of phonologically but not semantically similar words in the study results for completeness, but I will say in advance that the constraints on the set were such that the results for this set were less usable or reliable and so were left out of the majority of the analysis. An entirely different experiment may produce more usable results for discerning any phonaesthemic spreading effect – i.e., an effect whereby

words that have a phonaesthematic cluster are treated as having a similar tone even though they do not share the sense.

The initial set of study lemmas (represented by their citation form) is in Table 2.3.

Table 2.3: Initial set of study lemmas

Set	P	Freq	First	C	Freq	First	S	Freq	First
gl–	glow	11078	1000	glove	11632	1000	burn	35013	1000
	glare	7255	1250	glue	5883	1340	shine	12245	1000
	gleam	2779	1000	gland	2619	1631	scowl	2363	1340
	glisten	1667	1000	glucose	1514	1840	luster	985	1591
	glower	723	1513	glade	739	1522	radiant	929	1450
gl– totals		23502	1152.6		22387	1466.6		51535	1276.2
sn–	snort	3369	1366	snack	6312	1402	inhale	3439	1725
	snore	1986	1330	snail	1705	1000	nasal	3081	1425
	sneeze	1491	1493	snare	1170	1100	exhale	2504	1400
	snout	897	1220	snipe	737	1325	beak	1493	1220
	snivel	90	1325	snooker	122	1889	cavil	45	1548
sn– totals		7833	1346.8		10046	1343.2		10562	1463.6
spl–, spr–	spread	39312	1200	splendid	2597	1624	expand	26822	1475
	spray	12633	1626	spruce	2062	1378	wet	18837	1000
	sprinkle	7655	1400	sprig	1922	1359	scatter	8671	1154
	splash	6222	1699	spleen	488	1300	dampen	1407	1633
	splay	671	1330	splanchnic	23	1694	diverge	950	1665
spl–, spr– totals		66493	1451		7092	1471		56687	1385.4
–æʃ	crash	23120	1400	ash	8038	1000	slap	9680	1632
	splash	6222	1699	rash	2550	1000	collide/ collision	7093	1621
	slash	5036	1576	stash	2490	1794	immerse/ immersion	3715	1605
	mash	1767	1000	hash	1191	1655	pulp	2244	1400
	thrash	1567	1000	sash	823	1599	sever	2106	1382
–æʃ totals		37712	1335		15092	1409.6		24838	1528
–ərl	curl	9904	1447	pearl	9181	1340	curve	11925	1560
	swirl	5582	1425	earl	6398	1000	spiral	6183	1556
	whirl	3212	1290	squirrel	3321	1366	vortex	963	1653
	twirl	1740	1598	hurl	2974	1300	gyre/gyrate	382	1420
–ərl totals		20438	1440		21874	1251.5		19453	1547.25
–ʌmp	dump	12247	1300	jump	39367	1511	cluster	13570	1000
	slump	5301	1677	pump	17676	1420	knot	6693	1000
	clump	2782	1586	trump	5305	1555	ditch	4588	1045
	hump	1395	1708	chump	304	1680	subside	2085	1616
	rump	668	1425	sump	230	1425	backside	1052	1489
–ʌmp totals		22393	1539.2		62882	1518.2		27988	1230

As stated above, the sets are named as follows: **P** means they have both phonemic and semantic matching; **C** means they have the phonemic group but not the sense; **S** means they have the sense but not the phonemic group. **Freq** is the number of hits for that word in COCA (Davies 2008). **First** is the date of the earliest citation for that word in the *Oxford English Dictionary*; dates that are before AD 1000 (many of which are listed in *Oxford* simply as "OE") are given as 1000. Beneath each phonaestheme set is a total of the number of hits as well as an average of the first citation dates from the *OED*, which is included just for the sake of general comparison between the sets. One word, *splash*, is present in two sets, since it contains two of the study phonaesthemes. In the actual analysis it was counted only once.

It will be noted that two of the **C** set, *glucose* and *snooker*, were found to have their first citations in the 1800s. I initially included them for want of suitable others, with the idea that I might use them only in the most recent set of results. Those two were ultimately found to be problematic for other reasons as well (clear over- or under-representation in certain topic areas; for example, *snooker* is much used in British newspapers, which often report on snooker tournaments) and so they were excluded in the final analysis.

In general, over the course of the corpus research that is the meat of this thesis, certain lemmas were found at length to be of such slight or uneven representation as to be better excluded. A few were found to be over- or under-represented in certain genres for reasons unrelated to phonaesthematic considerations, including the two examples noted above. A few were found to give problematic results due to their appearance in such things as proper nouns, which are not subject to the same tone-based discretion as other words are (for example, if you are writing about a Mr. Burns, you cannot avoid using the word *Burns* even though nothing related to fire may be involved); it was found that such problematic results could not always be reliably excluded from the total count. Thus, a number of the initial set of words were removed from consideration in the final analysis, and were not surveyed in those genres that were surveyed last. As mentioned above, I also observed that

certain of the initial **S** set did not have a truly equivalent syntactic ambit to the phonaesthematic words, so I added derived forms and included their counts in the total. The final sets of words used for analysis is as follows; words excluded from the final calculation are in italics:

Table 2.4: Final set of study lemmas

P	**C**	**S**
glow	*glove*	*burn*
glare	glue	shine
gleam	*gland*	scowl
glisten	*glucose*	LUSTER
glower	glade	RADIANT/RADIATE
snort	*snack*	INHALE/INHALATION
snore	snail	NASAL/NASALITY
sneeze	snare	EXHALE/EXHALATION
snout	snipe	beak
snivel	*snooker*	CAVIL
spread	*splendid/splendor*	EXPAND/EXPANSION
spray	spruce	wet
sprinkle	*sprig*	**scatter**
splash	spleen	*dampen*
splay	*splanchnic*	DIVERGE
crash	ash	slap
splash	*rash*	COLLIDE/COLLISION
slash	*stash*	IMMERSE/IMMERSION
mash	hash	pulp
thrash	sash	SEVER
curl	*pearl*	curve
swirl	*earl*	SPIRAL
whirl	*squirrel*	VORTEX
twirl	hurl	*gyre/gyrate*
dump	*jump*	**cluster**
slump	pump	knot
clump	*trump*	ditch
hump	*chump*	SUBSIDE
rump	*sump*	**backside**

Splash is repeated, so one instance of it is italicized and excluded, but the other instance is retained. The ALL CAPITALS and **bolding** are motivated by a further factor

that I have mentioned in my literature review section: that one-syllable Germanic words have a known general greater tone of concreteness of sense on average and are typically seen as more basic, while polysyllabic words of classical origin are more associated with abstract and more formal usage. The **S** set includes words of both types: monosyllabic Germanic and polysyllabic classical. It also includes three of polysyllabic Germanic origin (*scatter*, *cluster*, *backside*). The **P** set is almost exclusively monosyllabic Germanic, a noteworthy finding in itself, as I did not intentionally restrict the set to such words. As I detail in chapter 3, I thus made multiple analyses for the sake of comparison: one comparing the **P** set to the entire **S** set, and one comparing it only to that subset of **S** that is monosyllabic and Germanic control words (i.e., excluding the polysyllabic classical words – the polysyllabic Germanic words were left in the control group, since there were two polysyllabic Germanic words in the phonaesthematic group). I will call the subset of **S** composed of the polysyllabic classical words (all coming from Romance languages, in this case) $\mathbf{S_R}$, and the Germanic remainder $\mathbf{S_G}$. In the table above, words in ALL CAPITALS are the the $\mathbf{S_R}$ subset. The **bolded words** are the polysyllabic Germanic words, which are included with the $\mathbf{S_G}$ set and the **P** set.

In the research, all common spellings of a lemma were used, for example *splendour* as well as *splendor* and *lustre* as well as *luster*. Inflected forms were counted as well: conjugations of verbs and plurals of nouns. The full set of forms surveyed is included in Appendix B. Well-made online corpora such as COCA and the British National Corpus (Davies 2004) allow search by lemma, which includes all inflected forms (I surveyed a sample of the results to confirm this, and occasionally re-searched an inflected form to confirm the results for it). Other corpora that have been assembled as searchable sets of text files, such as the ZEN corpus (Fries et al. 2004) and ad-hoc corpora assembled by me from Project Gutenberg (2015), were searched using Adobe Dreamweaver with regex variables to capture all inflected forms, and the results were visually scanned as a second failsafe.

2.4 Genres and corpora

I determined as a basic condition of this project that I would survey several genres over multiple time periods and would consider American and British corpora separately. American and British usage can be quite distinct in some cases, and I felt that the differences were worth observing and documenting, especially since corpora are available for both countries for at least some genres. I initially set the time periods to be circa 2000, circa 1900, and circa 1800, in each case a span of ±10 years. I considered the possibility of adding circa 1700 and circa 1600, but I found that usable and representative corpora were not in ample evidence, and that many of the study lemmas were not in common use in those time periods. The greater variation in spelling was also a risk factor – although modern-spelling editions of such things as the works as Shakespeare are available, other genres were not so forthcoming and would have had to be surveyed much more exhaustively and carefully. Moreover, many modern genres simply did not exist in those time periods. Novels per se were not a genre at the time of Shakespeare, nor were newspapers (as well, the total word count of the ZEN corpus of early English newspapers for its earliest time periods was insufficient to produce usable results).

Several factors conditioned my choice of genres to survey. I was limited to genres with corpora that were large enough to give statistically significant results over multiple time periods and in two countries and were not forbiddingly time consuming to assemble. Given that I was surveying lemmas, and in many cases low-frequency ones, rather than word types or syntactic construction, I needed much larger corpora for usable results than would be required for many other kinds of projects, even such detailed work as done by Biber (1995). At the same time, I needed the genres to be specific enough to give meaningful results, and yet general enough for those results to be usable and for a proper diachronic survey to be feasible – genres come and go over time, and some genres (for example poetry) have stylistically distinct sub-genres that shift in prevalence over time, muddying the results. I needed a representative set that would have distinctly different tones and stylistic approaches, but I also had to limit myself to genres that would not

automatically exclude some of my study lemmas or overuse others for reasons of subject matter. (Items in the such as *snooker* and *earl* proved to be overrepresented by orders of magnitude in British newspapers and British parliamentary speech, respectively, and were consequently removed from the final results.) Beyond all this, I wanted to match them as well as I could to registers studied in Biber (1995) so as to be able to compare usage of phonaesthemes with certain dimensions of variation as measured by Biber. Biber (1995, 87) divided the registers into written (Press reportage, Editorials, Press reviews, Religion, Skills and hobbies, Popular lore, Biographies, Official documents, Academic prose, General fiction, Mystery fiction, Science fiction, Adventure fiction, Romantic fiction, Humor, Personal letters, Professional letters) and spoken (Face-to-face conversation, Telephone conversation, Public conversations, debates, and interviews, Broadcast, Spontaneous speeches, Planned speeches). I did not plan to include any spoken genres, simply because they could not be surveyed reliably (if at all) in earlier time periods. As we will see, I did ultimately include British Hansard, which is a parliamentary record, but is composed principally of planned speeches, which are as much a literary as an oral genre.

I thus created an initial set of desirable genres and time periods, and then determined what was available in corpora to which I had access or which I could assemble. As I proceeded with the data collection from the corpora, I found that some that had seemed viable were not, and I came to consider others for addition. I found that some chronotopes (e.g., US 1800) were not accessible for some genres, and that the data for others forced a broader time period – several decades instead of two. Table 2.5 shows the genres I chose and what corpus I used for each genre, place, and time. In the following sections I explain the choices of genres and corpora I used for them, as well as what genres I discarded.

Table 2.5: Genres chosen and corpora used

Genre	1800 UK	1800 US	1900 UK	1900 US	2000 UK	2000 US
Fiction	CLMET subset (see 2.4.1) *1782–1826*	COHA: FICTION *1810s+1820s*	CLMET subset (see 2.4.1) *1890–1910*; also detective fiction from Project Gutenberg (see 2.4.1) *1841–1922*	COHA: FICTION *1890s+1900s*	BYU-BNC: w_fict_prose *1980–1993*	COCA FIC: Gen (book) *1990–2015*
Drama	CLMET subset (see 2.4.2) *1766–1835*					COCA: FIC: Movies *1990–2015*
Magazines		COHA: MAGAZINE *1810s+1820s*		COHA: MAGAZINE *1890s+1900s*	BYU-BNC: MAGAZINE *1980–1993*	COHA: MAGAZINE *1990–2009*
Newspapers	ZEN *1701–1791*			COHA: NEWSPAPER *1890s+1900s*	BYU-BNC: NEWSPAPER (also w_news_ tabld for tabloid subset; see 2.4.4) *1980–1993*	COCA: NEWSPAPER *1990–2015*
Humanities articles			CLMET subset (see 2.4.5) *1884–1920*		BYU-BNC: w_ac_hum_ arts *1980–1993*	COCA: ACAD: History + Humanities + Phil/Rel *1990–2015*
Hansard	Hansard *1810s*		Hansard *1900s*		Hansard *1990s*	

Legend:

CLMET: Corpus of Late Modern English Texts, version 3.0 (Diller, De Smet, and Tyrkkö 2011); COHA: Corpus of Historical American English (Davies 2010); BYU-BNC: British National Corpus, Brigham Young University interface (Davies 2004); COCA: Corpus of Contemporary American English (Davies 2008); ZEN: Zurich English Newspaper Corpus (Fries et al. 2004); Hansard: Hansard Corpus (Alexander and Davies 2015). Specific genre subsets are given as applicable after the corpus name, using the identifier used in the corpus.

2.4.1 Fiction

Fiction is a very easy genre to get data for. However, it is also a very broad genre. I was faced with a decision of whether to survey sub-genres of fiction or just to survey fiction as a whole. I found that my choice was constrained by the corpora available and their divisions and subdivisions, as well as by the existence or non-existence of certain kinds of fiction in certain time periods. For example, although

ample science fiction texts are available digitally for the 2000 time period, the genre was invented not much more than 100 years ago and was not enough in evidence at that time to be usable for data.

For the US 2000 set, I used the Corpus of Contemporary American English (COCA) (Davies 2008), a thoroughly parsed corpus of 520 million words from 1990 to the present. I chose to use the full time frame of the corpus for maximum statistical power as well as to avoid errors possible when specifying multiple search criteria. The corpus has an overall fiction genre, as well as subdivisions for book, journal, sci-fi/fantasy, and movies. I used the overall fiction genre.

For the UK 2000 set, I used BYU-BNC (Davies 2004), an interface created by Mark Davies of Brigham Young University for the British National Corpus; it contains 100 million words from the 1980s through 1993. This is slightly earlier than the COCA time period, but not far enough removed in time to be a major concern in the results. Its fiction section is subdivided into drama, poetry, and prose. I used the prose fiction genre and surveyed the entire time period.

For the US 1900 set, I used the Corpus of Historical American English (COHA) (Davies 2010), a corpus of 400 million words covering 1810 to 2009. It has a fiction genre with no subgenre divisions. I surveyed the fiction genre for the decades of the 1890s and 1900s (i.e., from 1890 to 1909).

For the UK 1900 set, I used the Corpus of Late Modern English Texts (CLMET), version 3.0 (Diller, De Smet, and Tyrkkö 2011), a corpus of 34 million words covering 1710 to 1920. This corpus is a downloadable corpus of text files with an index spreadsheet that includes the full list of works with a genre classification and the date of publication for each. I included all works classified as narrative fiction with publication dates from 1890 through 1910. This included 32 texts with a total word count of 3,200,050.

For the US 1800 set, I used COHA, fiction, 1810 through 1829. I couldn't use earlier dates, as the corpus starts with 1810.

For the UK 1800 set, I used CLMET, narrative fiction published from 1782 through 1826. It was necessary to expand the time frame in order to have a usable amount of data. The sample consisted of 18 texts with a total word count of 2,128,602.

For collateral information, I decided to investigate whether the first 1000 words of an action-oriented genre of fiction might use a higher proportion of phonaesthematic words than the entire body, given the types of opening sequences such works are often prone to. I was not able to restrict results from COCA, BYU-BNC, or COHA to the first 1000 words of works, so I used Project Gutenberg to compile a corpus of detective novels from the decades around 1900, and from that I created a corpus of the first 1000 words of each work.

The fiction genre subsumes Biber's five fiction registers, which, however, are for the most part fairly clustered in his results: general, mystery, sci-fi, adventure, and romantic.

2.4.2 Drama

Drama would seem to be a good genre for comparison due to its greater performativity, which might seem to call for more performative words. I found that BYU-BNC had a "drama" sub-genre of fiction, and COCA had a "movie" sub-genre; as well, CLMET had a set of plays, from which I could draw 31 dating from 1766 to 1835 with a total of 547,595 words. Movies are naturally not the same as stage plays, but I felt an instructive comparison could be made, and the question of performativity was sufficiently motivating to lead me to include these corpora. In the data gathering, however, I found that the results for the BYU-BNC corpus were insufficient – for example, there were 0 tokens of my 6 lemmas for the *gl*–phonaestheme – so I had to exclude that corpus. The remaining results are two widely separated and distinctly different genres, but both oriented to performance. I therefore decided to include those results for the value of the comparison they might bring.

2.4.3 Magazines

The genre of magazines is another very broad genre containing diverse sub-genres. It is present in COHA as a genre, but is not subdivided. It is subdivided in COCA, but I would not be able to compare the results across time periods. I thus chose to use COHA over all three time periods for the US. I surveyed three time periods: 1990–2009, 1890–1909, and 1810–1829. Although the BYU-BNC corpus has a magazine genre, I found that it was not suitable for use in the comparison.

This genre does not have a clear match in Biber's registers.

2.4.4 Newspapers

It would have been interesting to use individual newspapers of record, such as *The New York Times* and *The Times* of London, but I found that none of the ones I considered had usable search engines for my purpose, and most did not have searchable databases covering more than the most recent years. However, I was able to survey general corpora of newspapers for four chronotopes.

For the US 2000 set, I used COCA. It subdivides news into several sub-genres, but I did not make use of these because the other corpora used for this genre did not have the same subdivisions and because the results would not have been as reliable.

For the UK 2000 set, I used BYU-BNC. It, too, has multiple sub-genres, but for the same reasons as with COCA I did not use the sub-genres. However, given that the British newspapers include a tabloid genre not present in American newspapers that could make a difference in the results, I also surveyed that for a separate comparison to evaluate its effects on the overall British results.

For the US 1900 set, I used COHA, 1890–1909. It did not have sub-genres.

For the UK 1800 set, I used the ZEN (Zurich English Newspaper Corpus) database (Fries et al. 2004). The ZEN corpus covers early English newspapers published between 1661 and 1791, from the early issues of The London Gazette up to the period of the first publication of The Times. It consists of 349 complete newspaper

issues containing 1.6 million words. The corpus has one set per decade (for the years 1661, 1671, 1681, and so on through 1791). I used the entire 1701–1791 set. This is obviously both earlier and broader than the other samples; however, it was necessary to use the whole century in order to get statistically significant results. The comparison is thus not fine-grained, but as we will see in chapter 3, the results are of interest.

This genre subsumes Biber's registers of press reportage, editorials, and press reviews, as well as several others that may be found in newspapers; the specific set of registers varies from newspaper to newspaper and between time periods. The variation we will see in the results will thus mirror the varying generic composition of newspapers, not just varying tone in a specific register.

2.4.5 Academic articles in the humanities

This genre is one where Biber's register may be the broader. His delimitation is simply "academic prose." I chose to limit it to a more specific field, not just because I had the means to do so, and not just because the composition of a more general "academic" genre would have been inconsistent from corpus to corpus, making for a more problematic comparison, but also because I had hoped to include one or more other academic genres for the sake of comparison. As I detail in section 2.4.7, however, other genres turned out not to be viable for the purpose: some had insufficient data, some were not well defined, and some were too topic-specific in vocabulary and would have produced skewed results for reasons other than the question of phonaesthemes.

For the US 2000 set, I used COCA, limiting to the academic genre, specifically the subsets history, humanities, and philosophy/religion.

For the UK 2000 set, I used BYU-BNC, academic subset humanities-arts.

For the UK 1900 set, I used a subset of CLMET, 11 texts dating from 1884 to 1920, with a total of 757,034 words. The works are mostly philosophical works by such

authors as Russell, Whitehead, and Chesterton. I was not able to find or assemble suitable corpora for this genre for the other three chronotopes.

2.4.6 Hansard

In an earlier research paper (Harbeck 2014), I found that phonaesthematic words were much less common in legislative texts than in the language overall or in sci-fi fan fiction. For that effort, I used the Congress.gov database of legislation. However, that database covers only the most recent years, and equivalent databases are not available for earlier times or for the UK. Available historical records were of insufficient volume. Records of the judgements of the Supreme Court of the United States are available online; I assembled a 20-year span from 1790 to 1809, but I found that the data were insufficient, to the point where up to a full century's worth of judgements would be needed, and it was evident that although Supreme Court judgements cover many matters, certain subject areas (for example, shipping and property ownership) were heavily overrepresented, while other areas of detail were underrepresented.

However, a very substantial corpus does exist of the Hansard of the Parliament of the United Kingdom (Alexander and Davies 2015); it contains 1.6 billion words from 1803 to 2005. It is the record of speeches, debates, and pronouncements in Parliament. Such discourse does of course have some limitations of subject matter and treatment that can be expected to skew it slightly, but the limitations are not nearly as severe as for the legal judgements (indeed, I found that words such as *glow* and *glare* are at times more frequent in Hansard than in newspapers), and the historical development of the genre is of interest as well.

A similar usable set of records is not currently available for the United States. The Canadian government has a database of its Hansard from recent years, and the *Canadian Encyclopedia* maintains a database of the historical Canadian Hansard to the late 1800s, but I found that these were not properly searchable in the way required for my purposes. Therefore, I have included only the British Hansard.

The Hansard genre can be expected to correspond reasonably well with Biber's "planned speeches" genre.

2.4.7 Genres discarded

Medical articles

Both COCA and BYU-BNC have "medicine" sub-genres of the academic genre. I began a survey of these corpora but found that the subject matter caused terms such as *spread, spray, rash,* and *inhalation* to be strikingly overrepresented. I could not exclude them without make the results incomparable, and I could not exclude all such terms from all result sets without making the results much less viable. Thus I discarded the results.

Philosophical essays

This sub-genre of academic essays was at first my preference over the broader humanities genre. However, I found that the data available to me at this level of specificity were insufficient.

Science articles

I had hoped to include a science article genre such as might be represented by a historical corpus of one or more science journals such as *Science, Nature,* or *Scientific American* (although the latter is oriented to more general audiences). I found that usable text corpora were not readily available in sufficient quantity and could not be assembled within a reasonable project scope.

Legislation

See above under "Hansard" for the history of considerations and decisions for this genre.

Letters

Historical corpora of personal correspondence (written letters) are available. However, equivalent modern corpora are not, and we are also faced with the near-disappearance of the written letter per se as a genre. Email is available in great

quantity but is a different genre in many ways from the letters of ages past. A further forbidding issue is the unedited nature of personal correspondence, with much variation in spelling, especially in earlier times, which would make the results less reliable even with a multiplication of searched forms in an attempt to take in all possibilities. A future research effort focusing on the difference between electronic and written communication could be productive; such an effort would be informed by the results of the current effort more than it would contribute to them.

Poetry

BYU-BNC has a poetry genre, but COCA does not. Other sets of poetry are downloadable from Project Gutenberg, but the size and variety of the texts available did not promise to make the effort of downloading and preparing them worthwhile. As well, they do not match the time frames well; rather, they are divisible by genre: Romantic, Modern, Imagist, Metaphysical. This would make it a different sort of comparison, one not tracking the development of a genre but rather displaying the differences between more specific genres. I would suggest this as a project possibly worth undertaking in a separate research paper with different parameters, making use of my present results as a basis for comparison.

2.5 Summation

The necessary assembly of materials for the primary research exercise of this thesis was in itself of value. The evaluation of phonaesthemes produced clear evidence of their reality by demonstrating that there are several that meet the definition "a phonemic grouping that, within a language, correlates well above chance with a particular semantic quality in etymologically unrelated words." The words found to be associated with these phonaesthemes were also of the types identified in chapter 1: notably iconic and expressive. The selection of study lemmas confirmed that the number of lemmas per phonaestheme was a reasonable number; more would have been difficult to match reasonably. We will see in the next chapter that the set selected were sufficient for aggregate results. The exigencies involved in the choice of genres were instructive about the structural limits on work of this nature, but

also demonstrated that it is possible. We will see in the next chapter that the materials chosen were sufficient to produce interesting and viable results.

Chapter 3: Results and analysis

In this chapter, I present and analyze the results of the corpus surveys of frequency of the lemmata. The detailed results are in Appendix C; in this chapter I will look only at the aggregate results for the different sets of words for each genre, place, and time. I will give the full results in overview. I will then consider each genre's development over time individually. After that, I will look at the relations between genres within each time period. In each section, I will look at the relationship between phonaesthemes and genre in light of the data and external information, including some of the dimensions of variation as identified by Biber (1995), to identify the sort of role phonaesthemes play in genre definition: tone of use, level of use, nature of content, nature of communicative situation, and similar considerations. I will illustrate the points with examples from representative works. The data and observations will give a good understanding of the nature and function of phonaesthemes in communication.

I will use the short forms established in chapter 2 to refer to the different sets of study lemmas. The set containing phonaesthematic words will be referred to as **P**; the set containing words that have the phonemic clusters but not the semantic commonality will be **C**; the set having the semantic commonality but lacking the phonemic clusters will be **S**. Set **S** will be further subdivided into two mutually exclusive subsets that together compose the entire set: polysyllabic, Romance-derived words will be $\mathbf{S_R}$, and monosyllabic and Germanic words will be $\mathbf{S_G}$. I will use these initials to refer to the sets as such and will also use them in mathematical expressions to describe calculations. So, for instance, $\mathbf{P}/(\mathbf{P}+\mathbf{S_G})$ expresses the relative frequency of phonaesthematic words in that set that is the union of **P** (phonaesthematic words) and $\mathbf{S_G}$ (semantically related but non-phonaesthematic words that are monosyllabic and/or of Germanic origin).

3.1 Method

As stated in chapter 2, I searched all inflected forms of the study lemmas (see Appendix B) in the specified corpora. This gave me raw numbers. These numbers are not directly comparable between different corpora due to the different sizes of the corpora. A common way of making results such as these comparable between corpora is to express them as frequency relative to the total number of words in the corpus. I chose instead to analyze each set just by frequency relative to the total study lemmas. This was more appropriate for several reasons. First, different genres have different structural components that may add words that could not include the study lemmas. For example, Hansard has various pro forma procedural details as well as lists of names of members of parliament and similar items. Second, some genres are more prone to using more smaller words rather than fewer larger words, or to using circumlocutions or formal set phrasings that add words, without increasing the usage rates of the study lemmas correspondingly. Third, some genres will discuss certain topics or describe certain things more than others will. The study lemmas are focused on particular semantic areas that may be more common in one genre (or even one time period, in some cases) than in others. The **P** and **S** sets are in the same semantic areas and are generally fungible, so comparing relative frequency in just those sets should correct for over- or under-representation of that semantic area. Lastly, accurate total word counts were not equally available for all corpus subsets studied.

I used three different totals for analysis for a given genre and time period: **P**+**C**+**S**; **P**+**S** excluding **C**; and **P**+$\mathbf{S_G}$, which is **P**+**S** excluding polysyllabic Romance-derived words ($\mathbf{S_R}$). I will be focusing mainly on the latter two sets, which directly compare the semantically related phonaesthematic and non-phonaesthematic words with no consideration of the **C** set. The last **P**+$\mathbf{S_G}$ calculation will neutralize any length and word origin effect, and the difference between the two sets will allow us to have some sense of how much of the total effect may be accounted for by length and word origin. I will give a briefer look at the frequency of the **C** set relative to **P**+**C**+**S**; as I will discuss below, the **C** set did not produce results that were as usable or relevant

as the other sets, and its exclusion does not have an important on the results as they bear on the research question.

In my presentation of the data, with a few exceptions, I will show only the relative frequency of the phonaesthematic words. Since the non-phonaesthematic words are simply the remainder in the **P**+**S** and **P**+**S**$_{\mathbf{G}}$ sets, it would be redundant to display those results as well. I will also present the margin of error at 95% confidence, calculated as $0.98/\sqrt{n}$ where n is the total number of words of the study lemmas counted in that particular corpus, genre, and time period, so as to show which results are sufficiently different not to be the result of sampling error. This margin is one-half the confidence interval (CI); that is, the confidence interval is the margin above the data point plus the margin below it. This is the most reasonable approach to the numbers. This study is not a study of the frequency of all phonaesthematic words in the lexicon, or in use; it is a study of the difference in ratios of usage (phonaesthematic to non-phonaesthematic lemmas) between genres and time periods. As such, the null hypothesis is that all genres and time periods have the same ratio, whatever that ratio may be. The results of the corpus surveys are taken as samplings of an unlimited population of words for a given genre and time period; thus, the exercise is like a political poll, or like drawing coloured balls out of an enormous vat filled with them. The 95% confidence interval tells us the range within which the real frequency in the population will be, 19 times out of 20, given the sample we have drawn. On my analysis charts, I will show the edges of this interval flanking the data point so that we can see whether the differences between times, places, and corpora are significant or may be due to sampling error. In the overall results (section 3.2.1), I will not display the margin of error in the chart just for reasons of visual clarity (too many lines). I will include it in all charts per genre and per time period. The margins of error for the full results are in Table 6.

Naturally, the existence of a difference does not tell us the cause of the difference. I will thus take some time in my analysis of the results to consider factors other than the presence of phonaesthemes that could also account for the differences.

I am including the full data in Appendix C. The reader who is interested in the results for specific phonaesthemes or even specific lemmas can inspect the data and calculate totals and ratios as desired. I have elected not to present an analysis of the results at that level of detail, as the sample sizes for individual words and even individual phonaesthemes are mostly not large enough to produce statistically reliable results. However, in analyzing the overall results, I will consider the effect of individual items where they are salient.

3.2 Overall results

3.2.1 All lemmas

Table 3.1 shows the frequency of phonaesthematic lemmas (the **P** set) relative to all study lemmas (i.e., the union of all three sets, **P**+**C**+**S**), with the 95% confidence intervals (the margin of error plus or minus, calculated as $0.98/\sqrt{n}$ where n is **P**+**C**+**S** for that specific year and country).

Table 3.1: P/(P+C+S) with 95% confidence intervals

	1800 UK	1800 US	1900 UK	1900 US	2000 UK	2000 US
fiction	43.33±3.09%	30.92±1.66%	42.37±1.83%	41.77±0.68%	48.48±0.74%	45.68±0.56%
drama	29.84±5.52%					49.91±0.85%
magazine		21.58±4.16%		27.00±1.46%		37.26±0.82%
newspaper	8.66±6.45%			26.64±2.75%	52.44±1.17%	36.60±0.42%
humanities			23.42±6.58%		10.94±2.77%	13.28±0.89%
Hansard	9.76±4.84%		15.14±1.31%		24.12±0.60%	

The initial results tell a story of increasing use of phonaesthematic words in nearly all genres, the exception being the humanities articles, which have declined steeply. The UK fiction also has a slight decline from 1800 to 1900, and then a more notable incline from 1900 to 2000. The most striking increase is that of the UK newspapers between 1800 and 2000 (which we may remember is in fact between the 18th century and the past quarter-century). Overall, it stratifies, with all of the fiction at the highest numbers, US newspapers and magazines in the middle, a roughly parallel but much lower trend for the British Hansard, and the humanities articles

declining from medium numbers to the lowest results between 1900 and 2000. I will discuss these results in detail below by genre and by time period.

This set includes the results from the **C** set: those words containing the phonemic clusters but not having the semantic value associated with the phonaesthemes. Table 3.2 gives the results for **C**/(**P+C+S**), again with the 95% confidence intervals.

Table 3.2: C/(P+C+S) with 95% confidence intervals

	1800 UK	1800 US	1900 UK	1900 US	2000 UK	2000 US
fiction	13.55±3.09%	14.54±1.66%	8.15±1.83%	11.24±0.68%	10.47±0.74%	9.68±0.56%
drama	17.14±5.52%					15.37±0.85%
magazine		8.63±4.16%		11.38±1.46%		13.41±0.82%
newspaper	29.00±6.45%			17.49±2.75%	13.17±1.17%	13.08±0.42%
humanities			8.56±6.58%		6.31±2.77%	7.66±0.89%
Hansard	17.07±4.84%		13.42±1.31%		8.83±0.60%	

These results make for interesting reading but can be difficult to interpret, in part due to the small final set – more susceptible to the vagaries of individual words – and in part because the decline will be in some cases at least partly just the obverse of the increase in use of other types of words. Nonetheless, there are some interesting things worth noting.

First, there is no apparent phonaesthemic spreading effect – that is, these words are not seeing an increase in usage in the times and genres where phonaesthematic words are increasing in usage. In most sets, the effect is rather the opposite. The UK fiction, US magazine, and UK humanities sets do all roughly parallel the phonaesthematic sets in their trends, but the reason for this may be in some part related to the usage of polysyllabic Romance-derived words (see section 3.2.3). Given that the trends for these words are different from the trends for phonologically similar but phonaesthematic words, these data suggest that there may be some reality to the phonaesthematic effect we are seeking to discern. We will be seeing further data to confirm this.

It is tempting to speculate about causes of a decline in use of these words, but we should be careful not to put too much stock in these specific results; the final **C** set of words was diminished by necessity from the initial 29 to 12, and so do not constitute a very robust set. As well, at least some of the variation in these words can be accounted for by opposite variation in phonaesthematic words. For these reasons, and because – as we will see – the exclusion of these results does not have any important effect on the distribution and trends of the phonaesthematic results in the different genres across different times, they will be left out of the remainder of the discussion. We may note that they are all monosyllabic Germanic words, and so the trends applying to the monosyllabic Germanic words in the **C** set are likely to apply to them as well; however, I did not wish to merge this set with that set due to lack of semantic matching (and, of course, the presence of phonemic matching, although the results above indicate that that is not a concern of the type we may have anticipated).

3.2.2 Excluding C

Table 3.3: P/(P+S) with 95% confidence intervals

	1800 UK	1800 US	1900 UK	1900 US	2000 UK	2000 US
fiction	50.12±3.33%	36.18±1.80%	46.13±1.91%	47.06±0.72%	54.15±0.78%	50.58±0.59%
drama	36.02±6.07%					58.98±0.93%
magazine		23.62±4.35%		30.47±1.55%		43.03±0.88%
newspaper	12.20±7.65%			32.28±3.03%	60.40±1.25%	42.11±0.45%
humanities			25.62±6.88%		11.68±2.86%	14.38±0.92%
Hansard	11.76±5.31%		17.49±1.41%		26.46±0.63%	

Table 3.3 tells a story not appreciably different from Table 3.1. We see the same trends and the same stratification. I will present charts in the genre and time analysis sections below visually representing the data to make plain the trends and strate. To clarify how much of this may be due to the effects of word length and origin, as discussed in the literature review (see Systematicity in section 1.1.4 above), I have calculated $\mathbf{P}/(\mathbf{P}+\mathbf{S_G})$, which excludes the $\mathbf{S_R}$ (polysyllabic Romance-derived) set.

3.2.3 Excluding polysyllabic Romance-derived words

Table 3.4: P/(P+S_G) with 95% confidence intervals

	1800 UK	1800 US	1900 UK	1900 US	2000 UK	2000 US
fiction	60.84±3.66%	46.42±2.04%	52.57±2.04%	52.67±0.76%	59.41±0.82%	57.39±0.63%
drama	46.08±6.86%					65.14±0.97%
magazine		38.59±5.56%		42.94±1.84%		58.98±1.03%
newspaper	14.60±8.37%			48.49±3.71%	77.12±1.42%	64.29±0.56%
humanities			44.83±9.10%		22.99±4.01%	38.67±1.52%
Hansard	31.01±8.63%		30.93±1.88%		63.91±0.97%	

The exclusion of the $\mathbf{S_R}$ set removes much of the stratification. However, it does not change the overall trends; they are simply more tightly clustered together. (We should not forget, either, that some of the effect from the polysyllabic Romance words is likely to be the same difference in tone and expressivity that distinguishes the other non-phonaesthematic words from the phonaesthematic words; the stratification may owe much to the $\mathbf{S_R}$ set, but it is fair to expect that the phonaesthematic effect accounts for about as much of the difference between the **P** set and them as it does between the **P** set and the $\mathbf{S_G}$ set. The length and origin effects account for most or all of what remains after that has been taken into account.) We see that the distance between the news/magazine genres and the fiction genres is largely attributable to the Romance-derived words, and even the Hansard ends up clustered with nearly all of the others in 2000. We will see below that for newspapers and Hansard, one lemma in particular accounts for a large part of the difference, but the upshot of the analysis is not much changed when we exclude it. The exception to the closer clustering is the humanities articles, which have diverged – a trend consistent with the differentiation of academic prose in the "Involved versus Informational" and "Non-Abstract versus Abstract" dimensions in Biber (1995), as I discussed in section 1.2.2. There remains an overall upward trend in nearly all other genres, except UK fiction, which dips and returns. I will discuss the implications of these trends in the individual results by genre (section 3.3, below).

To clarify the weight of the $\mathbf{S_R}$ set, I have calculated in Table 3.5 its relative frequency in the different genres and times, with 95% CI.

Table 3.5: S_R/(P+S) with 95% confidence intervals

	1800 UK	1800 US	1900 UK	1900 US	2000 UK	2000 US
fiction	15.24±3.33%	18.85±1.80%	11.25±1.91%	9.45±0.72%	7.93±0.78%	10.72±0.59%
drama	18.10±6.07%					8.01±0.93%
magazine		35.43±4.35%		25.74±1.55%		23.41±0.88%
newspaper	11.69±7.65%			27.58±3.03%	18.82±1.25%	29.99±0.45%
humanities			39.19±6.88%		46.09±2.86%	58.00±0.92%
Hansard	51.46±5.31%		37.63±1.41%		53.42±0.63%	

As we have just observed above, the main effect of the polysyllabic Romance-derived words is stratification: certain genres are more prone to including them than others, and this stratification is persistent across the centuries.

The trends of polysyllabic Romance-derived words in the different genres are of potential interest in at least some cases. Their use in humanities academic articles is comparatively high, and has increased over the last century, which corresponds well with the decrease in use of phonaesthematic words. However, as we have seen above, this does not fully account for the decrease in phonaesthematic words; they decreased by comparison with their non-phonaesthematic Germanic counterparts as well.

The relative frequency of set $\mathbf{S_R}$ in the British Hansard has an interesting inflection; it dips from 1800 to 1900 and increases for 2000, a trend we also see for phonaesthematic words in that genre. When we look at the individual lemmas, however (see Appendix C), we have a better sense of the reason: use of most polysyllabic Romance lemmas decreased between 1990 and 2000, but use of *expand/expansion* increased sharply. This appears to be subject-specific. Without this item, the use of polysyllabic Romance words drops steeply. The use of phonaesthematic words retains its general contour but increases more steeply between 1900 and 2000. I will look at the results with that item excluded below.

The relative frequency of set $\mathbf{S_R}$ in US newspapers and magazines has generally declined; if we remove *expand/expansion*, the decline is steeper. It seems reasonable to infer that this item is overrepresented in the news media for many of the same reasons it is overrepresented in the Hansard: economic and political activity. Likewise, that one item accounts for the increase in the UK newspaper results for $\mathbf{S_R}$ – without it, they show a slight decrease. We should note, however, that this item does not cause the phonaesthematic results to decline over the same period; their upward trend remains, even if it is slightly less steep.

The relative frequency of set $\mathbf{S_R}$ starts comparatively low and stays low in the fiction and drama genres, which is unsurprising. There is a slight increase in the US fiction between 1900 and 2000; it is not accounted for by any single item – it is manifest across several lemmas. It would be interesting to explore how this breaks out between genres; scientific and medical details are certainly present in many modern genres of fiction. But they are also present in British fiction, and the UK fiction showed a decrease in this set from 1900 to 2000. We should be aware, however, that there may be some difference in composition of the corpora that may at least partially account for the difference. As we have seen in section 1, genre is open to variation and vagary in definition, and this is a possible hazard of research enterprises such as this thesis too. I will look at this question in more detail in section 3.3.1.

For completeness, I will include a chart and table of the trends for the $\mathbf{S_G}$ set (i.e., all monosyllabic words and those few words that are polysyllabic but Germanic).

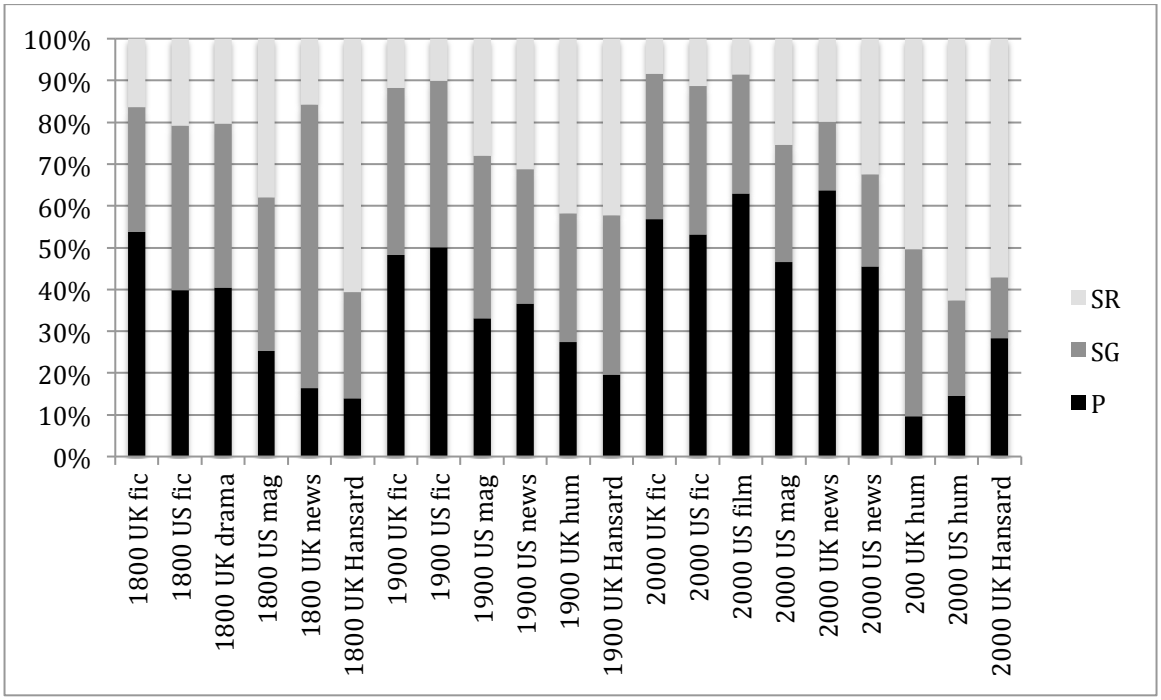
Table 3.6: S_G/(P+S) with 95% confidence intervals

	1800 UK	1800 US	1900 UK	1900 US	2000 UK	2000 US
fiction	27.89±3.33%	35.69±1.80%	38.23±1.91%	37.53±0.72%	33.12±0.78%	33.92±0.59%
drama	34.92±6.07%					26.71±0.93%
magazine		34.35±4.35%		35.88±1.55%		25.92±0.88%
newspaper	50.65±7.65%			28.29±3.03%	15.56±1.25%	20.33±0.45%
hum-art			28.83±6.88%		36.66±2.86%	21.06±0.92%
hansard	21.71±5.31%		33.81±1.41%		13.62±0.63%	

The most interesting thing about this set of data is the tight clustering. There are differences from genre to genre, but they are not as stark as seen for the Romance-derived words or for the phonaesthematic words. We have already seen that some of this variation can be accounted for by the *expand/expansion* effect in the politically focused genres (Hansard and newspapers). This data set confirms that an important part of the difference from genre to genre can be accounted for by the polysyllabic Romance-derived words, but it also shows that such variation as remains when those words are excluded is due to the phonaesthematic words, which differ from these other short Germanic words in having phonaesthemes and consequently in all the effects associated with that.

Figure 3.1 presents a synopsis of all proportions of sets **P**, $\mathbf{S_G}$, and $\mathbf{S_R}$ in each genre for each time period. In this view, a few things are more salient. We can see that while phonaesthemes were uncommon in the 1800 UK news, polysyllabic Romance words were also uncommon; the plain short Germanic words were the mainstay. We can see, too, that the 2000 Hansard, for all its Romance erudition (including a strong skew from the *expand/expansion* lemma – see section 3.3.6 below for data with that excluded), also has more phonaesthematic words than the humanities articles that are in the same range of Romance words. This is less surprising, perhaps, when we consider the performative nature of parliamentary speeches: they aim to be impressive and impactful, which means the polysyllabic Romance words to accomplish the former and the phonaesthematic words to add the latter. But the 1900 and 1800 Hansards lack the phonaesthematic aspect. A possible explanation is that more recent politicians may need to appeal to the common person more strongly than those of earlier eras. We will see examples below that support this hypothesis.

Figure 3.1: Proportions of all word types in all (P+S) results



3.3 Diachronic per genre

I will first look at what the results suggest about the historical development of the individual genres. For these analyses, I will present two charts for each genre, one showing **P**/(**P+S**) and the other showing **P**/(**P+S$_G$**). The absolute numbers are not comparable between the two calculations, as the **S$_G$** set is smaller than the **S** set, but I have scaled the Y axes so as to show the trends within similar relative scaling. Each chart will show the results as boxes showing the full 95% confidence interval for each result, with a line in the middle indicating the actual data point. I will not repeat the numerical results in tables, as they are given in the tables above (and in greater detail in Appendix C).

3.3.1 Fiction

Figure 3.2: Fiction across time, P/(P+S) with 95% CI

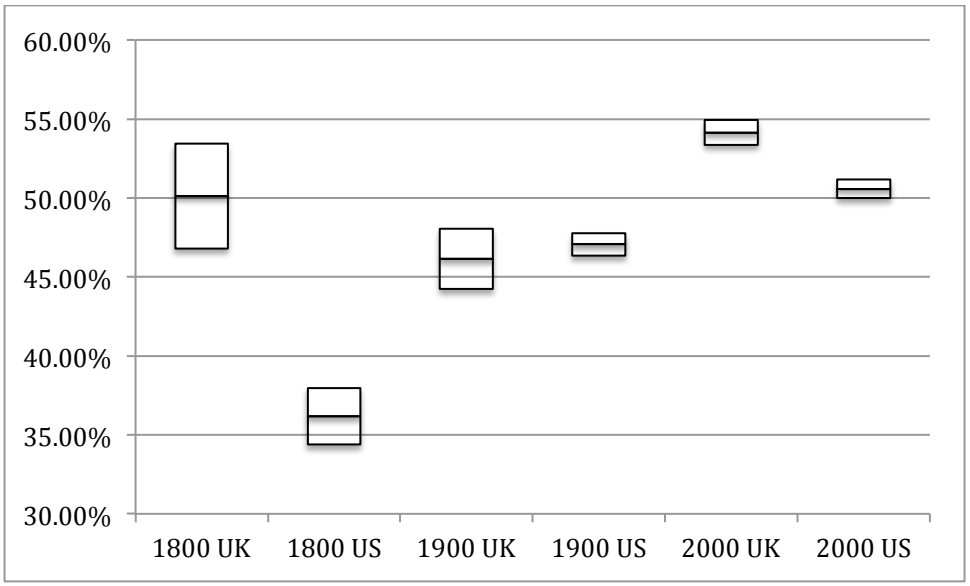
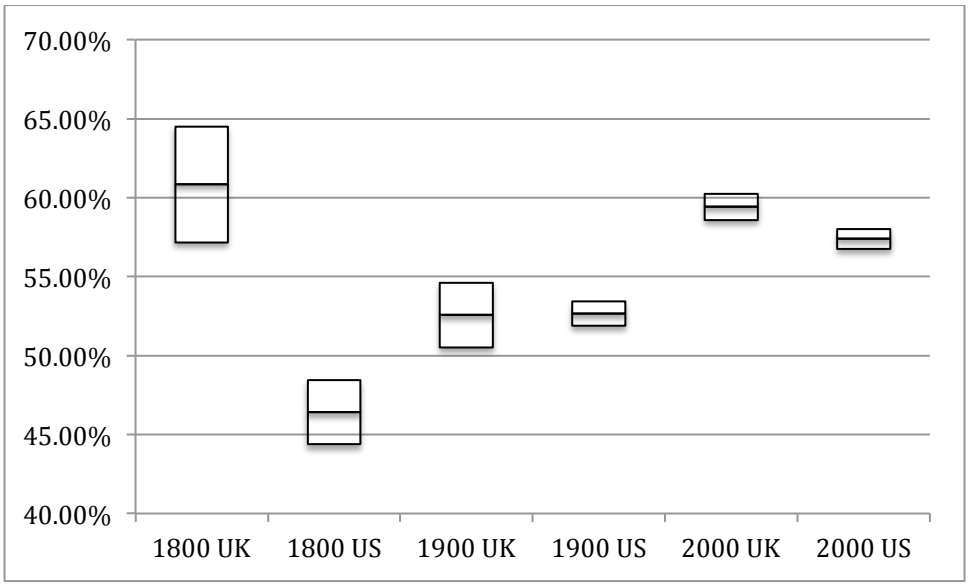


Figure 3.3: Fiction across time, P/(P+S_G) with 95% CI



We see from Figures 3.2 and 3.3 that the $\mathbf{S_R}$ set has an appreciable effect between 1800 and 1900 in the UK, but otherwise does not change the overall contour of the results: a dip and return for the UK results, and an upward trend across time for the US results. At 1900, both continents had effectively the same use of language in the

details that concern us, but at 1800 they were starkly different, the UK fiction having much more phonaesthematic usage, and in 2000 they are slightly different, the UK again leading in phonaesthematic usage, but not by as much.

We can notice interesting similarities of pattern when we compare these trends to Biber's (1995) historical charts of genres, which I introduced in section 1.2.3 and will reproduce here for ease of reference. Dimensions are "sets of co-occurring linguistic features, reflecting different functional underpinnings (e.g., interactiveness, planning, informational focus and explicitness)" (Biber 1995, 36); Biber has used multidimensional statistical analyses of linguistic features to identify co-occurring bundles of features, and has given these bundles (dimensions) names that characterize the type of difference they make. I have found two of Biber's dimensions to be relevant to my research: "Involved versus Informational Production" and "Non-abstract versus Abstract Style." Biber's chart of "Involved versus Informational Production" (289) shows a V shape for fiction over time:

Figure 3.4: "Involved versus Informational Production" over time for several genres (Biber 1995, 289)

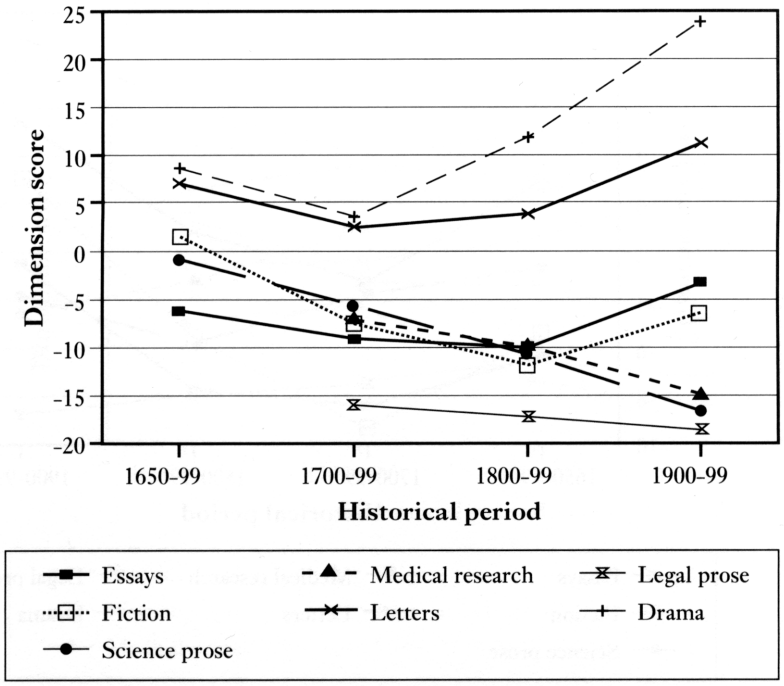


Figure 8.4 Historical change among seven written and speech-based registers along English dimension 1: 'Involved versus Informational Production'

The inflection point for this line is the 19th century. Biber's chart of "Non-abstract versus Abstract Style" (291) shows a line inflecting in the 18th century:

Figure 3.5: "Non-abstract versus Abstract Style" over time for several genres (Biber 1995, 291)

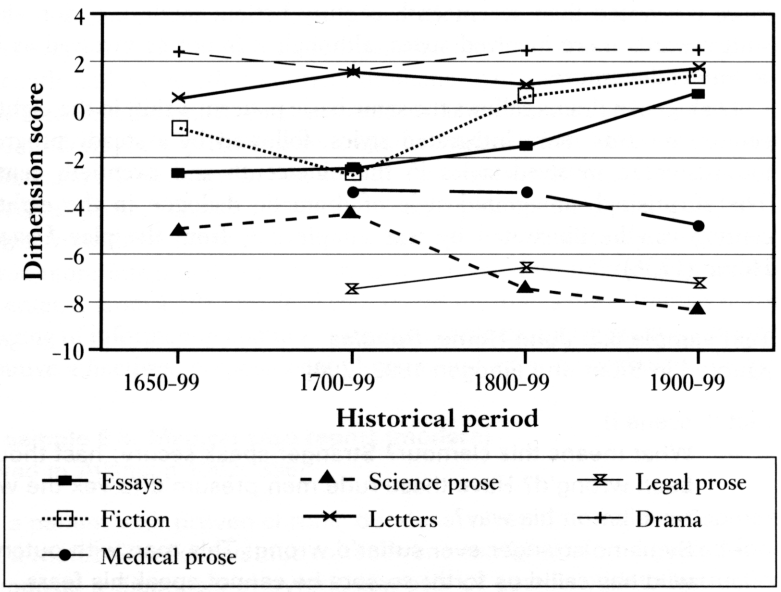


Figure 8.6 Historical change among seven written and speech-based registers along English dimension 5: 'Non-abstract versus Abstract Style'

We can see that the trend of phonaesthematic words in American fiction matches more closely the development from abstract to non-abstract, and in British fiction the trend matches more closely the variation between informational and involved. The questions we are faced with are whether the use of phonaesthemes is more characteristic of non-abstract style or of involved style – whether our British results or our American results are better matched to Biber's corpora, in short – or whether American dimensions of variation are different from British ones, with phonaesthemes working more with non-abstract style in the US but with involved style in the UK.

The first issue in addressing this is the composition of Biber's corpus – whether the authors are British, American, or both. Biber tells us what authors compose his corpora (1995, 88–89): in the 18th century, his fiction sources are Austen, Defoe, Fielding, Johnson, and Swift, British all; in the 19th century (to 1865), they are Dickens, Hawthorne, Kingsley, Melville, Mill, and Poe – largely but not exclusively British; in the modern era (since 1865), they are Harte, Hemingway, Lawrence,

Lewis, Orwell, Steinbeck, Twain, and Woolf, more evenly balanced and perhaps leaning slightly towards the Americans. Since the key difference is in the 1800 sets, we might expect a closer match with the British, suggesting that the involved-information dimension may be more pertinent, but the different time divisions reduce certainty on this.

Another question is of whether the sub-genres in the fiction sets are well matched. The subset of CLMET I used for the British 1800 results includes Austen, Lamb, Wollstonecraft, Scott, and Disraeli, among others – largely romance and adventure, and generally the same as Biber's set; for British 1900, it is an assortment of novels mostly by authors not well known today (Samuel Butler, Rudyard Kipling, and E.M. Forster are the best known among them), but all popular fiction, aimed at largely the same audiences as the 1800 set. The composition of the COHA, COCA, and BYU-BNC corpora is not given, although the sub-genres listed for the two modern corpora show a catholic assortment.

When we look at Biber's factorial structure for the two dimensions, reproduced here for convenience (1995, 142 and 163), we can see certain aspects that may tend to go with our phonaesthematic set of words.

Table 3.7: Co-occurring linguistic features in "Involved versus Informational Production" (Biber 1995, 142)

Table 6.1 *Co-occurring linguistic features on English dimension 1: 'Involved versus Informational Production.' (Features in parentheses have lower weights and are not used in the computation of dimension scores)*

Dimension 1	
'Involved Production'	
Positive features:	
Private verbs	0.96
THAT deletion	0.91
Contractions	0.90
Present tense verbs	0.86
Second person pronouns	0.86
DO as pro-verb	0.82
Analytic negation	0.78
Demonstrative pronouns	0.76
General emphatics	0.74
First-person pronouns	0.74
Pronoun IT	0.71
BE as main verb	0.71
Causative subordination	0.66
Discourse particles	0.66
Indefinite pronouns	0.62
General hedges	0.58
Amplifiers	0.56
Sentence relatives	0.55
WH questions	0.52
Possibility modals	0.50
Non-phrasal co-ordination	0.48
WH clauses	0.47
Final prepositions	0.43
(Adverbs)	0.42
'Informational Production'	
Negative features:	
Nouns	−0.80
Word length	−0.58
Prepositions	−0.54
Type–token ratio	−0.54
Attributive adjectives	−0.47
(Place adverbials	−0.42)
(Agentless passives	−0.39)
(Past participial postnominal clauses	−0.38)

In Table 3.7, which shows the factor loadings for the "Involved versus Informational Production" dimension, we see that amplifiers, emphatics, and demonstratives all factor in with "involved" production, and they are all more pointedly vivid and performative, as we have found phonaesthematic words are. We can see that short words go more with "involved" production (word length loads at –0.58), and the phonaesthematic words are short. But the words in sets **C** and $\mathbf{S_G}$ are also short, and

they do not pattern with the phonaesthematic words. Thus it is reasonable to expect that phonaesthemes would factor in strongly with the Involved dimension.

Table 3.8: Co-occurring linguistic features in "Non-abstract versus Abstract Style" (Biber 1995, 163)

Table 6.5 *Co-occurring linguistic features on English dimension 5: 'Non-abstract versus Abstract Style.' (Polarity reversed – see note 3)*

Dimension 5	
[No positive features]	
Negative features:	
Conjuncts	−0.48
Agentless passives	−0.43
Past participial adverbial clauses	−0.42
BY-passives	−0.41
Past participial postnominal clauses	−0.40
Other adverbial subordinators	−0.39

The factors for "Non-abstract versus Abstract Style," shown in Table 3.8, are not specific to phonaesthematic words, but in the direction of "abstract" (and thus away from our phonaesthematic results) they do remove the agent and generally take an approach that is less geared towards direct depiction of action. Phonaesthematic words are obviously geared more towards direct depiction of action, and so we would expect them to factor in with the non-abstract. Our results for fiction suggest that this is so, but possibly not as strongly as with the other factor, especially in the UK fiction, which would suggest that phonaesthemes function more strongly as indicators of the involvement of the speaker or writer – that is, as markers of personal expression and involvement.

This in turn could be read as suggesting that British fiction became more genteel and restrained from 1800 to 1900 and then rebounded, while American fiction was even more genteel and restrained in 1800 and advanced to a current state of expressivity that is still not quite at the level of the British. There is naturally a reasonable question of whether my study lemmas happen to be more common in British usage overall; the problem with doing a general survey of their frequency would be that even if we found this to be so, we would need a much fuller study of

other factors to assess the possibility that British speech is in general more involved and expressive (as my results for the most part suggest). Thus I will limit my observations here to what is manifest in the fiction results.

One important point to bear in mind is that these literatures did not come into existence in 1800; we are starting *in medias res*. Biber's charts go back farther in time, and they illustrate a point that Biber discerns (1995, 297): that "most registers evolved to become even more distinct from speech over the first 100–200 years of their history." Many of these genres had their inception in English in or around the 1600s. But whereas scholarly registers have diverged ever farther from the common speech, genres such as fiction and popular literature (e.g., magazines) shifted back towards oral styles with the rise of mass middle-class – and then lower-class – literacy (Biber 1995, 298). We need also to remember, however, that American fiction did not come into independent existence as a distinct national genre until around the time of the American Revolution.

Let us consider typical examples of passages from works of the different times and places. Consider this passage from *Pride and Prejudice* by Jane Austen (taken from the version in the CLMET corpus), first published in 1813:

> "Oh! my dear Mr. Bennet," as she entered the room, "we have had a most delightful evening, a most excellent ball. I wish you had been there. Jane was so admired, nothing could be like it. Everybody said how well she looked; and Mr. Bingley thought her quite beautiful, and danced with her twice! Only think of *that*, my dear; he actually danced with her twice! and she was the only creature in the room that he asked a second time. First of all, he asked Miss Lucas. I was so vexed to see him stand up with her! But, however, he did not admire her at all; indeed, nobody can, you know; and he seemed quite struck with Jane as she was going down the dance. So he inquired who she was, and got introduced, and asked her for the two next. Then the two third he danced with Miss King, and the two fourth with Maria Lucas, and the two fifth with Jane again, and the two sixth with Lizzy, and the *Boulanger*—"

This is just the sort of locution we have learned to expect from this period, with long complex sentences and many polite forms. Notwithstanding this, there are words such as *vexed* and *struck* peppered into it as well, and it has many short Germanic words to mix in with its *delightlful, beautiful, creature, admire,* and similar words.

Compare it with this passage from *The Coquette: Or, the History of Eliza Wharton* by Hannah Webster Foster, a very popular early American novel first published in 1797, serialized for general audiences before it was issued in book form:

> We arrived at Col. Farington's about one o'clock. The Col. handed me out of the carriage, and introduced me to a large company assembled in the Hall. My name was pronounced with an emphasis; and I was received with the most flattering tokens of respect. When we were summoned to dinner, a young gentleman in a clerical dress offered his hand, and led me to a table furnished with an elegant, and sumptuous repast, with more gallantry, and address than commonly fall to the share of students. He sat opposite me at the table; and whenever I raised my eye, it caught his. The ease, and politeness of his manners, with his particular attention to me, raised my curiosity, and induced me to ask Mrs. Laiton who he was? She told me that his name was Boyer; that he was descended from a worthy family; had passed with honor and applause through the university where he was educated; had since studied divinity with success; and now had a call to settle as a minister in one of the first parishes in a neighbouring state.

The scene is not so different in nature or politesse from the one in Austen, and the novel, being epistolary, is written in the voice of a woman writing a letter to her friend. But the vocabulary includes more longer and classically derived words and fewer short, direct words. It is worth noting that the speech of the British upper classes has been observed to be often more direct, and that of the middle classes has been seen to put on more airs (Mitford 1956); whether this is a factor between the patrician British set and the striving colonial set in these novels is something worth exploring.

British fiction circa 1900 included authors such as Rudyard Kipling, Samuel Butler, and E.M. Forster. Here is a passage from *The Way of All Flesh* by Samuel Butler, first published in 1903 (taken from the CLMET corpus):

> He was softened by Christina's winning manners: he admired the high moral tone of everything she said; her sweetness towards her sisters and her father and mother, her readiness to undertake any small burden which no one else seemed willing to undertake, her sprightly manners, all were fascinating to one who, though unused to woman's society, was still a human being. He was flattered by her unobtrusive but obviously sincere admiration for himself; she seemed to see him in a more favourable light, and to understand him better than anyone outside of this charming family had ever done. Instead of snubbing him as his father, brother and sisters did, she drew him out, listened attentively to all he chose to say, and evidently wanted him to say still more. He told a college friend that he knew he was in love now; he really was, for he liked Miss Allaby's society much better than that of his sisters.

The language is genteel, and has complex sentences and many longer words, but notice also the possibly phonaesthematic words *sprightly, flattered,* and *snubbing*.

American fiction circa 1900 included authors such as H.G. Wells, Henry James, and Mark Twain – in other words, quite a diversity. Here is a vivid moment from *The Age of Innocence* by Edith Wharton, first published in serialized form in 1920:

> As Madame Nilsson's "M'ama!" thrilled out above the silent house (the boxes always stopped talking during the Daisy Song) a warm pink mounted to the girl's cheek, mantled her brow to the roots of her fair braids, and suffused the young slope of her breast to the line where it met a modest tulle tucker fastened with a single gardenia. She dropped her eyes to the immense bouquet of lilies-of-the-valley on her knee, and Newland Archer saw her white-gloved finger-tips touch the flowers softly. He drew a breath of satisfied vanity and his eyes returned to the stage.

It is evocative, but uses words such as *mounted, mantled,* and *suffused*. Words such as *breath* and *glove* appear, but to less impact. Wharton describes a blush without using the word *blush*.

For contrast, however, consider the style of the pointedly rural dialect of Twain's *The Adventures of Huckleberry Finn*, first published in 1884:

> Pretty soon a spider went crawling up my shoulder, and I flipped it off and it lit in the candle; and before I could budge it was all shriveled up. I didn't need anybody to tell me that that was an awful bad sign and would fetch me some bad luck, so I was scared and most shook the clothes off of me. I got up and turned around in my tracks three times and crossed my breast every time; and then I tied up a little lock of my hair with a thread to keep witches away. But I hadn't no confidence. You do that when you've lost a horseshoe that you've found, instead of nailing it up over the door, but I hadn't ever heard anybody say it was any way to keep off bad luck when you'd killed a spider.

When we consider American literature of the time as a whole, it subsumes dramatically different sub-genres such as the above two. If we consider the oeuvres of individual authors, we are sure to get starkly different results; the present study can serve as a basis for comparison for future efforts examining sub-genres and specific authors.

It goes without saying that literature of the present day is, if anything, even more varied. Entire genres of fiction have arisen in the time, and literacy has become near-universal. But even in works in the same general area as those cited above – chronicles of romances and relationships (the Twain is obviously an exception) – we can see important differences. Here is a passage from Mark Helprin's 2012 *In Sunlight and in Shadow* (page 10):

> She was a flow of color. Her hair trapped the sun and seemed to radiate light. It moved in the wind at the nape of her neck and where it had come loose, but was otherwise gloriously up in a way that suggested self-possession and

> formality and yet also exposed most informally the beauty of her shoulders. She wore a blouse with a low collar that even across the gap he could see was embroidered in pearl on white, and the glow of the blouse came not only from its nearly transparent linen but from the woman herself. The narrowing at her waist, a long drop from her shoulders, was perfect and trim.

The sentences are long and the language may seem genteel to us, but consider: *flow, trapped, gloriously, glow, drop, trim* – the vocabulary does not insulate itself in the same way as many works of earlier times did.

Let us take as a last example a popular contemporary British author, Philippa Gregory, a historian who writes historical novels. You might expect them to emulate the speech represented in fiction of earlier times, but you would be mistaken if you did. Rather, they give a modern impression of it at most. Here is a passage from her 2004 novel *The Queen's Fool* (page 41):

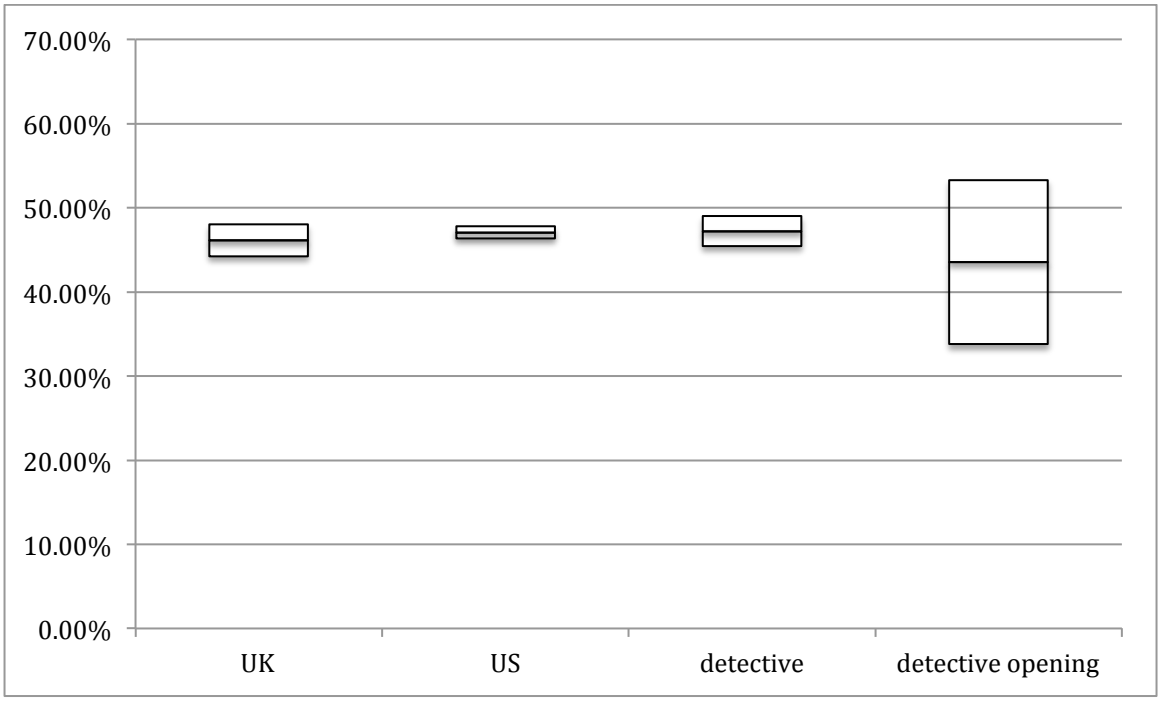
> She was a woman in her thirty-seventh year, but she still had the pretty coloring of a girl: pale skin and cheeks which flushed rosy pink. She wore the hood set back off her square honest face and showed her hair, dark brown with a tinge of Tudor red. Her smile was her great charm: it came slowly, and her eyes were warm. But what struck me most about her was her air of honesty. She did not look at all like my idea of a princess—having spent a few weeks at court I thought everyone there smiled with hard eyes and said one thing and meant the opposite. But this princess looked as if she said nothing that she did not mean, as if she longed to believe that others were honest too, that she wanted to ride a straight road.

There are comparatively few long or ornate words. The sentences are not exceedingly short, but they are much less complex. There is no shying away from *flushed* or *straight* or non-phonaesthematic monosyllables such as *dark, hard,* and *square*. This book, in its time, might be seen as in the same genre as *Pride and Prejudice*, and yet the language usage has evolved quite a bit – and towards plainer

speech. This is an important consideration: plainness of speech, and the extent to which high-iconicity words are considered "plainer" speech (and so more basic and also less erudite or élite) than words that are more purely arbitrary.

In hopes of identifying a structural effect in usage of phonaesthemes – greater use in one part of a structured work than another – I compiled a corpus of detective fiction from the late 1800s and early 1900s and a separate corpus of just the first 1000 words from each work. This was necessary because no similar corpus isolating the opening passages of works could be made from COCA, BYU-BNC, or COHA, and the works in the CLMET sets would not have given a sub-corpus of sufficient size. The results, however, show no statistically significant difference between them (results are presented as bars covering the 95% confidence interval, with a split in colour at the data point). Figure 3.6 shows **P**/(**P**+**S**) for the UK 1900 and US 1900 fiction sets, the full detective fiction set, and the set containing just the first 1000 words of each work in the detective fiction set. I have not created a separate figure showing data for the **P**/(**P**+$\mathbf{S_G}$) set, as the results are not significantly different.

Figure 3.6: Detective fiction and first 1000 words of detective fiction compared to all fiction, 1900: P/(P+S)



While we may have different expectations of detective fiction as a genre from some other genres of fiction, the sample in this case came from the years when it was first emerging as a distinct genre. As we observed in section 1.2, genres do not emerge *ex nihilo* or spring fully formed from the brow of the writer; they are based on existing genres. At 1900, fiction had come to a reasonably consistent state between the US and UK, and detective fiction appears not to have differentiated itself in language usage, at least in the aspect we are examining. Consider this example, the opening two paragraphs of *The Sign of the Four* by Arthur Conan Doyle:

> Sherlock Holmes took his bottle from the corner of the mantel-piece and his hypodermic syringe from its neat morocco case. With his long, white, nervous fingers he adjusted the delicate needle, and rolled back his left shirt-cuff. For some little time his eyes rested thoughtfully upon the sinewy forearm and wrist all dotted and scarred with innumerable puncture-marks. Finally he thrust the sharp point home, pressed down the tiny piston, and sank back into the velvet-lined arm-chair with a long sigh of satisfaction.
>
> Three times a day for many months I had witnessed this performance, but custom had not reconciled my mind to it. On the contrary, from day to day I had become more irritable at the sight, and my conscience swelled nightly within me at the thought that I had lacked the courage to protest. Again and again I had registered a vow that I should deliver my soul upon the subject, but there was that in the cool, nonchalant air of my companion which made him the last man with whom one would care to take anything approaching to a liberty. His great powers, his masterly manner, and the experience which I had had of his many extraordinary qualities, all made me diffident and backward in crossing him.

Here we have a scene that graphically describes a man injecting cocaine into his arm, the opening image of a detective novel, and yet it has a comparatively small proportion of phonaesthematic or similarly short and punchy words – *thrust* is the

most salient – and quite a lot of lengthy locutions such as "Again and again I had registered a vow that I should deliver my soul upon the subject."

The same experiment repeated with a larger corpus of works of a more distinctly action-oriented genre might produce different results, and this is a study I would like to conduct in future. The current results serve only to reinforce the general impression we have gotten of fiction circa 1900, which is comparatively low in use of phonaesthemes, even as it describes things that could be put in quite graphic terms. We may be temped to characterize this as a "Victorian" attitude, manifesting the rise of the middle class and its emphasis on decorum in the effort to climb socially. This gives us a view of phonaesthematic words as comparatively indecorous, a trait which cannot escape involvement in genre definition.

3.3.2 Drama/film

Figure 3.7: Drama and film across time, P/(P+S) with 95% CI

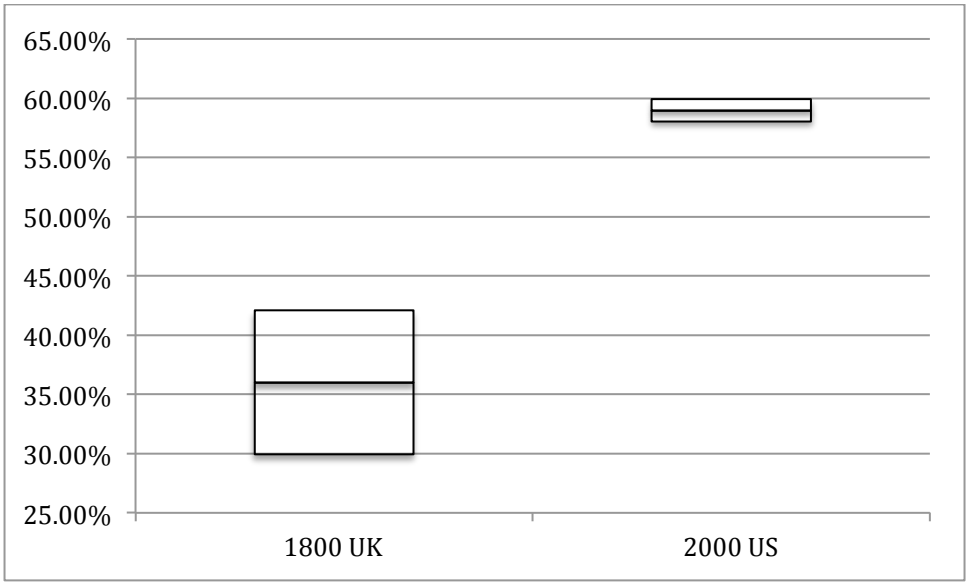
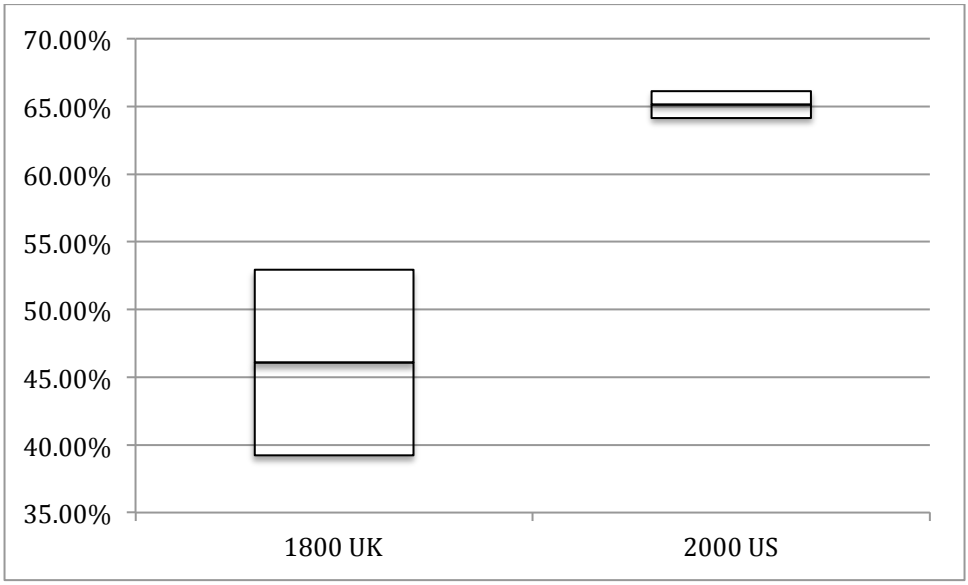


Figure 3.8: Drama and film across time, P/(P+S_G) with 95% CI



Figures 3.7. and 3.8 show the difference between drama in the UK in 1800 and movie scripts in the US in 2000. These are two performative genres, but we can see that performativity is not the only or even the most important factor in phonaestheme usage. These genres make representations of speech behaviour, after all, and they do so within the bounds of a literary genre; they are not transcriptions of actual speech. The difference between the two figures shows us that in the development from stage drama in 1800 to film in 2000, the influence of polysyllabic Romance-derived words may have made some difference in our results, but not an important one. What we *can* discern undeniably from both figures is that there is a significant increase in rates of usage of phonaesthematic words from the one set to the other. This ought not to be surprising, given the starkly different approaches to theatrical representation between the two. Consider this speech from Richard Brinsley Sheridan's *The School for Scandal* of 1777 (taken from the CLMET corpus):

> Sir Peter, I cannot expect you will credit me; but the tenderness you expressed for me, when I am certain you did not know I was within hearing, has penetrated so deep into my soul, that could I have escaped the mortification of this discovery, my future life should have convinced you of my sincere repentance. As for that smooth tongued hypocrite, who would

> have seduced the wife of his too credulous friend, while he pretended an honourable passion for his ward, I now view him in so despicable a light, that I shall never again respect myself for having listened to his addresses.

Compare to it this fatherly advice from the 2007 movie *Juno* (Cody 2007), one of the longest single-character speeches in the script:

> It's not easy, that's for sure. Now, I may not have the best track record in the world, but I have been with your stepmother for ten years now, and I'm proud to say that we're very happy. . . . In my opinion, the best thing you can do is to find a person who loves you for exactly what you are. Good mood, bad mood, ugly, pretty, handsome, what have you, the right person will still think that the sun shines out your ass. That's the kind of person that's worth sticking with.

The dialogue is "realistic" but not an accurate representation of ordinary speech; it's too clever. There is a genre effect, and a writerly desire to make it appealing – like life, but more interesting.

Is it reasonable to present US films as having developed from UK drama? In truth, the dramatic tradition of England in 1800 was the common dramatic tradition of the two countries; few plays had yet been written in America, and British plays held the stage and influenced American drama for some time (Banham 1988, 1014). The development of drama progressed towards less formal language over the years in both England and the United States, as well as elsewhere; naturalism was the order of the day by the early 1900s, although speech even in naturalist dramas of the time was not strongly reflective of truly natural speech patterns. George Bernard Shaw was considered a leading modernist and realist playwright, but even when describing action his characters made speeches such as the following from *Arms and the Man* (Shaw 2015):

> He did it like an operatic tenor—a regular handsome fellow, with flashing eyes and lovely moustache, shouting a war-cry and charging like Don Quixote

> at the windmills. We nearly burst with laughter at him; but when the sergeant ran up as white as a sheet, and told us they'd sent us the wrong cartridges, and that we couldn't fire a shot for the next ten minutes, we laughed at the other side of our mouths. I never felt so sick in my life, though I've been in one or two very tight places. And I hadn't even a revolver cartridge—nothing but chocolate. We'd no bayonets—nothing. Of course, they just cut us to bits. And there was Don Quixote flourishing like a drum major, thinking he'd done the cleverest thing ever known, whereas he ought to be courtmartialled for it. Of all the fools ever let loose on a field of battle, that man must be the very maddest. He and his regiment simply committed suicide—only the pistol missed fire, that's all.

There is no lack of expressive language, including phonaesthematic words such as *flashing* and *burst*, but it's a far cry from standard dialogue in current movies or even in many current plays.

From the time of the first "talkie" movies, the speech was very similar in style to that of the stage plays of the day, likely in part because stage playwrights were also writing some of the film scripts and in part because stage plays were the available model for the new genre. Not until mid-late-century playwrights such as the British playwright Harold Pinter and the American David Mamet (who at times recorded overheard conversations in public places such as bars and used them as guidance) did speech more frankly emulate the rhythms, locutions, and discontinuities of everyday speech, but even in the present time plays do not usually exactly mirror ordinary speech. Movie scripts evolved in parallel. But the speech style has become more colloquial, less formal, less erudite-seeming, and the increase in frequency of phonaesthematic words illustrates that. Again we see these words as part of a more natural and immediate style, and less pointedly formal or decorous. The speech is not striving to be disinterested or to match what is considered the most educated model of speech, but is rather intended to be more directly engaging. The interesting point with regard to phonaesthemes is that although plays circa 1800 were no less

performances than movies circa 2000 – indeed, being live performances, they may arguably have been more so – the usage of phonaesthemes has not been consistent. It is a question not of the actually performative nature of the occasion but rather of what is considered appropriate from a literary genre perspective. Frank expressivity and verbal iconicity is more accepted in the modern time in scripts just as in fiction.

Scripts are not just dialogue, however; there is another component: the directions, telling the actors, directors, and designers what to do and how. Stage directions are fairly limited in drama circa 1800; they are mostly simple notes such as “Enter,” “Exit,” and “Aside,” and actions are generally to be inferred from the dialogue, as with this from *The School for Scandal*:

> Walk in, Gentlemen, walk in; Trip give chairs; sit down Mr. Premium, sit down Moses. Glasses Trip; come, Moses, I'll give you a sentiment. "Here's success to usury." Moses, fill the gentleman a bumper.

Scene descriptions similarly just name the location. Thus we do not see an important effect on usage of phonaesthemes. Modern movie scripts, on the other hand, have much fuller directions, such as this from *Juno*:

> We push in over Bleeker sleeping in his car-bed towards the window. We look out onto the lawn to find Juno and Leah running back to the Previa, hopping in, and screeching off.

Such directions are generally direct, declarative, and frankly descriptive, including liberal use of phonaesthematic words; they are a distinct register from the dialogue. The genre “movie script” contains both registers: dialogue and directions. Movie-goers experience only the dialogue verbally, but the script is the literature under consideration, and the directions evidently add to the overall phonaesthematic count. An interesting question meriting further study is whether there is an interaction effect: whether movies with more phonaesthemes in the directions have more in the dialogue.

Stage directions in modern plays are not usually as replete as in movies (I say this from experience, having read several hundred plays in the course of my graduate career in drama), although they are more detailed than in plays of Sheridan's era; a play script is expected to allow some degree of interpretive freedom on the part of the director, designer, and actors in each production, whereas a movie script is more precise directions for a single performed output. Movie scripts are often adaptations from novels, too, and in such cases they may incorporate descriptive text taken from the novel.

The difference between plays of circa 1800 UK and movies of circa 2000 US shows us that literal performativity is not necessarily the most important factor in usage of phonaesthemes. Rather, it is personal expressivity, expected levels of usage, and social standards as expressed through the literature. The functional needs of movie directions do play a role, however, so while the fiction results have shown us that description does not require phonaesthemes, the movie script direction results show us that it can license them.

3.3.3 Magazines

Figure 3.9: US magazines across time, P/(P+S) with 95% CI

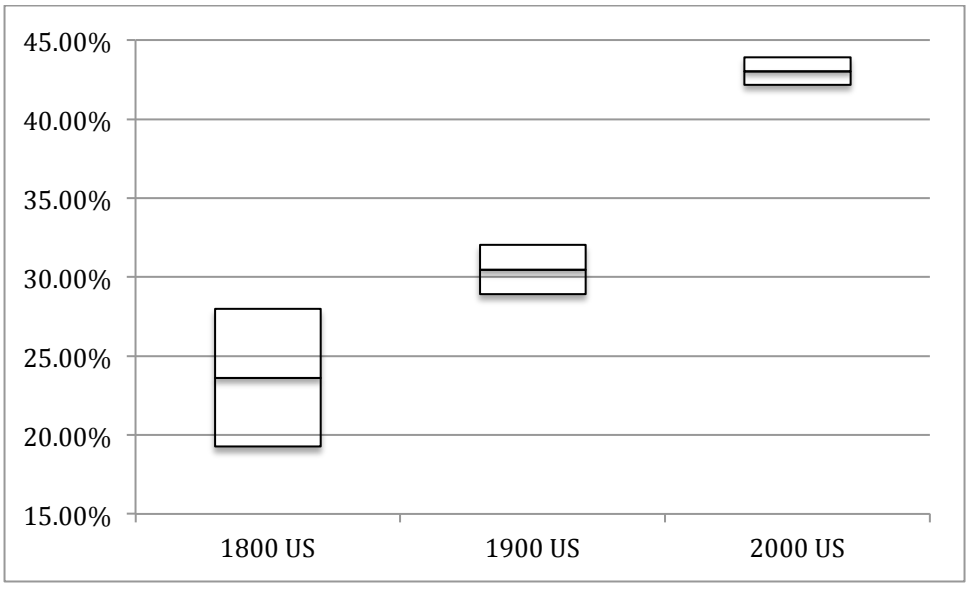
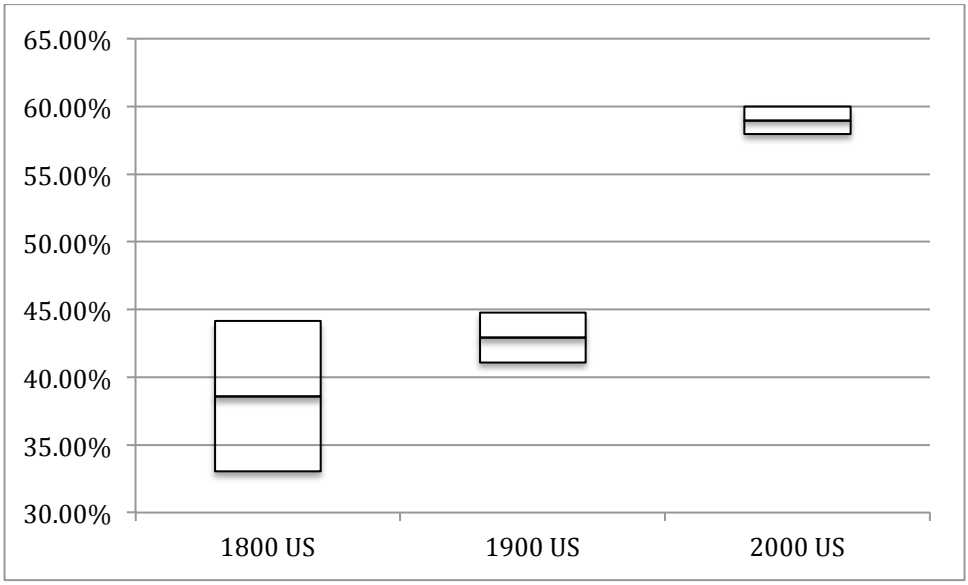


Figure 3.10: US magazines across time, $P/(P+S_G)$ with 95% CI



When we compare Figures 3.9 and 3.10, we see that the effect of the $\mathbf{S_R}$ set in American magazines makes a difference in 1900 but does not change the overall trend, which shows an increasing use of phonaesthematic words. We have seen that the $\mathbf{S_R}$ relative frequency is greater in general for magazines than for fiction; this is likely an effect of the substantial nonfiction component of magazines, which may have an approach more of fact analysis (informational, abstract) and less of engaged description (involved, non-abstract) – even without the influence of the $\mathbf{S_R}$ set, magazines still have a slightly lower frequency of phonaesthematic words than fiction. The magazines match more closely the essay genre in Biber's analysis (1995; see above) fairly well in both dimensions, involvement and non-abstractness: lower on the scale than fiction, and trending upward more steeply in the 1900–2000 century than in the 1800–1900 century. Nonetheless, they do roughly parallel the fiction in their general upward trend, consistent with other popular literary genres.

As we have seen in chapter 1 and above in this chapter, phonaesthematic words are not only more iconic and "vivid," which can mean involved and non-abstract, they can also seem more basic and less appropriate to the more polite levels of communication. It may be tempting to attribute the upward trend in phonaesthematic usage that we are seeing in magazines and other genres to a

decreasingly "erudite" approach in culture in general, but we ought to remember that literacy rates have risen substantially over the time period in question (Snyder 1993). Whereas readership of magazines in 1800 would have been limited to a comparatively high socioeconomic subset of the population, and would reflect the norms and distinctions of that status, they can now be read by a much broader population base. Indeed, the shift in level of the texts could as readily be an indication of an increase in erudition in the general populace: people who previously would not even have been reading are now reading, and while the tone of the literature has become less rarefied, this may simply be an averaging of the previous high (abstract, uninvolved) tone with a lower level that has been elevated to the point at which it can be added. So the increase in use of phonaesthematic words could be seen as a recognition of that segment of the population that sees such words as more commonly acceptable, and as an invitation to them to read. We will see a very clear manifestation of just this trend below in our analysis of newspapers and of Hansard, and a reacting counter-trend in humanities articles.

3.3.4 Newspapers

Figure 3.11: Newspapers across time, P/(P+S) with 95% CI

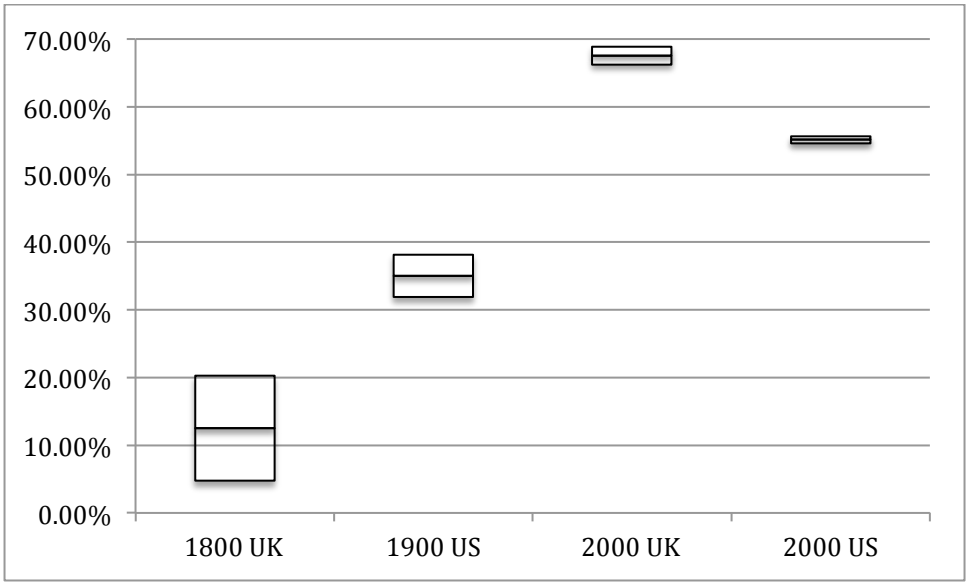
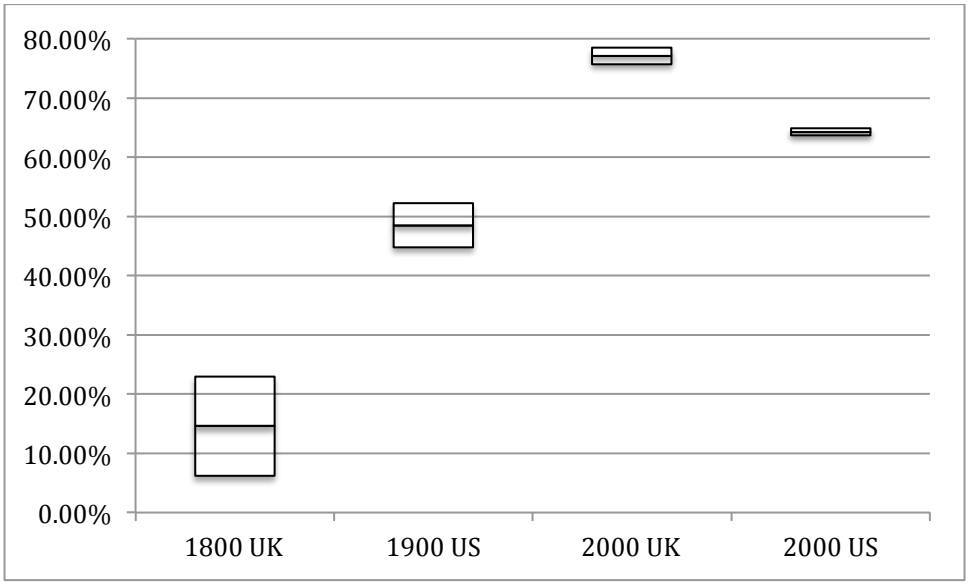


Figure 3.12: Newspapers across time, P/(P+S_G) with 95% CI



I have removed the *expand/expansion* data point for this calculation, since, as observed above, that single item is quite disproportionate in the 2000 results. We can see that in this case the $\mathbf{S_R}$ set makes no important difference – it has a small noticeable effect in the US 1900 corpus, but it does not change the overall pattern of relation. And that pattern is striking: UK newspapers have increased in use of phonaesthemes since 1800; so have American newspapers, but not by as much, and UK newspapers currently outstrip them by a clear margin. (Remember, however, that the UK 1800 set of newspapers is actually *ending* in 1800 and includes nearly a full century before 1800, due to the limitations of the corpus.)

The reasons for these results are subject to speculation, but we should make note of the existence of a sub-genre of tabloids in the UK that are not present as such in the American newspaper sphere – the US has its "supermarket tabloids" such as the *National Enquirer*, *Star*, and *Weekly World News*, but these are not considered part of the same genre as daily newspapers, whereas the UK tabloids such as the *Sun* and the *Mirror* are a very strong force in the daily newspaper business, and they trade on much more lurid – vivid, expressive, direct – approaches to the stories. They are exactly *not* the *Times* or the *Guardian*, and their market depends on this clear distinction. This was not the nature of the business 200 years ago, when those who

could read and could afford newspapers were a smaller set. To see what effect this tabloid genre has on the genre of newspapers as a whole, I have surveyed the tabloid subset of the UK 2000 newspapers to compare it with US newspapers, UK newspapers overall, and UK newspapers *excluding* tabloids. The results are in Figures 3.13 and 3.14.

Figure 3.13: Newspapers, US and UK compared with UK tabloids and UK minus tabloids, 2000, P/(P+S) with 95% CI

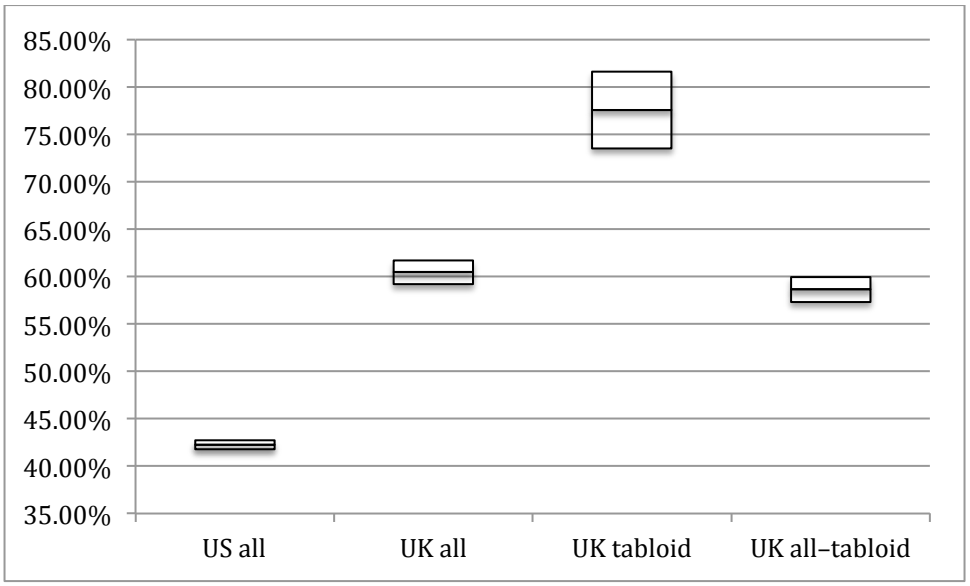
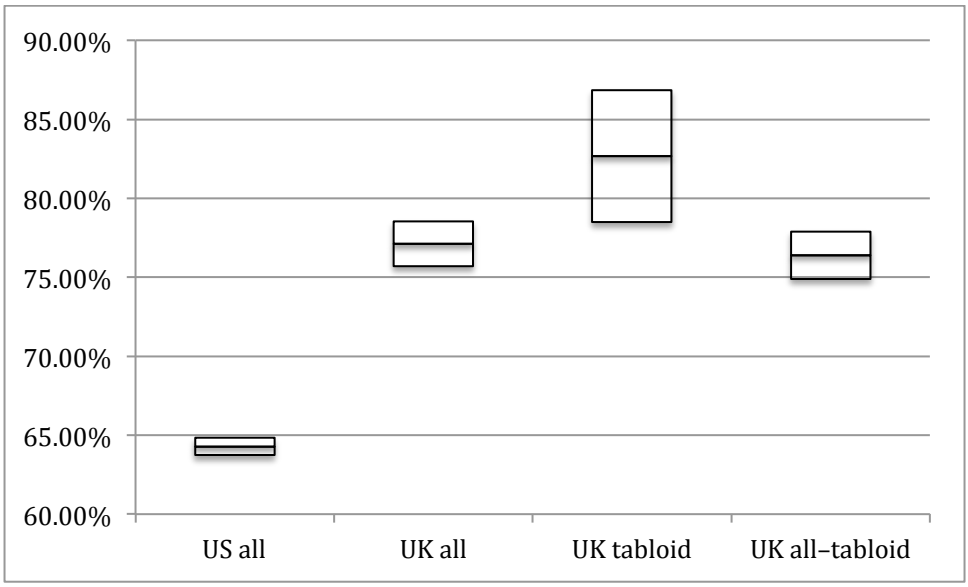


Figure 3.14: Newspapers, US and UK compared with UK tabloids and UK minus tabloids, 2000, $P/(P+S_G)$ with 95% CI



(Note that the US results are to the left, rather than the right, in these charts.) It is clear that the UK tabloids use extremely few polysyllabic Romance-derived words (see Appendix C for full details), much fewer than the other newspapers. Once set $\mathbf{S_R}$ is excluded, the results are closer, but there is still a thin statistically significant margin between the tabloids and the remainder (and an equally thin overlap of confidence intervals between the tabloids and the total results including the tabloids). The tabloids therefore give results very much as expected, but we can also see that they do not account for very much of the difference between the US and UK results. Even excluding the tabloids, the differences in phonaestheme usage between US and UK are noteworthy.

We should also remember the observation by Biber (1995, 297) that genres showed a general trend over their first 100 to 200 years to become more distinct from oral genres, which can help account for the level of the newspapers of earlier periods: as literature, they would have been expected to be more erudite and decorous than the main run of speech. The modern British newspapers are very distinct from oral usage in their structure and usage, but they are unblushingly full of iconic,

expressive usages. Consider an example such as this reportage from *The Telegraph* – certainly not a tabloid – on January 23, 2016:

> The Tory truce over Europe began to crumble on Saturday as David Cameron faced an angry backlash from MPs over his attitude to his party in the referendum campaign.

We will not want to miss the contribution of the headline, another register and structural element within the genre, to the overall expressive tendency: "Europe: the gloves are off as Tory rift widens." Headlines need to be short and punchy, and so they are very likely important contributors to the overall use of phonaesthemes in this genre. They also set the tone for the articles; a punchy headline gives licence to similar punchiness in the article, and indeed the reader might be disappointed to find the article much more reserved in usage than the headline. Parallel corpora of headlines and article bodies were not available for this study, but this is another potentially productive avenue for future research.

Compare the political reportage above with this crime report from a British newspaper of January 3, 1791, from the ZEN Corpus:

> On Sunday evening as Mrs. Sherrard, wife of Mr. Thomas Sherrard, watch-maker, Houndsditch, and her daughter, were coming up Petticoat-lane, Whitechapel, they were stopped at the corner of Gravel-lane, by three fellows, who dragged them a short way up Cock-hill, and robbed Mrs. Sherrard of half a guinea, two shillings, and a gold ring; and Miss Sherrard, of four shillings and a gold hair-worked locket. They behaved in a very indecent manner to Miss Sherrard, and on Mrs. Sherrard rebuking them, they attempted to serve her in the same manner, but were prevented by some persons coming up.

It is not merely literary but even studiedly dispassionate and understated. We do see *dragged them*, but then we see *behaved in a very indecent manner*, *rebuking*, and *attempted to serve her in the same manner*. It seems it is only barely proper to make

mention of the crime at all, and would be entirely unacceptable to speak more plainly and graphically of it.

We see from these results that newspapers have followed the same general upward trend as other genres. They have a variety of characteristics that distinguish them from genres such as fiction, to be sure, but the use of phonaesthemes is not especially salient among those. This suggests some independence of phonaestheme usage from Biber's two dimensions that I have cited. But no single factor loads at 100% in any dimension – nothing correlates exactly with an axis of variation. So it is not surprising that phonaesthemes exhibit some level of independence.

3.3.5 Humanities articles

Figure 3.15: Humanities articles across time, P/(P+S) with 95% CI

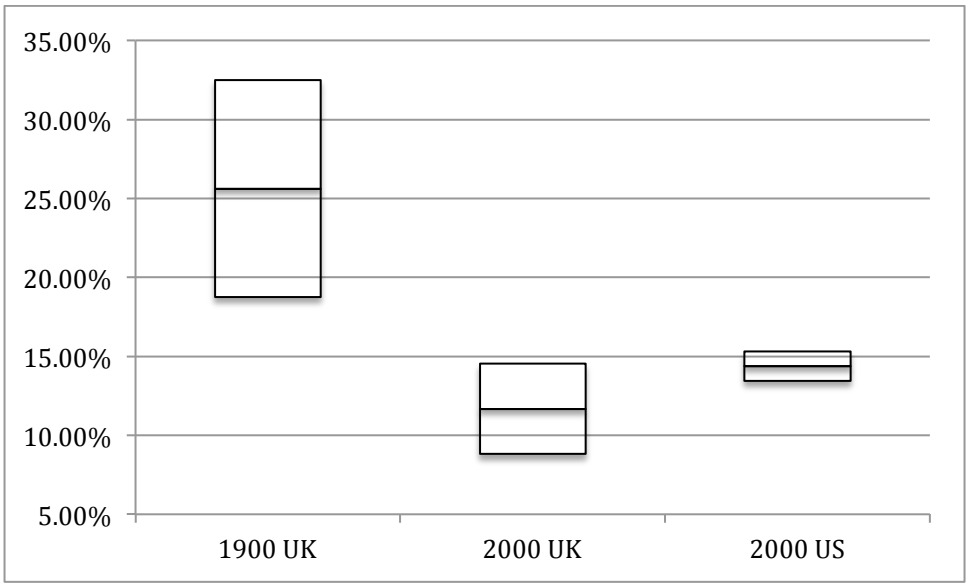
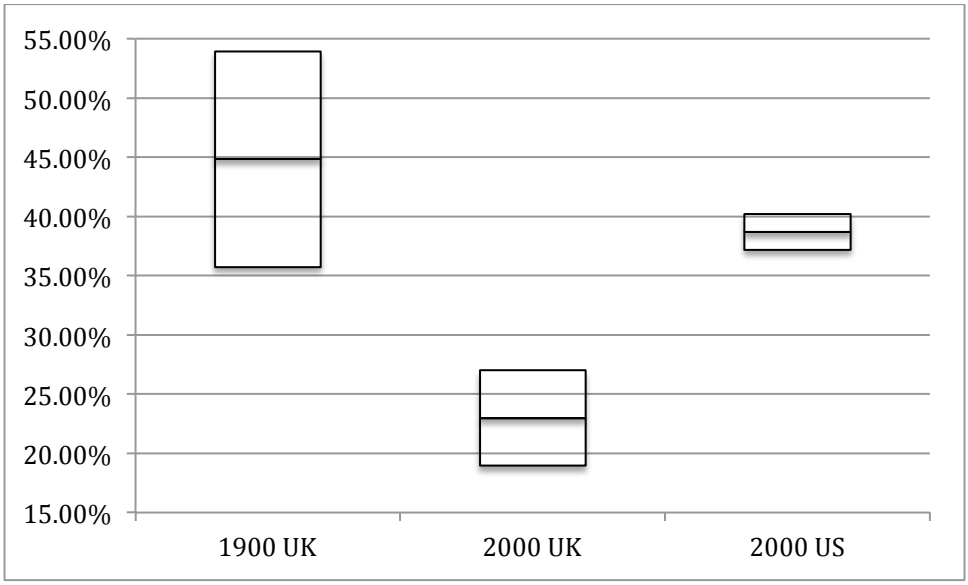


Figure 3.16: Humanities articles across time, P/(P+S_G) with 95% CI



When we compare Figures 3.15 and 3.16, we can see a clear effect of the poylysyllabic Romance-derived words, especially in the US 2000 humanities articles. We can also see a sharp decline from 1900 to 2000 in the UK humanities articles. The difference between 1900 and 2000 sets is at least partly explainable by an important difference in medium: the 2000 sets are composed of journal articles, whereas the 1900 set is composed of books – full-length works or books of essays. These are more in the line of what are now called "thinkpieces," whereas journal articles, even in the humanities, are expected to wear their sources and their academic approach on their sleeves. This is a field where my personal experience (in the course of completing an MA and PhD in a humanities discipline) attests that an article can be simultaneously praised for readability and criticized for informality. A certain level of abstract informationality is expected for *bona fides*.

Scholarly journals are an important part of modern academic discourse, and most scholarly journals in current publication had their starts in the 20th century. Biber (1995, 299) notes of scientific journals that they at first took a narrative presentation, but have over the years striven to obscure the human element and to present an air of disinterested objectivity, presenting information as facts existing independently of observers and presenting research processes and results as having

been done without mention of an agent – use of agentless passives has increased over time as one aspect of the "abstract" dimension Biber has identified (Biber 1995, 163, 291). Humanities journals appear to have borrowed to some extent on the model of the sciences. Thus the humanities journal article in 2000 owes some of its genesis to humanities books in 1900 and earlier, but some also to science articles, just as a thesis in linguistics does. The trajectory of phonaestheme usage in our humanities article results is consequently unsurprising. It is interesting that it is more strongly so in the UK than in the US; this may have some relation to the more class-stratified society of Britain, and it may be in reaction against the very demotic usages in popular media such as newspapers. There may also be greater senses of class distinction in the UK than in the US. Eschewing indecorous words helps to preserve the class status as well as to minimize the element of the observer and reporter – the articles do not resemble an involved personal recounting.

3.3.6 Hansard

Figure 3.17: British Hansard across time, P/(P+S) with 95% CI

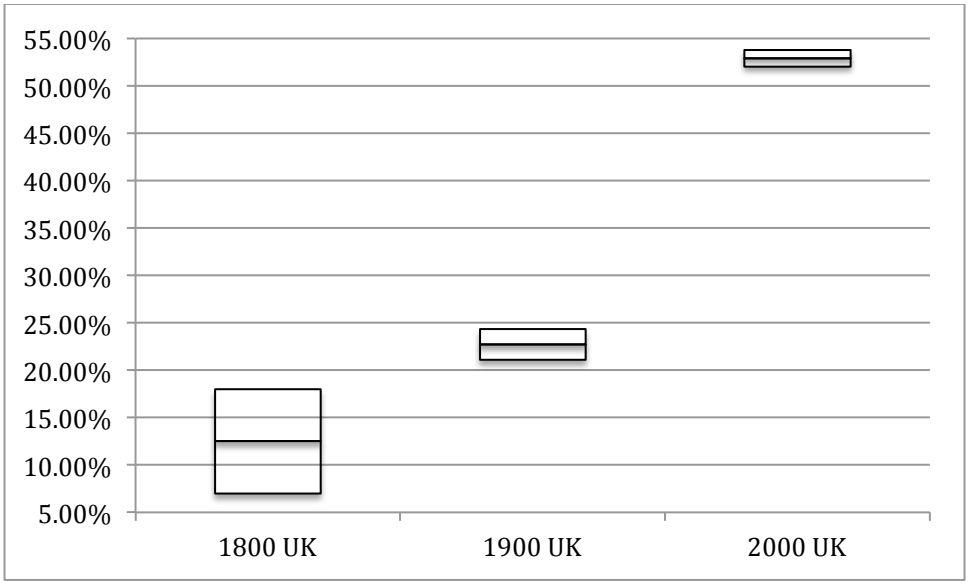
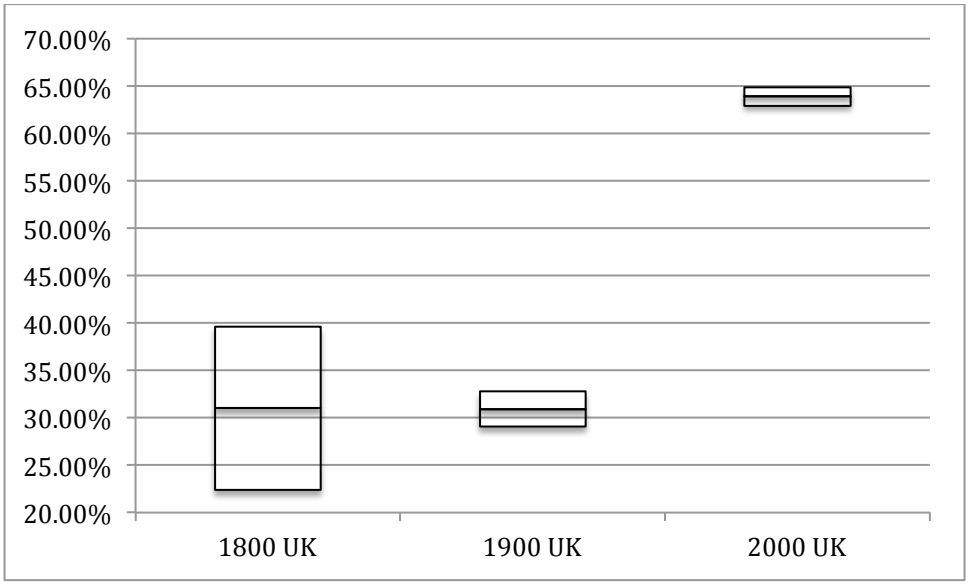


Figure 3.18: British Hansard across time, P/(P+S_G) with 95% CI



In making these figures, as with Figures 3.13 and 3.14, I have excluded the *expand/expansion* data point because of its disproportionate rise and its extreme influence in the 2000 data set: from 588 per 10,000 words in 1800 to 2300 in 1900 to 4999 – 50% of the data set – in 2000. By itself it made up 68% of the non-phonaesthematic word results in 2000, and led to calculations that produced a misleadingly low relative frequency of monosyllabic and Germanic words. With that data point in, the the **P**/(**P+S**) results are 11.76%, 17.49%, and 26.46% for 1800, 1900, and 2000, respectively; with it excluded, the **P**/(**P+S**) results are 12.50%, 22.71%, and 52.92%. This is well above the humanities and only slightly below the other genres for 2000. A further point of interest is that the $\mathbf{S_R}$ set has an important influence in the change from 1800 to 1900: when it is included, the two periods are statistically different, with the 1900 set showing a higher proportion of phonaesthemes than 1800, but when it is excluded there is no difference. This means that polysyllabic Romance-derived words formed a significantly more important part of the vocabulary in 1800. We can also see that their influence has waned substantially in 2000, with the exception of *expand/expansion*.

This signals a shift in the style and audience of parliamentary speeches. The audience for a parliamentary speech is, of course, first of all parliament, but in 1800

and 1900, it was, with rare exceptions, first, last, and only parliament. In 2000, with television cameras and a sophisticated news media looking for sound bites, and with people from a broad spectrum of backgrounds holding seats in parliament, the veneer of gentlemanly debate has largely been peeled off in favour of a more brightly coloured appliqué of high-impact language. Let us compare passages of speeches from the three times:

> I have no wish, and in one view certainly have no right, to speak with slight or disparagement of the abilities of the hon. gentlemen. Individually considered, they are all men of cultivated minds, of liberal education, of good natural endowments, not unread in the history of their country, not unpractised in its business, not unprovided with those talents and acquirements which are necessary for the conducting of business in this House. But if I am to speak of them collectively, as men forming the council which is to guide the affairs of a great empire, which is to rule the world in a crisis like the present, I must say, from whatever causes it arises, that they are weakness itself. I really believe the country will perish in their hands. I believe the hon. gentlemen will fairly see us out; that we shall not outlive their administration; that they will prove, as I believe, I once before took the liberty of remarking to them, the *Angustuli* in whose hands the empire will fall. There is an old joke which we may remember, of Cicero's, who when some person had ceased to be Consul on the same day on which he had been made, observed, that the person in question might tell of a prodigy which few of his predecessors could boast, for, that the sun had never set during his consulate. I wish that something equally prodigious may not be found in the history of the hon. gentlemen, and that it may not be to be said of them hereafter, that their administration lasted as long as the country. —*The Right Honourable W. St. Mawes Windham, November 23, 1803*

> Last year we had the Report of a Committee on the training of officers, and a most grave and serious condition of affairs was shown to have existed in the

> past, and, possibly, even to exist in the present. Many distinguished officers of the highest rank—some of whom, at all events, had been in a position to bring influence to bear on the problem—were unanimous in stating that the junior officers were lamentably wanting in military knowledge, and in the desire to acquire that knowledge, and they added that keenness was out of fashion. It does not seem to have occurred to these distinguished officers that it was the duty of the senior officers to see that those under them were trained to understand and to perform those duties on active service which are the very object of their existence as officers. For, if it was not for the certainty of war from time to time there would be no necessity for an Army at all, and if they themselves had been keen, in all human probability the young officers, too, would have been keen. —*Colonel Long (Worcestershire, Evesham), March 9, 1903*
>
> I do not think that I could speak in this debate without at least referring to speed cameras, which are hugely controversial, not least in Norfolk right at the moment. The most prominent recent case involves my constituent, Mrs. Jenny Mason. In January 2003, Mrs. Mason's 42-year-old son Andrew was killed in a head-on collision with a car that had been attempting, unsafely, to overtake a lorry on the A1066 in my constituency. Mr. Mason did not have a chance. He flashed his headlights twice. He braked hard and veered into the verge. Unfortunately he crashed into the car. He was catapulted into the air, struck the roof of the car and landed in the middle of the road. He was a motor cycle instructor and had an advanced driving certificate. He had been due to give his daughter away at her wedding four days later. His wife, Mairi, describes the impact of his death simply: "His mum lost her eldest son; I lost my husband and our three daughters their father. It has blown our lives apart." —*Mr. Richard Bacon (South Norfolk), September 8, 2004*

These are all quotations from prepared speeches, and while there is some selection effect, a speech such as the 2004 example would have been even more inconceivable

in 1803 than the 1803 example would be in 2004. The complexity of sentence structure is as obvious as the vocabulary, and the set of people described and appealed to is equally patent. As the composition of Biber's dimensions suggests, phonaesthematic words tend to go together with shorter, more direct sentences, and as our sources and data have shown, phonaesthematic words are expressive in an almost taking-the-necktie-off kind of way. The genre has developed in a distinctly demotic direction, as have politics in Britain (and elsewhere) generally.

3.3.7 Summary of historical development trends

Taking the historical trends in total, we have seen that most of them have increased in usage of phonaesthematic words since 1800, although humanities articles have decidedly not, and British fiction declined and rebounded. The development of English genres with regard to usage of phonaesthemes has been broadly consistent with the historical development on Biber's dimensions of "Involved versus Informational Production" (1995, 289) and "Non-abstract versus Abstract Style" (1995, 291): an increase since 1800 in the popular genres, but a decrease in the learnèd texts. The association of phonaesthemes with these dimensions may be variable over time and between countries, perhaps as a function of association of expressivity with the speaker (involved versus informational) versus with the text (abstract versus non-abstract); there may also be effects of social class codes, as mentioned above in relation to fiction. Whatever the case, there is a clear bifurcation developing, and phonaesthemes are playing an evident role in that – and clearly skewing away from the "learnèd," and this is not simply a function of word length and origin.

Other genres, when good corpora are available, could also be compared; one that would be most interesting to see would be business writing, which has a reputation for privileging expensive-sounding words but also has a reputation for liking "impactful" usages. Which way would it trend in usage of phonaesthemes? This is a potential future avenue of research, but it will by necessity start with the assembly

of a usable corpus. The work of giving it a basis for comparison has been done in this thesis.

3.4 Synchronic across genres

Looking at the genres compared synchronically in the different time periods will give us another angle of insight. We have looked at their relative historical development, but that does not clearly display the relationships of genres one to another within each time period. The overall tables do show that, but not with as much visual clarity, which is not possible with that much information. They are also at least slightly misleading in two genres, newspapers and Hansard, for the *expand/expansion* data point, which is overrepresented in the 2000 time period for reasons external to the research question. I did not wish to exclude it from the data overall because it was a usable data point for the other genres and helped match the **P** and **S** sets semantically. For the sake of comparison, however, I will exclude that data point from all genres for the charts in this section.

In the following figures I will show, as I did in section 3.3, the data points as the midpoint in floating bars extending over the confidence interval, and I will show two figures for each time period, **P**/(**P**+**S**) and **P**/(**P**+$\mathbf{S_G}$).

3.4.1 Circa 1800

Figure 3.19: Genres in 1800, P/(P+S) with 95% CI

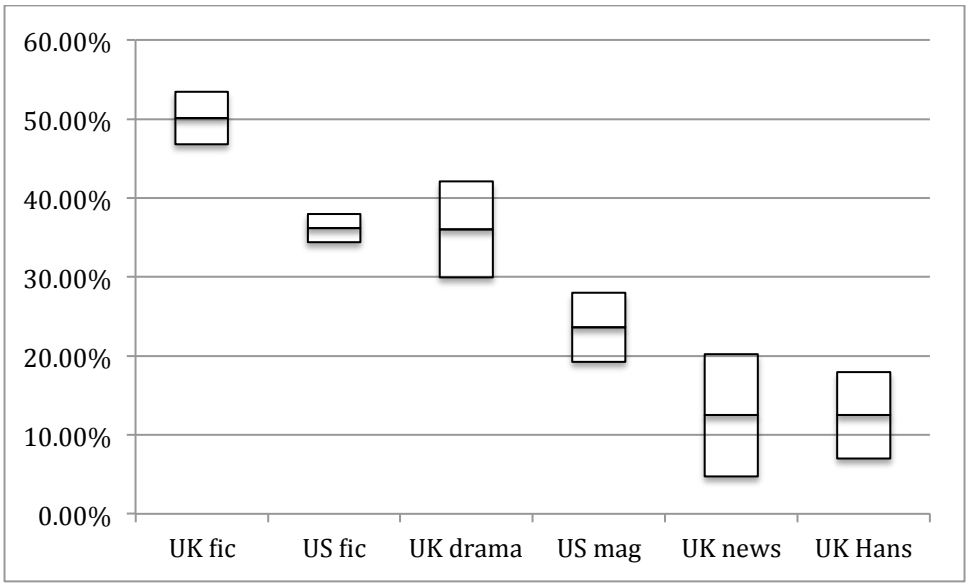
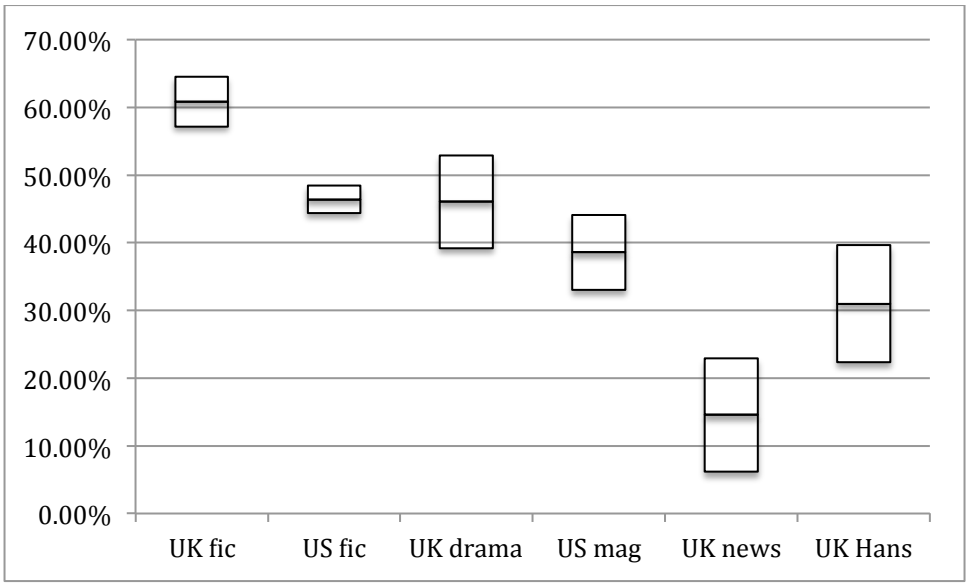


Figure 3.20: Genres in 1800, P/(P+S_G) with 95% CI



We can see clearly here where there are significant differences. UK fiction, in spite of its genteel tone, prevails distinctly over all others in use of phonaesthemes. US fiction and UK drama are even, which is quite interesting – not so much that they would be even but that the UK fiction would not be even with them, and would in

fact have higher usage rates than the drama. One possible cause of this is the lack of stage directions and scene descriptions in the plays; as I mentioned above, drama texts of the time consisted almost exclusively of dialogue, whereas novels included lengthy descriptions, and phonaesthematic words are often of the sort to be used in descriptions. But we have also seen that descriptions may be used in dialogue. There is also a question of audience – whether the audience for the UK novels was not exactly the same as that for UK plays or for US novels.

We see that the magazines are lower on the scale than the fiction and drama, but this is evidently in large part due to the use of polysyllabic Romance-derived words; the difference is not statistically significant once those are excluded. The newspapers, on the other hand, gain greater separation from the others in the $\mathbf{P}/(\mathbf{P}+\mathbf{S_G})$ figure, as they make comparatively little use of polysyllabic Romance-derived words. The parliamentary speeches in Hansard show considerable use of polysyllabic Romance-derived words, and when those are excluded the phonaesthematic words are seen to make up a greater proportion of the remainder (at least probably: there is a slight overlap in the confidence intervals). The overall results show a clear gradation across the genres; we can see the distinctions in Biber's (1989 and 1995) historical comparisons, and phonaesthemes have a clear relationship with the two axes I have brought into discussion here. Genres distinguish themselves from each other syntactically and lexically due to the demands of topic, medium, structure, and audience, and also due to systematic reinforcement and social classes of producers and receivers. Phonaesthemes clearly play a role in all of this; they carry not only an iconic effect but also a freight of tone and level.

3.4.2 Circa 1900

Figure 3.21: Genres in 1900, P/(P+S) with 95% CI

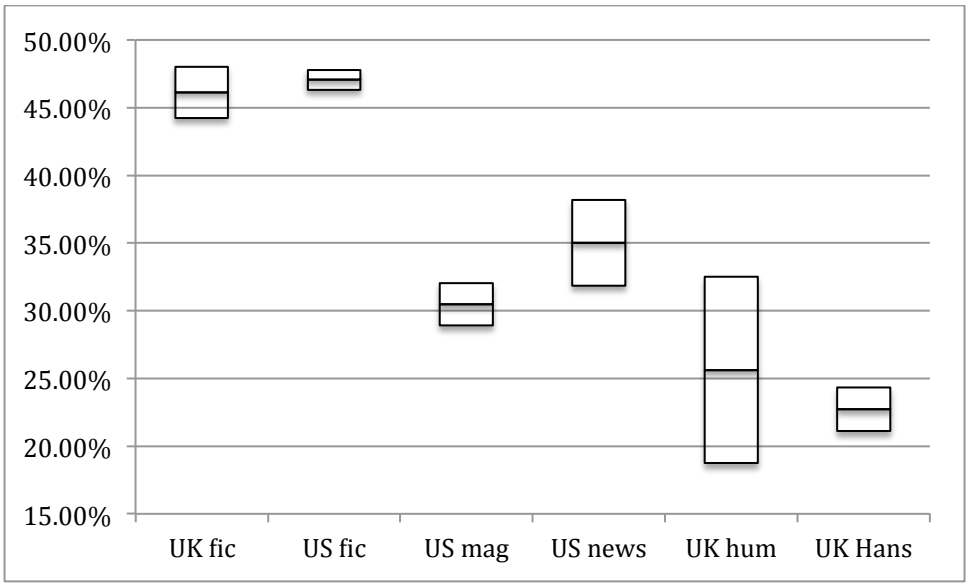
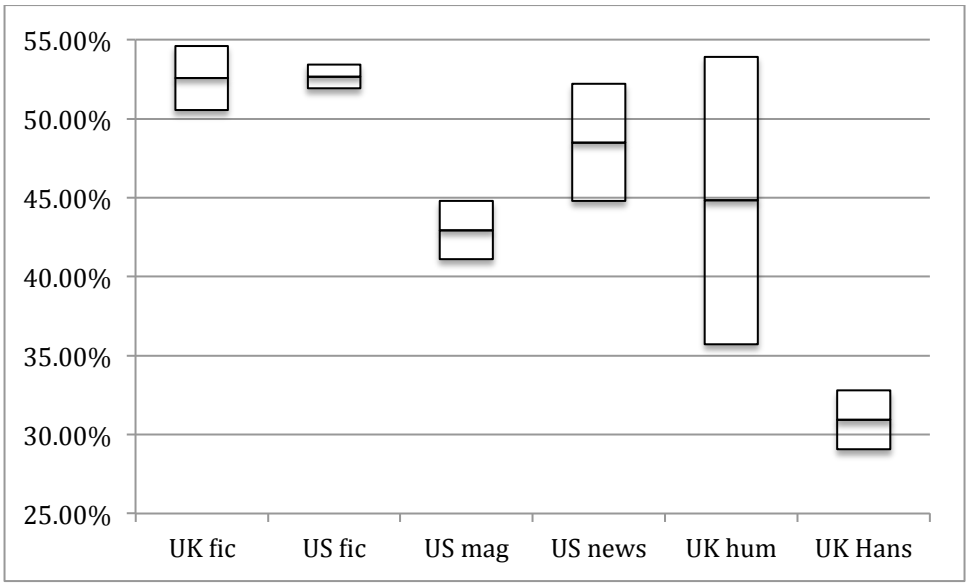


Figure 3.22: Genres in 1900, P/(P+S_G) with 95% CI



The stratification in the **P**/(**P**+**S**) figure (3.21) is, as we have seen, somewhat diminished in the **P**/(**P**+**S**$_G$) figure (3.22), but there is still some differentiation. The confidence intervals are very worthy of attention here for what they show to be significant and non-significant differences. The humanities sample has a large CI due

to its small size, but the difference between the **P+S** and **P+S$_G$** results is nonetheless significant – a large part of the vocabulary difference lies in the polysyllabic Romance-derived words. The Hansard is more clearly set off from the others. The magazines and newspapers are close, but distinct, with the newspapers more prone to use of phonaesthematic words, perhaps at least in part due to the nature of what they describe (although we have seen, from the early newspaper example, that it is possible to describe something quite violent using very unexpressive language). The fiction, in spite of what we have observed above about its level relative to 1800 and 2000, remains the most prone to using phonaesthemes.

3.4.3 Circa 2000

Figure 3.23: Genres in 2000, P/(P+S) with 95% CI

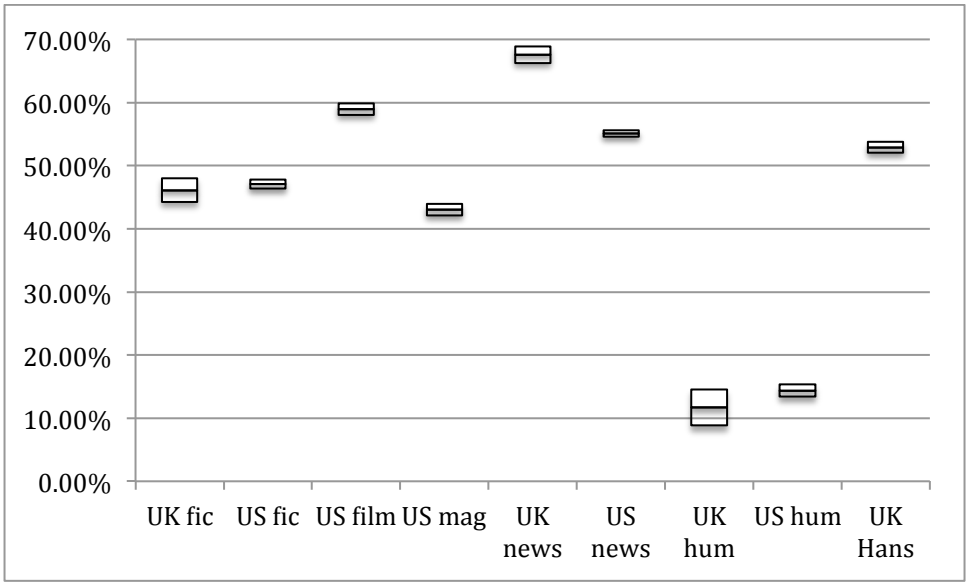
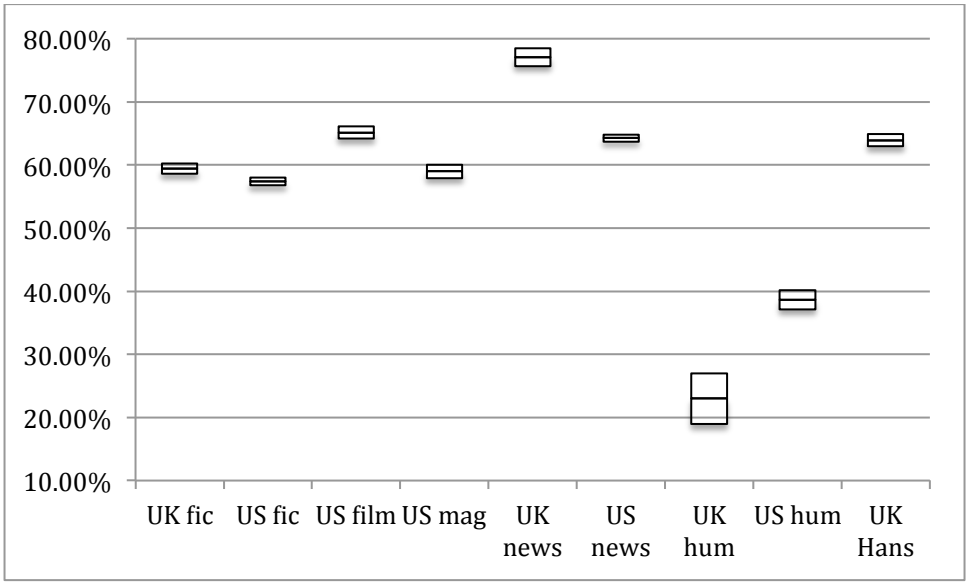


Figure 3.24: Genres in 2000, P/(P+S_G) with 95% CI



The most striking thing in the 2000 results is the size of the distinction between the humanities essays and the other results. It is interesting to note, too, how much more of the difference the $\mathbf{S_R}$ set accounts for in the US humanities articles than in the UK ones. This may be an effect of differing compositions of the genres, however.

The confidence intervals allow us to see that although the stratification is less in the $\mathbf{P}/(\mathbf{P}+\mathbf{S_G})$ figure, it is still significant. The Hansard, news, and film are above all the others, with the UK news holding the top spot by a clear margin. As we have seen above, this owes some but not nearly all to the tabloids. Fiction is no longer holding the top spot, which is an interesting finding relative to the development of fiction as a genre as well as to the others – fiction is now holding a comparatively slightly more "literate" and "learnèd" place, and so, with respect to phonaesthemes, a less expressive one, than the popular press. I use the terms "literate" and "learnèd" with full awareness of their implications: the stratification we see, and the qualities manifest in the text samples we have seen, show us that this is a distinction not just of expressivity but of level. Phonaesthematic words are not just childlike; they may be child*ish* in the eyes of writers and readers.

Let us look again now at the stratification Biber (1995) observed in relation to the dimensions "Involved versus Informational Production" (146) and "Non-abstract versus Abstract Style" (165) – I include these charts here again for convenience:

Figure 3.25: "Involved versus Informational Production" across genres (Biber 1995, 146)

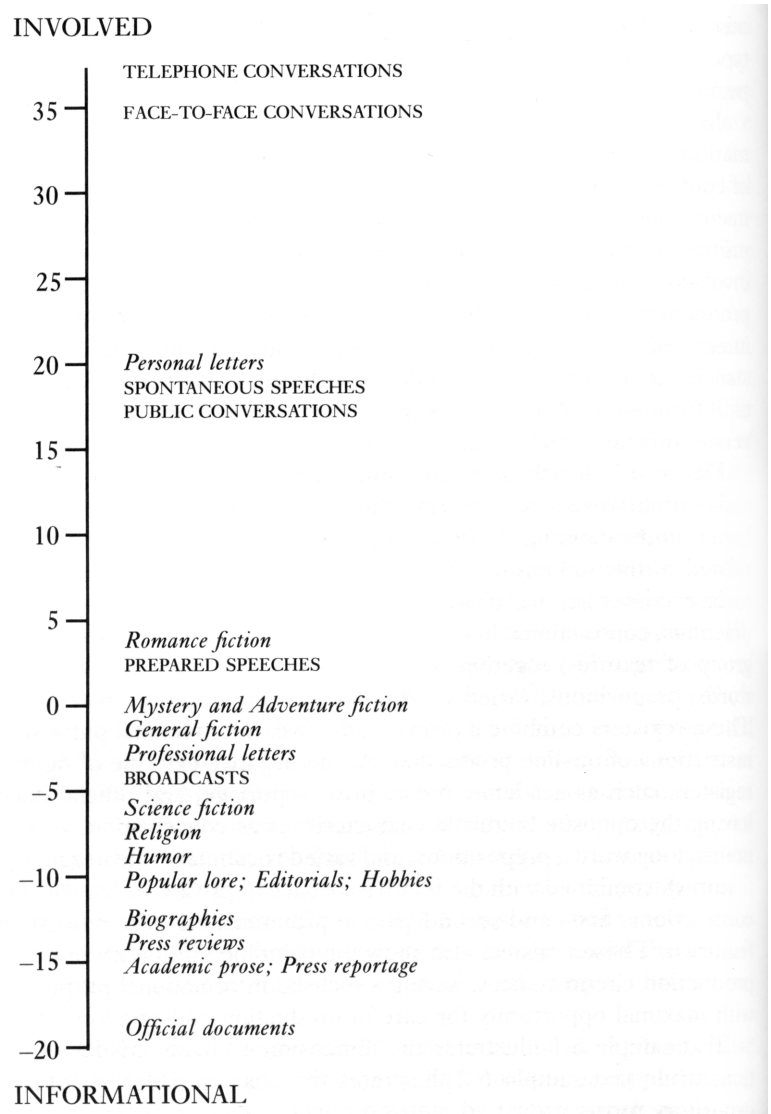


Figure 6.1 Mean scores of English dimension 1 for twenty-three registers: 'Involved versus Informational Production.' ($F = 111.9$, $p < 0.0001$, $r^2 = 84.3\%$)

Figure 3.26: "Non-abstract versus Abstract Style" across genres (Biber 1995, 165)

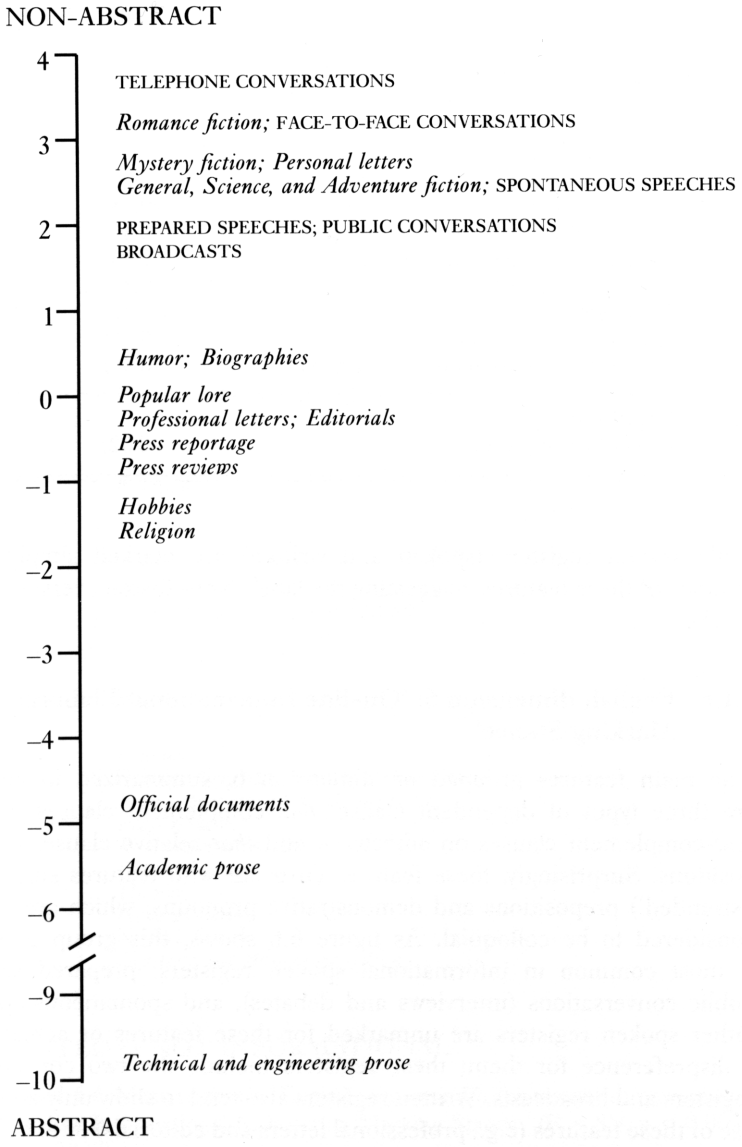


Figure 6.5 Mean scores of English dimension 5 for twenty-four registers: 'Non-abstract versus Abstract style' (Polarity reversed). ($F = 28.8$, $p < 0.0001$, $r^2 = 58.0\%$)

The results I have gotten are to some extent a match to Biber's results, but there is a striking departure: Biber's "press reportage" and "press reviews" are *more* informational and abstract than fiction and prepared speeches, whereas the newspaper genre as a whole as we have seen it makes *greater* use of phonaesthemes. This shows again that phonaestheme usage does not inevitably correlate with either dimension. Newspapers strive to appear impartial and fact-

based, and so will use diverse strategies to reduce the appearance of personal involvement of the authors; at the same time, however, they are in the business of engaging the audience and selling papers. In a business where "If it bleeds, it leads" is a well-known axiom, it is not surprising to see more vivid and evocative language used, even at the risk of seeming less learnèd – if indeed that is a risk: newspapers often present themselves as voices for the "little guy" or the common people to hold those in power to account, and so they would not want to seem elitist even as they strive to seem objective and authoritative.

3.4.4. Summary of synchronic comparisons

In every time period, there is discernible stratification in usage levels of phonaesthematic words between genres. The stratification is not identical over time, however. Genres such as newspapers and parliamentary speeches that showed low usage of phonaesthematic words in 1800 have come to the top by 2000; I have discussed in section 3.3 some likely reasons for this shift. Humanities articles have trended in the opposite direction, as we have seen. The clustering is tightest in the 1900 sets, which may suggest a greater homogeneity but may also mask a considerable diversity by averaging it out, as we have seen with American fiction; it may also reflect the adaptations necessary to accommodate and invite an increasingly broad populace.

What the data we have seen tell us about the relation between phonaesthemes and genre is that their expressivity seems to relate to frankness and decorousness, and to work against an impression of disinterested intellectual pursuit. They tend to work in concert with involved rather than informational production, though not invariably so, and with non-abstract rather than abstract presentation, though again not absolutely. We may have expected that they would factor in more where there is literal performance, but we have seen that this is not the key detail; the expectations of genre, as informed by its origins, standards, and social situation, are the key matters. Phonaesthemes appear to serve as *bona fides* for the common touch, and consequently are dispreferred in more elitist genres. Structural requirements of

genres can come into play as well – for example directions in movie scripts and headlines in newspapers; the full extent of this is matter for further study.

Chapter 4: Conclusion

My research question has been **"Does the presence of phonaesthemes in words play a role in the constitution and evolution of genres?"** The results of my research indicate that it does.

Phonaesthemes are phonemic clusters such as /spl–/ and /–æʃ/ that are associated at greater-than-chance frequency with specific semantic areas in etymologically unrelated words; they are generally seen as more vivid and expressive than the average word. Phonaesthemes are connected to sound symbolism, onomatopoeia, and ideophones through their systematicity and through their iconicity – an iconicity that is not necessarily strictly imitative but that at least draws on qualities of sound symbolism, which may be gestural, proprioceptive, or broadly imitative of temporal or sonic qualities, directly or metaphorically applied. Phonaesthematic words are not a special class of words, and are not especially impervious to inflections or to sound change over time, but phonaesthemes do have an apparent influence on sense and usage through the same kind of systematic imitation-based effects seen more generally in language. They have a function that is similar to that of morphemes and pseudomorphemes, but they are not compositional in the same way; they may or may not be present, and if they are they overlap with morphemes. Their semantics, too, are not as clear-cut as with morphemes; they are more impressionistic and probabilistic. Nonetheless, experiments have shown them to have a psychological reality in reception as well as production. They are present in many (possibly all) languages, but they are not identical between languages, though there are familial effects. We do find that phonaesthemes appear most commonly in shorter, earlier-learned words, and so in English very often in Germanic words. It is not that words from other sources do not contain phonaesthemes, simply that they are less likely to. The reasons for this have to do with systematicity but also with the

tone and level of usage of these words, which are seen as more fit for storybooks than for dissertations.

It is intuitively appealing to hypothesize that phonaesthemes would play a role in register and genre construction and distinctions. The terms *register* and *genre* are used variously loosely and often in overlapping ways; I am using *register* to refer to a set of patterns of morphosyntax and lexis used in a specific situation, and *genre* to refer to a text intended for a certain purpose and audience, having an expected structure as well as one or more expected registers. Both genre and register exist at many different levels; *fiction* is a genre, as is *science fiction*, as is *steampunk science fiction*. Genre may be constituted institutionally (e.g., various types of fiction and poetry) or may have an ad hoc existence (e.g., *management notices in condominium elevators*, which is a subgenre of the larger genres *business communication* and *public advisories*). Use of phonaesthemes will set the tone for a particular usage, will indicate a disposition towards the content (a vivid description, for example), and will also say something about the utterer, the intended hearer, and the relationship between the two; it is thus a natural aspect of consideration in register. But it is also worth seeing from the perspective of genre, not just because genres make use of registers but because genres have an existence of their own, with mutual influence from the different structural aspects they make use of. Genres can be examined at many different levels, although something such as use of phonaesthematic words requires a level with sufficient data in order to have any statistical authority. Genres come into existence on the basis of other genres, and develop also in contrast with other genres. They vary in numerous ways, and it can be seen that many features tend to co-vary because they have the same or similar effect and/or social profile of usage. Thus those who write or speak in a genre will choose certain features of language – morphosyntactic and lexical, and also, as we see, phonological and semantic – to serve their purpose and effect and to display what genre they are making use of, and at the same time those features will feed back into the tone and function of the genre. Genres develop over time according to social circumstances and needs as well as material circumstances (for example, the greater need and

expectation for directions in movie scripts than in stage play scripts). It stands to reason that phonaesthemes should vary in usage within a genre over time and between genres at any given time.

My research results have shown that this is the case. Most genres have increased in use of phonaesthemes over the past two centuries, although scholarly writing as exemplified by humanities articles has headed strongly in the opposite direction. But there is stratification: newspapers, movie scripts, and political speeches all appear to use more phonaesthematic words than fiction or magazines, although in previous times fiction used more and newspapers were quite restrained in their use. We have seen that this variation is not simply an effect of word length and origin; since the lemmas chosen for study were matched for frequency of use, we have also seen that it is not simply an effect of frequency of use. At the same time, we have seen that words containing a phonaesthematic sound cluster but not having a sense related to the phonaestheme do not vary in the same way, meaning that there is not some other purely phonological effect operating. We have seen that use of phonaesthemes corresponds fairly well, but not absolutely, with dimensions of variation that have been discerned, specifically with involved rather than informational production and with non-abstract rather than abstract style. We have considered factors that may play into the use of phonaesthemes, and effects they may have in return – not simply (or even perhaps primarily) vividness of description but level and tone of usage. We found in the literature review that there is a tone and level effect from the shorter, earlier-learned Germanic words as opposed to the longer, later-learned, and largely classically derived words, but the data here show that there is more than that to the effect of phonaesthemes. It suggests that they may be seen as unrestrained and lacking in gentility or decorum. Injudicious intensive use of them in hopes of producing a polished effect may sustain attention but may also provoke hostile or derisive responses when seen in an inappropriate context, but in any event they have undeniable impact. That is to say, if you dump a clump of phonaesthematic words into your work, you may hope to make it glisten, but while your readers are unlikely to snore, they may glare, curl

their lips, and snort if you use them where they don't fit in. But however you slice it, phonaesthematic words make quite a splash.

The results of this research project thus support the reality of phonaesthemes and give a clearer sense of their actual effect in use, and also provide a solid basis for future research efforts to build on. Possible future research efforts include differences in phonaestheme usage between specific genres of fiction and even individual authors; differences between different parts of works in genres, such as the beginnings of action-oriented novels versus the remainder, or between headlines and body text in newspapers; differences between electronic and written communication, and between different forms of electronic communication, for instance whether the brief and lively performances of Twitter or the detailed descriptions of personal emails are more likely to enfranchise phonaestheme use; differences in phonaestheme use between different genres of poetry, and the extent to which they match the prevailing standards in other literature at the time and/or reflect a contrary proclivity of the poets; relative frequency of phonaestheme usage in business writing, advertising, and other demotic genres; relative frequency in other academic genres; and experiments with live respondents testing responses to varying frequencies of phonaestheme use in texts of particular genres. Research on phonaesthemes, and on sound symbolism and iconicity in language more generally, is a young but burgeoning field. With the currently rapid expansion of available corpus data and computing power, such relatively impressionistic aspects of language as phonaesthetics are certain to be understood far better, and with them the fuller nature of language and its usage.

Bibliography

Abelin, A. (1999) Studies in sound symbolism. Göteborg University dissertation.

Alexander, M., and Davies, M. (2015–) *Hansard Corpus 1803–2005.* Available at www.hansard-corpus.org.

Alpher, B. (1994) Yir-Yiront ideophones. In Hinton, Nichols, and Ohala (1994), 161–177.

Atkinson, D. (1999) *Scientific discourse in sociohistorical context: the Philosophical Transactions of the Royal Society of London, 1675–1975.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Austerlitz, R. (1994) Finnish and Gilyak sound symbolism – an interplay between system and history. In Hinton, Nichols, and Ohala (1994), 249–260.

Bakhtin, M.M. (1986) *Speech Genres and Other Late Essays*. Ed. Caryl Emerson and Michael Holquist. Austin: U of Texas P.

Banham, M., ed. (1988) *The Cambridge Guide to World Theatre.* Cambridge, UK: Cambridge University Press.

Bawarshi, A. (2000) The Genre Function. *College English* 62(3): 335–360.

Bednarek, M. (2006) *Evaluation in media discourse: analysis of a newspaper corpus.* London: Continuum.

Bergen, B.K. (2004) The psychological reality of phonaesthemes. *Language* 80(2):290–311.

Biber, D. (1986) Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings. *Language* 62: 384–414.

Biber, D. (1988) *Variation across speech and writing*. Cambridge, UK: Cambridge University Press.

Biber, D. (1995) *Dimensions of Register Variation: A cross-linguistic comparison*. Cambridge, UK: Cambridge University Press.

Biber, D., and Conrad, S. (2009) *Register, Genre, and Style.* Cambridge, UK: Cambridge University Press.

Biber, D., and Finegan, E. (1989) Drift and the evolution of English style: A history of three genres. *Language* 65: 487–517.

Biber, D., and Finegan, E., eds. (1994) *Sociolinguistic Perspectives on Register.* New York: Oxford University Press.

Bloomfield, L. (1933) *Language*. New York: Henry Holt.

Bolinger, D. (1950) Rime, assonance and morpheme analysis. *Word* 6: 117–136.

Bolinger, D. (1968) *Aspects of Language*. New York: Harcourt, Brace & World, Inc.

British National Corpus. Accessed at www.natcorp.ox.ac.uk.

Bruthiaux, P. (1996) *The discourse of classified advertising: exploring the nature of linguistic simplicity.* Oxford: Oxford University Press.

Cassidy, K., and Kelly, M. (1991) Phonological information for grammatical category assignments. *Journal of Memory and Language* 30: 348–369.

Cassidy, K., Kelly, M., and Sharoni, L. (1999) Inferring gender from name phonology. *Journal of Experimental Psychology: General* 128: 362–381.

Chapman, R.L. (1977) *Roget's International Thesaurus*, fourth edition. Fitzhenry & Whiteside.

Charles, M. (2006) The construction of stance in reporting clauses: a cross-disciplinary study of theses. *Applied Linguistics* 27(3): 492–518.

Cody, D. (2007) *Juno*. Movie script accessed at https://finearts.uvic.ca/writing/websites/writ218/screenplays/award_winning/Juno.html

Cope, B., & Kalantzis, M. (Eds.) (1993) *The powers of literacy: A genre approach to teaching writing*. London: Falmer Press.

Davies, M. (2004–) *BYU-BNC*. (Based on the British National Corpus from Oxford University Press). Available at corpus.byu.edu/bnc/.

Davies, M. (2008–) *The Corpus of Contemporary American English (COCA): 520 million words, 1990–present.* Available at corpus.byu.edu/coca/.

Davies, M. (2009) The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2): 159–190.

Davies, M. (2010–) *The Corpus of Historical American English (COHA): 400 million words, 1810–2009.* Available at corpus.byu.edu/coha/.

Diller, H., De Smet, H., and Tyrkkö, J. (2011) *The Corpus of Late Modern English Texts (CLMET),* version 3.0. Available via https://perswww.kuleuven.be/~u0044428/.

Dingemanse, M., Blasi, D.E., Lupyan, G., Christiansen, M.H., and Monaghan, P. (2015) Arbitrariness, Iconicity, and Systematicity in Language. *Trends in Cognitive Sciences* 19(10): 603–615.

Doyle, A.C. (2011) *The Sign of the Four.* Project Gutenberg, www.gutenberg.org. Last updated March 2, 2011.

Dubrow, Heather. (1982) *Genre*. London: Methuen.

Eckert, P. (2010) Affect, Sound Symbolism, and Variation. *University of Pennsylvania Working Papers in Linguistics* 15(2), Article 9.

Ferguson, C.A. (1983) Ferguson, Charles. Sports Announcer Talk: Syntactic Aspects of Register Variation. *Language in Society* 12: 153–172.

Ferguson, C.A. (1994) Dialect, Register, and Genre: Working Assumptions About Conventionalization. In Biber and Finegan (1994), 15–30.

Fitneva, S.A., Christiansen, M.H., and Monaghan, P. (2009) From sound to syntax: phonological constraints on children's lexical categorization of new words. *Journal of Child Language* 36: 967–997.

Flowerdew, L. (2005) An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: countering criticisms against corpus-based methodologies. *English for Specific Purposes* 24(3): 321–332.

Foster, H.W. (1797) *The Coquette: Or, the History of Eliza Wharton.* Available online at digital.library.upenn.edu/women/foster/coquette/coquette.html

Freedman, A., and Medway, P., eds. (1994) *Genre and the New Rhetoric*. Bristol: Taylor.

Fries, U., Lehmann, H.M., Ruef, B., Schnieder, P., Studer, P., auf dem Keller, C., Nietlispach, B., Engler, S., Hensel, S., and Zeller, F. (2004) *ZEN: Zurich English Newspaper Corpus, Version 1.0.* Available at http://es-zen.uzh.ch/. Zürich: University of Zürich.

Frye, N. (1957) *Anatomy of Criticism: Four Essays*. Princeton: Princeton University Press.

Geenberg, K.R. (2010) "Poor baby, you got a booboo!": Sound symbolism in Adult Baby Talk (ABT). Paper presented at NWAV 39. Accessed at www.academia.edu/1694840/_Poor_baby_you_got_a_boo-boo_Sound_symbolism_in_adult_Baby_Talk

Genette, Gerard. (1992) *The Architext: An Introduction*. Berkeley: University of California Press.

Gergory, P. (2004) *The Queen's Fool.* New York: Simon & Schuster.

Gross, A.G., Harmon, J.E., and Reidy, M.S. (2002) *Communicating science: the scientific article from the 17th century to present.* Oxford: Oxford University Press.

Halliday, M.A.K. (1978) *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. London: Arnold.

Hamano, S. (1994) Palatalization in Japanese sound symbolism. In Hinton, Nichols, and Ohala (1994), 148–157.

Harbeck, J. (2014) Do Phonaesthemes Make a Splash in Language Change? Unpublished research paper, York University.

Helprin, M. (2012) *In Sunlight and in Shadow.* New York: Mariner Books.

Hinton, L., Nichols, J., and Ohala, J.J., eds. (1994) *Sound Symbolism*. Cambridge University Press.

Hoad, T.F., editor. (1996) *Oxford Concise Dictionary of English Etymology.* Oxford: Oxford University Press.

Hock, H.H., and Joseph, B.D. (1996) *Language History, Language Change, and Language Relationship: An introduction to historical and comparative linguistics.* Mouton de Gruyter.

Householder, F.W. (1946) On the problem of sound and meaning, An English phonestheme (discussion in *Language* 23 Feb 1946).

Hundt, M., and Mair, C. (1999) "Agile" and "Uptight" Genres: The Corpus-based Approach to Language Change in Progress. *International Journal of Corpus Linguistics* 4(2): 221–242.

Hutchins, S.S. (1997) The psychological reality, variability, and compositionality of English phonesthemes. Dissertation Abstracts International 59, 4500B. University Microfilms no. AAT 9901857.

Hyland, K. (1998) *Hedging in Scientific Research Articles.* Amsterdam: John Benjamins.

Hyland, K. (1999) Talking to Students: Metadiscourse in Introductory Coursebooks. *English for Specific Purposes* 18(1): 3–26.

Hyland, K., and Tse, P. (2005) Hooking the reader: a corpus study of evaluative that in abstracts. *English for Specific Purposes* 24(2): 123–139.

Jameson, Fredric. (1981) *The Political Unconscious: Narrative as a Socially Symbolic Act*. Ithaca: Cornell University Press.

Jamieson, K.M. (1975) Antecedent Genre as Rhetorical Constraint. *Quarterly Journal of Speech* 61: 406–415.

Jespersen, O. (1922) *Language: Its nature, development and origin.* Allen and Unwin.

Joseph, B.D. (1994) Modern Greek *ts*: beyond sound symbolism. In Hinton, Nichols, and Ohala (1994), 222–236.

Kaufman, T. (1994) Symbolism and change in the sound system of Huastec. In Hinton, Nichols, and Ohala (1994), 63–75.

Kelly, M., Springer, K., and Keil, F. (1990) The relation between syllable number and visual complexity in the acquisition of word meanings. *Memory and Cognition* 18: 528–536.

Kroch, A., Santorini, B., and Delfs, L. (2004) The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). Department of Linguistics, University of Pennsylvania. Available at www.ling.upenn.edu/hist-corpora/.

Lee, D.Y.W. (2001) Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology* 5(3): 37–72.

Library of Congress. (2014) Congress.gov. Database of legislation and congressional records. Accessed at beta.congress.gov.

MacDonald, S.P. (2005) The language of journalism in treatments of hormone replacement news. *Written Communication* 22(3): 275–297.

Magnus, M. (2001) What's in a word? Evidence for phonosemantics. University of Trondheim dissertation.

Malkiel, Y. (1994) Regular sound development, phonosymbolic orchestration, disambiguation of homonyms. In Hinton, Nichols, and Ohala (1994), 207–221.

Marchand, H. (1959) Phonetic symbolism in English word formations. *Indogermanische Forschungen* 64: 146–68.

Markell, N.N., and Hamp, E.P. (1960) Connotative meanings of certain phoneme sequences. *Studies in Linguistics* 15: 47–61.

Maurer, D., Pathman, T., and Mondloch, C.J. (2006) The shape of boubas: sound–shape correspondences in toddlers and adults. *Developmental Science* 9(3): 316–322.

McGurk, H., and MacDonald, J. (1976) Hearing lips and seeing voices. *Nature* 264(5588): 746–748.

Miller, C.R. (1984) Genre as Social Action. *Quarterly Journal of Speech* 70: 151–167.

Mitford, N., ed. (1956) *Noblesse Oblige.* London: Hamish Hamilton.

Monaghan, P., Christiansen, M.H., and Chater, N. (2007) The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology* 55: 259–305.

Monaghan, P., Mattock, K., and Walker, P. (2012) The Role of Sound Symbolism in Language Learning. *Journal of Experimental Psychology Learning, Memory, and Cognition* 38(5): 1152–1164.

Monaghan, P., Shillcock, R.C., Christiansen, M.H., and Kirby, S. (2014) How arbitrary is language? *Philosophical Transactions of the Royal Society B* 369: 20130299, 1–12.

Nielsen, A., and Rendall, D. (2011) The sound of round: Evaluating the sound-symbolic role of consonants in the classic takete-maluma phenomenon. *Canadian Journal of Experimental Psychology* 65: 115–124.

Nygaard, L.C., Cook, A.E., and Namy, L.L. (2009) Sound to meaning correspondences facilitate word learning. *Cognition* 112: 181–186.

Ohala, J.J. (1994) The frequency code underlies the sound-symbolic use of voice pitch. In Hinton, Nichols, and Ohala (1994), 325–347.

Otis, K., and Sagi, E. (2008) Phonoaesthemes: a corpora-based analysis. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (ed.s B.C. Love, K. McRae, V.M. Sloutsky), pp. 65–70. Austin, TX: Cognitive Science Society.

Oxford English Dictionary. (2015) Oxford University Press. Accessed at www.oed.com.ezproxy.library.yorku.ca.

Perniss, P., Thompson, R.L., and Vigliocco, G. (2010) Iconicity as a general property of language: evidence from spoken and signed languages. *Frontiers in Psychology* 1: Article 227, 1–15.

Project Gutenberg. (2015) Available at www.gutenberg.org.

Ramachandran, V.S., and Hubbard, E.M. (2001) Synaesthesia: a window into perception, thought and language. *Journal of Consciousness Studies* 8: 3–34.

Reaser, J. (2003) A quantitative approach to (sub)registers: the case of 'Sports Announcer Talk'. *Discourse Studies* 5(3): 303–321.

Reder, S. (1981) The written and the spoken word: influence of Vai literacy on Vai speech. In Scribner, S., and Cole, M. *The Psychology of Literacy*, 187–199. Cambridge MA: Harvard University Press.

Reilly, J., and Kean, J. (2007) Formal distinctiveness of high- and low-imageability nouns: Analyses and theoretical implications. *Cognitive Science* 31: 1–12.

Reilly, J., Westbury, C., Kean, J., and Peelle, J.E. (2012) Arbitrary Symbolism in Natural Language Revisited: When Word Forms Carry Meaning. *PLoS ONE* 7 (August 6), e42286.

Rhodes, R.A., and Lawler, J.M. (1981) Athematic Metaphors. In *Papers from the Seventeenth Regional Meeting of the Chicago Linguistic Society*, ed.s R. Hendrick, C. Masek, and M.F. Miller, 318–342.

Rhodes, R.A. (1994) Aural images. In Hinton, Nichols, and Ohala (1994), 276–292.

Rosmarin, A. (1985) *The Power of Genre*. Minneapolis: University of Minnesota Press.

Sapir, E. (1911) Diminutive and augmentative consonant symbolism in Wishram. *Handbook of American Indian Languages Bureau of American Indian Ethnography*, Washington, D.C. Bulletin 40(1): 638–646.

Sapir, E. (1929) A Study in Phonetic Symbolism. *Journal of Experimental Psychology* 12: 225–239.

Sereno, J.A. (1994) Phonsyntactics. In Hinton, Nichols, and Ohala (1994), 263–275.

Shaw, George Bernard. (2015) *Arms and the Man.* Project Gutenberg, www.gutenberg.org. Last updated June 21, 2015.

Snyder, T., ed. (1993) *120 Years of American Education: A Statistical Portrait.* National Center for Education Statistics. Extracts available online at https://nces.ed.gov/naal/lit_history.asp

Steen, G. (1999) Genres of discourse and the definition of literature. *Discourse Processes* 28, 109–120.

Stotesbury, H. (2003) Evaluation in research article abstracts in the narrative and hard sciences. *Journal of English for Academic Purposes* 2(4): 327–341.

Taylor, A., Nurmi, A., Warner, A., Pintzuk, S., and Nevalainen, T. (2006) The York-Helsinki Parsed Corpus of Early English Correspondence (PCEEC). Department of Linguistics, University of York. Oxford Text Archive, first edition. Available at www-users.york.ac.uk/~lang22/PCEEC-manual/index.htm.

Thibodeau, P.H., Bromberg, C., Hernandez, R., and Wilson, Z. (2014) An Exploratory Investigation of Word Aversion. Paper presented at COGSCI 2014: The Annual Meeting of the Cognitive Science Society.

Todorov, T. (1990) *Genres in discourse*. Trans C. Porter. New York : Cambridge University Press.

Traunmüller, H. (1996) Sound symbolism in deictic words. *TMH-Quarterly Progress and Status Report* 37(2): 147–150.

Twain, M. (Clemens, S.L.) (2015) *The Adventures of Huckleberry Finn.* Project Gutenberg, gutenberg.org. Updated October 5, 2015.

Ure, J. (1982) Introduction: approaches to the study of register range. *International Journal of the Sociology of Language* 35: 5–24.

Vande Kopple, W.J. (1998) Relative Clauses in Spectroscopic Articles in the Physical Review, Beginnings and 1980: Some Changes in Patterns of Modification and a Connection to a Possible Shift in Style. *Written Communication* 15(2): 170–202.

Vilha, M. (1999) *Medical writing: modality in focus.* Amsterdam: Rodopi.

Voeltz, F.K.E., and Kilian-Hatz, C. (2001) *Ideophones*. John Benjamins.

Walker, J.A. (2010) *Variation in Linguistic Systems.* New York: Routledge.

Wharton, E. (2008) *The Age of Innocence.* Project Gutenberg, gutenberg.org. Posted on August 12, 2008.

Appendices

Appendix A. Phoneme selection scores

For each table, each column with relatedness scores is headed with the semantic common feature; some words have more than one. The **Root** column lists the lemmas; in some cases a word has multiple entries because there are multiple homonyms with different meanings and etymological sources. For words that are obviously related, such as *placable* and *placate*, only one is listed in the table. For words that are etymologically related to other words but not necessarily obviously so, the word is included but greyed out and the **Related** column indicates which other word the word is related to; these words are not scored or included in the calculation. Some tables have a **Comment** column, where I indicate such things as where a word is included in spite of not being in the dictionary of reference. The **Absolute** row gives the sum of scores. The **Relative** row gives the relative score, which is the absolute score divided by the number of scores (not including unscored related words).

bl–

Root	loud, air-induced sound	Related
blab	1	
black	0	
bladder	0	
blade	0	
blame	0	
blanch	0	
blancmange		blank
bland	0	
blandish		bland
blank	0	
blanket		blank
blare	1	
blarney	0	
blasé	0	
blaspheme	0	
blast	0.5	
blatant		bleat
blather	0.5	

blaze	0	
blaze	0	
blaze	1	
blazon	0	
bleach		bleak
bleak	0	
bleak	0	
blear	0	
bleat	0	
bleed	0	
blemish	0	
blench		blink
blend	0	
blende	0	
blenny	0	
blesbok		blaze
bless	0	
blight	0	
blighter		blight
blimp	0	
blind	0	
blindfold	0	
blink	0	
bliss	0	
blister	0	
blithe	0	
blithering		blather
blitz	0	
blizzard	0	
bloat	0	
blob	0	
block	0	
blond	0	
blood	0	
bloom	0	
bloom	0	
bloomer	0	
blossom		bloom
blot	0	
blotch	0	
blouse	0	
blow	1	
blow	0	
blowzy	0	
blub	0	
blubber	0	
bludgeon	0	
blue	0	

bluff	0	
bluff	0	
blunder	0	
blunderbuss		blunder
blunt	0	
blur	0	
blurb	0	
blurt	0.5	
blush	0	
bluster	0	
Absolute	**5.5**	
Relative	**0.08**	

fl–

Root	2 dimensional extended	loose motion	Related
flabbergast			flabby
flabby	0	0	
flaccid	0	0.5	
flag	0	0	
flag	1	0	
flag	0	0	
flagellant	0	1	
flageolet	0	0	
flagitious	0	0	
flagon			flask
flagrant	0	0	
flail	0	1	
flair	0	0.5	
flak	0	0.5	
flake	1	0	
flambeau			flame
flamboyant			flame
flame	0	0.5	
flamen	0	0	
flamingo			flame
flan	1	0	
flange	1	0	
flank	1	0	
flannel	0.5	0	
flap	1	1	
flare	0	0.5	
flash	0	0	
flask	0	0	
flat	1	0	
flatter			flat
flatulent	0	0	

flaunt	0	0	
flautist	0	0	
flavine	0	0	
flavour	0	0	
flaw	0	0	
flaw	0	0	
flax	0	0	
flay	1	0	
flea	0	0	
fleam	0	0	
fleck	0.5	0	
fledge	0	0	
flee	0	0	
fleece	0	0	
fleer	0	0	
fleet	0	0	
fleet	0	0	
flense	0	0	
flesh	0	0	
fletcher	0	0	
fleur-de-lis			flower
fleury			flower
flews	0	0	
flex	0	0	
flibbertigibbet	0	0.5	
flick	0	0	
flicker	0	0.5	
flight			fly
flimsy	0	0.5	
flinch	0	0	
flinders	0	0	
fling	0	0.5	
flint	0	0	
flip			flap
flippant	0	0	
flirt	0	1	
flit	0	1	
flitch	0.5	0	
flitter			flit
float	0	0	
flocculent	0	0	
flock	0	0	
flock	0	0	
floe	1	0	
flog	0	1	
flood	0	0	
floor	1	0	
flop			flap

flora			flower
florescence			flower
florid			flower
florin			flower
florist			flower
floss	0	0	
flotation			float
flotilla			float
flotsam			float
flounce	0	0	
flounce	0	1	
flounder	0.5	0	
flounder	0	1	
flour			flower
flourish			flower
flout	0	0	
flow	0.5	0	
flower	0	0	
flu	0	0	
fluctuate			flow
flue	0	0	
fluent			flow
fluff	0	0.5	
fluid			flow
fluke	0.5	0	
fluke	1	0	
fluke	0	0	
flummery	0	0	
flummox	0	0	
flump	0	1	
flunkey	0	0	
fluor			flow
flurry	0	1	
flush	0	0	
flush			flow
fluster	0	0.5	
flute	0	0	
flutter	0.5	1	
fluvial			flow
flux			flow
fly	0	0	
fly	0	0	
phlebotomy	0	0	
phlegm	0	0	
phlogiston			phlox
phlox	0	0	
Absolute	**14.5**	**16.5**	
Relative	**0.15**	**0.17**	

fr–

Root	chaos; excrescence	Related
fracas	1	
fraction	0	
fragile		fraction
fragrant	0	
frail	0	
frail	0	
framboesia	0	
frame	0	
franc	0	
franchise		frank
francolin	0	
franc-tireur		frank
frangible		fraction
frangipane	0	
frank	0	
frankincense	0	
franklin		frank
frankpledge		frank
frantic	1	
frap	0	
frass	0	
frater	0	
fraternal		frater
fratricide		frater
fraud	0	
fraught		freight
fraxinella	0	
fray	0.5	
fray	1	
freak	1	
freckle	0.5	
free	0	
freesia	0	
freeze	0	
freight	0	
frenzy	1	
frequent	0	
fresco	0	
fresh	0	
freshet		fresh
fret	0	
fret	0.5	
friable		

friar		frater
fricandeau	0	
fricassee	0	
fricative		friction
friction	0.5	
friend		free
frieze	0	
frieze	0	
frigate	0	
fright	0	
frigid	0	
frill	1	
fringe	1	
frippery	1	
frisk	1	
fritillary	0	
fritter		fry
fritter	1	
frivolous	1	
frizz	1	
frizzle	1	
fro		from
frock	0	
frog	0	
frog	0	
frolic	1	
from	0	
frond	0	
front	0	
frontispiece		front
frore		freeze
frost	0	
froth	1	
frou-frou	1	
froward	0	
forwn	0	
frowzy	0	
fructify		fruit
frugal	0	
fruit	0	
fruition		fruit
frumenty	0	
frump	0	
frustrate		fraud
frustum	0	
frutescent	0	
fry	0	
fry	0	

phrase	0	
phrenetic		frenzy
phrenology		frenzy
Absolute	**18**	
Relative	**0.25**	

gl–

Root	light	smoothness	Related
glabrous	0	1	
glacé	0.5	0.5	
glacial			glacé
glacis	0	0	
glad	0	0	
glade	0	0	
gladiator	0	0	
gladiolus			gladiator
glair	0	0	
glaive			gladiator
glamour	0.5	0	
glance	0.5	0	
gland	0	0	
glare	1	0	
glass	1	0	
glaucoma	0.5	0	
glaucous			glaucoma
glaze			glass
gleam	1	0	
glean	0	0	
glebe	0	0	
glee	0	0	
glen	0	0	
glenoid	0	0	
glib	0	1	
glide	0	1	
glim			gleam
glimmer			gleam
glimpse	0.5	0	
glint	1	0	
glissade	0	1	
glisten	1	0	
glister			glisten
glitter	1	0	
gloaming	1	0	
gloat	0	0	
globe	0	0	
glomerate	0	0	

glory	0.5	0	
gloss	1	0.5	
gloss	0	0	
glossary			gloss
glottis			gloss
glove	0	0	
glow	1	0	
glower	0.5	0	
gloxinia	0	0	
gloze	0	0	
glucose	0	0	
glue	0	0	
glum	0	0	
glume	0	0	
glut	0	0	
gluteus	0	0	
gluten	0	0	
glutton	0	0	
glycerine			glucose
glyph	0	0	
glyptic			glyph
Absolute	**12.5**	**5**	
Relative	**0.27**	**0.11**	

kl–

Root	**abrupt onset**	**together**	**Related**
clack	1	0	
clad			clothe
claim			clear
clairvoyance			clear
clam	0	1	
clamber			climb
clamour			claim
clamp	0	1	
clan	0	1	
clandestine	0	0.5	
clang	1	0	
clangour	0.5	0	
clank	1	0	
clap	1	1	
clap	0	0	
clarence	0	0	
clarendon	0	0	
claret			clear
clarify			clear
clarion			clear

clarity			clear
clash	1	0.5	
clasp	0	1	
class	0	0.5	
classic			class
clatter	1	0	
clause	0	0	
claustral			close
claustrophobia			close
clavichord			clavier
clavicle			clavier
clavier	0	0	
claw	0	0	
clay	0	0	
claymore	0	0	
clean	0	0	
clear	0	0	
cleat	0	0	clot
cleave	0	0	
cleave	0	1	
clef			clavier
cleft			cleave
clematis	0	0	
clement	0	0	
clench	0	1	
clerestory			clear
clergy			clerk
cleric			clerk
clerk	0	0	
clever	0	0	
clew	0	0	
cliché	0	0	
click	1	0	
client	0	0	
cliff	0.5	0	
climacteric			climax
climate	0	0	
climb	0	0	
clinch			clench
cling	0	1	
clinic	0	0	
clink	1	0	
clinker			clink
clinometer			clinic
clip	0	1	
clip	1	0	
clipper			clip
clique			click

cloaca	0	0	
cloak	0	0	
clock	0.5	0	
clod			clot
clog	0	0	
cloisonné			close
cloister			close
closet			close
closure			close
clot	0	1	
cloth	0	0	
cloud	0	0	
clough	0	0	
clout			clot
clove			cleave
clover	0	0	
clown	0	0	
cloy	0	0	
club	0.5	0.5	
cluck	1	0	
clue			clew
clumber	0	0	
clump	0.5	1	
clumsy	0.5	0	
cluster	0	1	
clutch	0	1	
clutter			clot
clypeus	0	0	
chlorine	0	0	
klepht	0	0	
kleptomania	0	0	
klipspringer	0	0	
kloof	0	0	
Absolute	**13**	**15**	
Relative	**0.19**	**0.22**	

kr–

Root	bent	noisy impact	clenching or restriction	Related
crab	0	0	0.5	
crack	0	1	0	
cracknel				crack
cradle	0	0	0	
craft	0	0	0	
crake	0	0	0	
cram	0	0	0.5	
crambo	0	0	0	

cramp	0	0	0.5	
cranberry	0	0	0	
crane	0	0	0	
cranium	0	0	0	
crank	1	0	0	
cranky				crank
crannog	0	0	0	
cranny	0	0	0.5	
crape	0	0	0	
crapulous	0	0	0	
crash	0	1	0	
crasis	0	0	0	
crass	0	0	0	
cratch	0	0	0	
crate	0	0	0	
crater	0	0	0	
cravat	0	0	0	
crave	0	0	0.5	
craven	0	0	0	
craw	0	0	0.5	
crawl	0	0	0	
crayfish	0	0	0	
crayon	0	0	0	
craze	0	0	0	
creak	0	0.5	0	
cream	0	0	0	
crease	0	0	0.5	
create	0	0	0	
creche	0	0	0	
credence	0	0	0	
credo				credence
creed				credence
creek	0	0	0	
creel	0	0	0	
creep	0	0	0	
cremate	0	0	0	
crenate				crenellate
crenellate	0.5	0	0	
creole	0	0	0	
creosote	0	0	0	
crepe	0	0	0	
crepitation	0	0.5	0	
crepuscular	0	0	0	
crescendo				crescent
crescent	1	0	0	
cress	0	0	0	
crest	0.5	0	0	
cretaceous	0	0	0	

cretin	0	0	0	
cretonne	0	0	0	
crevice	0	0	0.5	
crew	0	0	0	
crewel	0	0	0	
crib	0	0	0	
cribbage	0	0	0	
crick	1	0	0.5	
cricket	0	0.5	0	
cricket	0	0.5	0	
crikey	0	0	0	
crime	0	0	0	
crimp	0.5	0	1	
crimson	0	0	0	
cringe	0	0	0.5	
cringle	0	0	0	
crinkle	0.5	0	0.5	
crinoid	0	0	0	
crinoline	0	0	0.5	
cripple	0	0	0	
crisis	0	0	0	
crisp	0.5	0.5	0	
criss-cross	0	0	0	
criterion	0	0	0	
critic	0	0	0	
croak	0	0.5	0	
crochet	0.5	0	0	
crock	0	0	0	
crocket	0	0	0	
crocodile	0	0	0	
crocus	0	0	0	
croft	0	0	0	
cromlech	0	0	0	
crone	0	0	0	
crony	0	0	0	
crook	1	0	0	
croon	0	0	0	
crop	0	0	0	
croquet	0	0	0	
croquette	0	0.5	0	
crosier				cross
cross	0	0	0	
crotch	0.5	0	0.5	
crotchet	1	0	0	
croton	0	0	0	
crouch				crotch
croup	0	0	0	
croup	0	0	0	

croupier				croup
crow	0	0	0	
crowd	0	0	1	
crown	0	0	0	
crucial				cross
crucible				cross
crucifer				cross
crucifix				cross
crude	0	0	0	
cruel	0	0	0.5	
cruet				crock
cruise				cross
crumb	0	0	0	
crumpet	0	0	0	
crumple				cramp
crunch	0	1	1	
crupper				crop
crural	0	0	0	
crusade				cross
crush	0	1	1	
crust	0	0	0	
crustaceous				crust
crutch	0.5	0	0	
crux				cross
cry	0	0	0	
cryptic	0	0	0	
crystal	0	0	0	
chrestomathy	0	0	0	
chrism	0	0	0	
chromatic				chrome
chrome	0	0	0	
chronic	0	0	0	
chronicle				chronic
chrysalis	0	0	0	
chrysanthemum				chrysalis
chrysolite				chrysalis
chrysoprase				chrysalis
kraal	0	0	0	
kriegspiel	0	0	0	
kris	0.5	0	0	
krummhorn	0.5	0	0	
krypton	0	0	0	
Absolute	**10**	**7.5**	**11**	
Relative	**0.08**	**0.06**	**0.09**	

pl–

Root	abrupt onset	1-dimensional thick	Related
placable	0	0	
placard	0	0	
place	0	0	
placebo	0	0	
placenta	0	0	
placer	0	0	
placid			placebo
placket	0	0	
plafond	0	0	
plagal	0	0	
plagiary	0	0	
plague	0	0	
plaice	0	0	
plaid	0	0	
plain	0	0	
plaint	0	0	
plaintiff	0	0	
plait	0	0.5	
plan	0	0	
plane	0	0	
planet	0	0	
plangent	0	0	
plank	0	1	
plankton	0	0	
plant	0	0	
plantain	0	0	
plantigrade	0	0	
plaque	0	0	
plash	1	0	
plasma	0	0	
plaster	0	0	
plastic	0	0	
plastron	0	0	
plate	0	0	
plateau	0	0	
platen	0	0	plate
platform	0	0	plate
platitude	0	0	
platoon	0	0	
platter	0	0	plate
plaudit	0	0	
plausible	0	0	
play	0	0	
plea	0	0	
pleach	0	0	

plead	0	0	
pleasance	0	0	please
please	0	0	
pleat	0	0	
plebeian	0	0	
plectrum	0	0	
pledge	0	0	
pledget	0	0	
plenary	0	0	
pleonasm	0	0	
pleiosaurus	0	0	
plethora	0	0	
pleura	0	0	
plexus	0	0	
pliable	0	0	pliers
pliers	0	0	
plight	0	0	
plimsoll	0	0	
plinth	0	0	
plod	0	0	
plop	1	0	
plosive	1	0	
plot	0	0	
plough	0	0	
plover	0	0	
pluck	1	0	
plum	0	0	
plumage			plume
plumb	0	0	
plumbago			plumber
plumber	0	0	
plume	0	0.5	
plummet			plumb
plump	0	0	
plumule	0	0	
plunder	0	0	
plunge	0	0	
pluperfect			plus
plural			plus
plus	0	0	
plush	0	0	
plutocracy	0	0	
plutonic	0	0	
pluvial	0	0	
ply	0	0	
Absolute	**4**	**2**	
Relative	**0.05**	**0.02**	

skr–

Root	complex onset with white noise component	grating impact or sound	clenching or restriction	Related
scrabble	0	0.5	0	
scrag	0	0	0.5	
scramble	0	0	0.5	
scrannel	0	0	1	
scrap	0	0	1	
scrape	0.5	1	1	
scratch	1	1	0.5	
scrawl	0	0	0	
scream	1	0	0	
scree	0.5	0.5	0	
screech	1	1	0	
screed	0	0	0	
screen	0	0	0	
screw	0	0	0	
scribble				scribe
scribe	0	0	0	
scrimmage	0	0.5	0.5	
scrimp	0	0	1	
scrimshaw	0	0	0	
scrip	0	0	0	
script				scribe
scrivener				scribe
scrofula	0	0	0	
scroll	0	0	0	
scrotum	0	0	0.5	
scrounge	0	0	1	
scrub	0.5	1	0.5	
scrub	0	0	0	
scruff	0	0	0	
scrum				scrimmage
scrumptious	0	0	0	
scrunch	1	0.5	1	
scruple	0	0	0	
scrutator	0	0	0.5	
scry	0	0	0	
Absolute	**5.5**	**6**	**9.5**	
Relative	**0.18**	**0.19**	**0.31**	

skw–

An extra column has been added to exclude uncommon words, as this category has a notable portion of words that are unknown to most English speakers.

Root	compressed	Related	compressed (uncommon words excluded)
squab	0		0
squabble	0		0
squad		squadron	
squadron		square	
squails	0		
squalid	0		0
squall	0		0
squaloid	0		
squamose	0		
squander	0		0
square	0		0
squarrose	0		
squash	1		1
squash	0		0
squat	1		1
squaw	0		0
squawk	0		0
squeak	0		0
squeal	0		0
squeamish	0		0
squeegee		squeeze	
squeeze	1		1
squelch	1		1
squib	0		0
squid	0		0
squiffy	0		0
squiggle	0		0
squill	0		
squinch	0		
squint	1		1
squire	0		0
squirm	0		0
squirrel	0		0
squirt	0.5		0.5
squish	1		1
squit		squirt	
squitch	0		
Absolute	**6.5**		**6.5**
Relative	**0.20**		**0.25**

sl–

Root	smoothly wet	Related
slab	0	
slack	0	
slade	0	
slag	0.5	
slake	0.5	
slam		
slander	0	
slang	0	
slant	0	
slap	0	
slash	0	
slat	0	
slate	0	
slattern	0.5	
slaughter	0	
slave	0	
slaver	1	
slay	0	
sled	0	
sleek	0.5	
sleep	0	
sleet	0.5	
sleeve	0	
sleight	0	
slender	0	
sleuth	0	
slew	0	
slice	0	
slick	1	
slide	1	
slight	0	
slim	0	
slime	1	
sling	0	
slink	0	
slip		slippery
slip	0	
slippery	1	
slit	0	
slither	0.5	
slobber	1	
sloe	0	
slog	0	
slogan	0	
sloop	0	

slop	1	
slope	0	
slosh		slush
slot	0	
sloth	0	
slouch	0	
slough	1	
slough	0	
sloven	0	
slow	0	
sludge	1	
slug	0.5	
sluice	1	
slum	0	
slumber	0	
slump	0	
slur	0	
slurry	1	
slush		sludge
slut	0	
sly	0	
slype	0	
Absolute	**14.5**	
Relative	**0.23**	

sn–

Root	**nose**
snack	0
snaffle	0.5
snag	0
snail	0
snake	0
snap	0
snare	0
snark	0
snarl	0.5
snatch	0
sneak	0
sneer	0.5
sneeze	1
snick	0
sniff	1
snicker	0.5
snip	0
snipe	0
snivel	1

snob	0
snood	0
snook	1
snooker	0
snoop	0.5
snooze	0
snore	1
snorkel	0.5
snort	1
snot	1
snout	1
snow	0
snub	0.5
snuff	0
snuff	1
snug	0
Absolute	**12.5**
Relative	**0.36**

spl–

Root	complex onset with white noise component	wet and messy	division	Related
splanchnic	0	0	0	
splash	1	1	1	
splay	0.5	0.5	1	
spleen	0	0	0	
splendid	0	0	0	
splice	0	0	1	
splint				splinter
splinter	0.5	0.5	1	
split	0.5	0	1	
splodge	0.5	1	0	
splotch	0.5	1	0	
splurge	0	0.5	0	
splutter	1	1	0	
Absolute	**4.5**	**5.5**	**5**	
Relative	**0.38**	**0.46**	**0.42**	

spr–

Root	complex onset with white noise component	disarry, spreading	Related
sprain	0	0.5	
spraints	0	0	
sprat	0	0	
sprawl	0	1	
spray	1	1	

spread	0	1	
spree	0	1	
sprig	0	0	
spring	1	0.5	
springe			spring
sprinkle	0.5	1	
sprint	0.5	0.5	
sprit			sprout
sprite	0	0	
sprocket	0	0	
sprout	0.5	0.5	
spruce	0	0	
spry	0	0	
Absolute	**3.5**	**7**	
Relative	**0.22**	**0.44**	

str–

Root	1 dimensional, flexible	effort or constraint	Related
strabismus	0	0	
straddle			stride
strafe	0	1	
straggle			stretch
straight			stretch
strain	0	1	
strain	0	0	
strait			strict
strand	0	0	
strand	1	0	
strange	0	0	
strangle	0	1	
strap	1	0.5	
strategy	0	0	
stratum	0	0	
straw	0.5	0	
stray	0	0	
streak	0.5	0.5	
stream	0.5	0	
street	1	0	
strength			strong
strep			strophe
stress			strict
stretch	1	1	
strew			straw
striate	0	0	
strict	0	1	
stride	0	0.5	

strident	0	0	
strife	0	1	
strigose	0	0	
strike			streak
string	1	0	
stringent			strain
strip	1	0	
stripe	1	0	
strive			strife
stroke			strike
stroll	0	0	
stroma	0.5	0	
strong	0	1	
strontia	0	0	
strop			strap
strophe	0.5	0.5	
structure	0	0	
struggle	0	1	
strum	0	0	
struma	0	0	
strumpet	0	0	
strut	0.5	0	
struthious	0	0	
strychnine	0	0.5	
Absolute	**10**	**10.5**	
Relative	**0.26**	**0.27**	

tw–

Root	**twisting motion; rotatory**	**Related**
twaddle	0	
twain		two
twang	0	
twat	0	
tweak		twitch
twee	0	
tweed	0	
tweedle	0	
tween	0	
tweet	0	
tweezers	0	
twelve		two
twenty		two
twerp	0	
twice		two
twiddle	0.5	
twig	0	

twilight	0	
twill		two
twin		two
twine	1	two
twinge	0	
twinkie	0	
twinkle	0.5	
twirl	1	
twist	1	
twit	0	
twitch	0.5	
twite	0	
twitter	0.5	
twizzle		twiddle
two	0	
Absolute	**5**	
Relative	**0.21**	

–ərl

Root	**spinning or spiralling**	**Comment**
birl	1	not in Ox Etym
burl	0.5	not in Ox Etym
churl	0	
curl	1	
earl	0	
furl	1	
girl	0	
hurl	0	
knurl	0	not in Ox Etym
pearl	0.5	
purl	1	
purl	1	
skirl	0	
swirl	1	
twirl	1	
whirl	1	
merle	0	
squirrel	0	
note: omitting words formed on –al suffix plus productive root (e.g., demurral)		
Absolute	**9**	
Relative	**0.5**	

–æp

Root	**surface**	**sharp sound**	**Related**	**Comment**

burlap	0.5	0		
cap	1	0		
chap	0.5	0		
chap	0	0		
clap	0	1		
clap	0	0		
crap	0	0		not in Ox Etym
dewlap			lap	
fap	0	1		not in Ox Etym
flap	1	0.5		
frap	0	0		
gap	0.5	0		
handicap			cap	
hap	0	0		
kidnap	0	0		
lap	1	0		
lap	0.5	0.5		
lap	1	0		
map	1	0		
mishap			hap	
nap	0	0		
nap	1	0		
overlap			lap	
pap	0.5	0		
pap	0	0		
rap	0.5	1		
rap	0	0		
recap	0	0		not in Ox Etym
sap	0.5	0		
sap	0	0		
scrap	0	0		
scrap	0	0		
slap	0.5	1		
snap	0	1		
strap	1	0		
tap	0	0		
tap	0.5	1		
trap	0	0		
trap	0	0		
wrap	1	0		
yap	0	1		
zap	0	1		not in Ox Etym
Absolute	**12.5**	**9**		
Relative	**0.33**	**0.24**		

–æʃ

Root	hit, fragments	Related	Comment
abash	0		
ash	0		
ash	0		
balderdash	0		
bash	1		
brash	0.5		
calabash	0		
calash	0		
cash	0.5		
clash	1		
crash	1		
crash	0		
dash	1		
eyelash		lash	
fash	0		
flash	0.5		
gash	0.5		
gnash	1		
hash	0		
lash	1		
lash	0		
mash	1		
midrash	0		
plash	1		
potash	0		
rash	0.5		
rash	0		
sash	0		
sash	0		
slash	1		
smash	1		
splash	1		
stash	0		not in Ox Etym
succotash	0		
thrash	1		
trash	0		
Absolute	**14.5**		
Relative	**0.41**		

–ætʃ

Root	hold, come to hold	Related	Comment
bandersnatch	0		not in Ox Etym
batch	0.5		

catch	1		
cratch	0		
dispatch	0		
hatch	0		
hatch	0		
hatch	0		
latch	1		
match	0.5		
match	0		
nuthatch		hatch	
potlatch	0		not in Ox Etym
scratch	0		
snatch	1		
thatch	0		
kaffeeklatsch	0		
attach	0		
detach		attach	
Absolute	**4**		
Relative	**0.24**		

–up

Root	**curve**	**Related**	**Comment**
coop	0		
coup	0		
droop	1		
goop	0		not in Ox Etym
hoop	1		
hoop	0		
loop	1		
nincompoop	0		
poop	0		
scoop	0.5		
sloop	0		
snoop	0		
stoop	1		
stoop	0		
swoop	1		
troop	0		
troupe	0		
whoop	0		
croup	0		
croup	0		
group	0		
recoup	0		
roup	0		
soup	0		

drupe	0	
dupe	0	
stupe	0	
stupe	0	
Absolute	**5.5**	
Relative	**0.20**	
Relative*	**0.37**	*spelled *oop* only

–ap

Root	cessation of motion	Related	Comment
atop		top	
bop	1		not in Ox Etym
caltrop	1		
chop	1		
chop	0		
chop	0		
clop	1		not in Ox Etym
cop	0.5		
crop	0		
drop	1		
flop	1		
fop	0		
hop	1		
hop	0		
knop	0		
lollipop	0		
lop	0.5		
lop	0		
mop	0		
orlop	0		
plop	1		
pop	0		
pop	0		
prop	0		
shop	0		
slop	0		
sop	0		
stop	1		
strop	0		
swop	0		
top	0.5		
top	0		
whop	1		
Absolute	**11.5**		
Relative	**0.36**		

–sl

Root	frenzied or chaotic action	Related	Comment
hassle	1		not in Ox Etym
tousle	1		
tussle	1		
apostle	0		
bristle	0.5		
bustle	1		
bustle	0		
castle	0		
epistle	0		
forecastle		castle	
gristle	0		
hustle	1		
jostle	0.5		
nestle	0		
pestle	0		
rustle	0.5		
thistle	0		
throstle	0		
trestle	0		
whistle	0		
dismissal		missal	
missal	0		
vassal	0		
dossal	0		
counsel	0		
mussel	0		
tassel	0		
twissel	0		
vessel	0		
fossil	0		
fissile	0		
missile	0		
sessile	0		
consul	0		
cancel	0		
chancel	0		
parcel	0		
tercel	0		
council	0		
pencil	0		
stencil	0		
Absolute	**6.5**		
Relative	**0.17**		

–ʌmp

Root	heaviness and bluntness	3 dimensional solid
bump	1	1
chump	0	0
clump	1	1
dump	0.5	0
dump	1	0
flump	1	0
frump	0	0
grump	0.5	0
hump	1	1
jump	0	0
lump	1	1
lump	0	0
lump	0	0
mugwump	0	0
mump	0.5	1
plump	0.5	0
plump	0.5	1
pump	0	0
pump	0	0
rump	1	1
slump	1	0
stump	0.5	1
sump	0	0
thump	1	0
trump	0	0
trump	0	0
Absolute	**12**	**8**
Relative	**0.46**	**0.31**

–ʌst

Root	surface formation	force	Related
adjust			just
august	0	0	
bust	0	0	
bust	0	0.5	
combust	0	0	
crust	1	0	
disgust	0	0	
dust	1	0	
gust	0	1	
just	0	0	
lust	0	0.5	
must	0	0	

must	0	0.5
must	0.5	0
robust	0	0
rust	1	0
thrust	0	1
trust	0	0
Absolute	**3.5**	**3.5**
Relative	**0.21**	**0.21**

random 1

Root	type of person
advocate	1
agnostic	1
alumnus	1
andante	0
arctic	0
beige	0
between	0
blot	0
chaplain	1
cigar	0
crank	1
efficacious	0
enema	0
equestrian	1
era	0
execute	0
fieldfare	0
flautist	1
fount	0
harrier	0
I	0.5
incur	0
inter	0
intern	1
item	0
jab	0
lawyer	1
live	0
malachite	0
mystery	0
nightmare	0
oryx	0
peer	1
perjure	0
prime	0

quinary	0
rue	0
saltation	0
scope	0
silage	0
slipper	0
standard	0
stripe	0
temporal	0
traipse	0
umbel	0
uncle	1
virus	0
wake	0
wodge	0
Absolute	**11.5**
Relative	**0.23**

random 2

Root	plant	resembling white fabric
allure	0	0
attire	0	0.5
bassinet	0	0.5
bent	0	0
chenille	0	1
chrism	0	0
entity	0	0
examine	0	0
fascine	0	0
fizz	0	0
gloxinia	1	0
hexad	0	0
ingeminate	0	0
lean	0	0
mademoiselle	0	0
moon	0	0.5
nopal	1	0
odometer	0	0
pall	0	1
pendent	0	0
platform	0	0
puff	0	0.5
rare	0	0
ravel	0	0
rod	0.5	0
round	0	0

sapient	0	0
shemozzle	0	0
sinecure	0	0
smooth	0	1
soft	0	1
souse	0	0
spell	0	0
squash	1	0
stagnant	0	0
steenbok	0	0
taraxacum	1	0
tint	0	0
tonneau	0	0
treasure	0	0
upsides	0	0
vacant	0	0
various	0	0
villein	0	0
vizor	0	0
warn	0	0
went	0	0
win	0	0
wormwood	1	0
yacht	0	0
Absolute	**5.5**	**6**
Relative	**0.11**	**0.12**

Appendix B. Forms of lemmas surveyed

ph+s	Forms	ph–s	Forms	s–ph	Forms
glow	glow, glowed, glowing, glows	glove	glove, gloves; gloved	burn	burn, burned, burning, burns, burnt
glare	glare, glared, glaring, glares	glue	glue, glued, gluing, glues, glueing	shine	shine, shined, shining, shines, shone
gleam	gleam, gleamed, gleaming, gleams	gland	gland, glands	scowl	scowl, scowled, scowling, scowls
glisten	glisten, glistened, glistening, glistens	glucose	glucose	luster	luster, lusters; lustre, lustres
glower	glower, glowered, glowering, glowers	glade	glade, glades; gladed	radiant/ radiate	radiant, radiate, radiating, radiates, radiated
snort	snort, snorted, snorting, snorts	snack	snack, snacked, snacking, snacks	inhale/ inhalation	inhale, inhaled, inhaling, inhales; inhalation, inhalations
snore	snore, snored, snoring, snores	snail	snail, snails	nasal/ nasality	nasal, nasals; nasality
sneeze	sneeze, sneezed, sneezing, sneezes	snare	snare, snared, snaring, snares	exhale/ exhalation	exhale, exhaled, exhaling, exhales; exhalation, exhalations
snout	snout, snouts	snipe	snipe, sniped, sniping, snipes	beak	beak, beaked, beaks
snivel	snivel, snivels, sniveled, sniveling (and 2-l versions)	snooker	snooker, snookered, snookering, snookers	cavil	cavil, caviled, cavils, cavilling, cavilled, cavilling
spread	spread, spreading, spreads, spreaded	splendid /splend or	splendid; splendor, splendors, splendour, splendours	expand/ expansion	expand, expanded, expanding, expands; expansion, expansions
spray	spray, sprayed, spraying, sprays	spruce	spruce, spruced, spruces, sprucing	wet	wet, wetting, wets, wetted, wetter, wettest
sprinkle	sprinkle, sprinkled, sprinkles, sprinkling	sprig	sprig, sprigs, sprigged, sprigging	scatter	scatter, scattered, scattering, scatters
splash	splash, splashed, splashes, splashing	spleen	spleen, spleens	dampen	dampen, dampens, dampened, dampening
splay	splay, splayed, splaying, splays	splanch nic	splanchnic	diverge	diverge, diverged, diverges; diverging
crash	crash, crashed, crashing, crashes	ash	ash, ashes; ashed	slap	slap, slapped, slapping, slaps
splash	splash, splashed, splashing, splashes	rash	rash, rashes	collide/ collision	collide, collided, colliding, collides; collision, collisions
slash	slash, slashed, slashing, slashes	stash	stash, stashed, stashing, stashes	immerse/ immersion	immerse, immersed, immersing, immerses; immersion, immersions
mash	mash, mashes, mashing, mashed	hash	hash, hashed, hashing, hashes	pulp	pulp, pulps, pulping, pulped
thrash	thrash, thrashed, thrashing, thrashes	sash	sash, sashes	sever	sever, severs, severed, severing
curl	curl, curled, curling, curls	pearl	pearl, pearled, pearling, pearls	curve	curve, curved, curving, curves
swirl	swirl, swirled, swirling, swirls	earl	earl, earls	spiral	spiral, spirals, spiraled, spiralled, spiraling, spiralling
whirl	whirl, whirled, whirling, whirls	squirrel	squirrel, squirrels	vortex	vortex, vortexing, vortexes, vortices, vortexed
twirl	twirl, twirled,	hurl	hurl, hurled, hurling,	gyre/	gyre, gyres; gyrate, gyrated,

	twirling, twirls		hurls	gyrate	gyrating, gyrates
dump	dump, dumps, dumped, dumping	jump	jump, jumped, jumping, jumps	cluster	cluster, clustered, clustering, clusters
slump	slump, slumped, slumping, slumps	pump	pump, pumped, pumping, pumps	knot	knot, knotted, knotting, knots
clump	clump, clumped, clumping, clumps	trump	trump, trumped, trumping, trumps	ditch	ditch, ditched, ditches, ditching
hump	hump, humped, humping, humps	chump	chump, chumps	subside	subside, subsided, subsides; subsiding
rump	rump, rumps, rumped	sump	sump, sumps	backside	backside, backsides

Appendix C. Corpus survey results

The results tables show the results for all items included in final results. Lines are left empty for lemmas that were excluded from final results. This is to maintain the parallel structure across the lists. The lemmas are matched by group – e.g., *gl–*, *sn–* – and not by individual item; for the sake of efficiency I have left them blank where they were in the original tables, since unequal numbers were excluded from each group. I have put lemmas in the $\mathbf{S_R}$ set in SMALL CAPITALS for visibility.

The **Raw results** tables show the actual numbers. The **Proportional results: all sets** tables show the number of items per 10,000 out of the sum of all included study lemmas in all three groups. The numbers are out of 10,000 so as to eliminate decimals (they are rounded to the nearest whole number). The **Proportional results: P+S** tables show the number of items per 10,000 out of the sum of just the phonaesthematic and semantically matched non-phonaesthematic groups (i.e., of those that have the semantic commonality with or without the phonemic commonality). A separate table has not been provided for results just from the subset that excludes polysyllabic Romance-derived words, as it would not provide substantial additional information, but a summary line is provided at the bottom of the third table showing proportions just of that subset, i.e., $\mathbf{P}/(\mathbf{P}+\mathbf{S_G})$. (The lines showing proportion of polysyllabic Romance lemmas and monosyllabic and Germanic lemmas are proportions of the entire set.) Note that the $\mathbf{P}/(\mathbf{P}+\mathbf{S_G})$ results are not the sum of the proportional results for **P** and the proportional results for $\mathbf{S_G}$; they are a recalculation of **P** as proportional to just $\mathbf{P}+\mathbf{S_G}$ rather than **P**+**S**.

Fiction

Raw results

P lemma	1800 UK	1800 US	1900 UK	1900 US	2000 UK	2000 US	C lemma	1800 UK	1800 US	1900 UK	1900 US	2000 UK	2000 US	S lemma	1800 UK	1800 US	1900 UK	1900 US	2000 UK	2000 US
glow	124	284	193	1489	927	1443														
glare	49	75	114	747	1145	1521	glue	1	8	9	81	125	328	shine	93	519	360	3195	1148	1433
gleam	110	199	201	1153	631	584								scowl	6	30	32	322	352	633

glisten	15	57	49	392	176	331								LUSTER	28	192	23	35	35	38				
							glade	33	31	17	163	70	61	RADIANT/ RADIATE	13	51	97	671	85	410				
snort	6	25	35	239	421	804								INHALE/ INHALATION	6	28	11	103	118	544				
snore	5	21	30	164	194	267	snail	2	5	23	34	59	90	NASAL/ NASALITY	2	8	9	34	38	91				
sneeze	1	3	19	60	72	167	snare	26	69	33	175	65	106	EXHALE/ EXHALATION	7	29	12	91	86	508				
snout	3	1	9	13	60	79	snipe	0	2	4	25	26	48	beak	6	21	20	79	114	136				
snivel	1	3	6	17	21	16								CAVIL	2	6	6	20	4	7				
														EXPAND/ EXPANSION	35	59	40	254	226	574				
spray	6	48	39	331	334	684	spruce	1	13	5	106	116	124	wet	44	231	228	1148	1347	2555				
sprinkle	14	43	24	218	101	241								scatter	54	196	111	882	452	738				
							spleen	10	33	4	37	17	30											
														DIVERGE	3	15	10	38	6	15				
crash	18	37	69	702	872	1247	ash	27	212	61	605	609	758	slap	3	23	43	348	609	1518				
splash	3	11	85	389	498	741								COLLIDE/ COLLISION	0	9	29	169	159	262				
slash	1	7	14	120	179	401								IMMERSE/ IMMERSION	4	32	12	62	90	125				
mash	1	2	7	39	59	102	hash	1	5	3	45	31	80	pulp	0	0	3	27	47	182				
thrash	4	13	16	126	155	162	sash	5	19	21	178	86	120	SEVER	4	99	16	161	78	132				
curl	38	121	155	1001	1100	2154								curve	4	19	123	574	690	1143				
														SPIRAL	4	4	11	88	180	307				
whirl	13	93	81	907	334	545								VORTEX	1	5	8	40	51	35				
twirl	6	8	14	78	103	339	hurl	26	90	29	512	293	285											
dump	2	4	17	146	456	1036								cluster	25	66	47	423	243	521				
slump	2	1	1	21	380	573	pump	4	19	25	371	351	935	knot	27	76	80	525	437	924				
clump	7	13	28	232	190	296								ditch	16	59	51	263	266	420				
hump	4	6	11	72	93	165								SUBSIDE	44	119	39	195	243	234				
rump	2	1	0	9	52	89								backside	2	2	0	0	138	182				
Grand totals	**435**	**1076**	**1217**	**8665**	**8553**	**13987**		**136**	**506**	**234**	**2332**	**1848**	**2965**		**433**	**1898**	**1421**	**9747**	**7242**	**13667**				
P+C+S	**1004**	**3480**	**2872**	**20744**	**17643**	**30619**																		

Proportional results: all sets

P lemma	1800 UK	1800 US	1900 UK	1900 US	2000 UK	2000 US	C lemma	1800 UK	1800 US	1900 UK	1900 US	2000 UK	2000 US	S lemma	1800 UK	1800 US	1900 UK	1900 US	2000 UK	2000 US
glow	1235	816	672	718	525	471														
glare	488	216	397	360	649	497	glue	10	23	31	39	71	107	shine	926	1491	1253	1540	651	468
gleam	1096	572	700	556	358	191								scowl	60	86	111	155	200	207
glisten	149	164	171	189	100	108								LUSTER	279	552	80	17	20	12
							glade	329	89	59	79	40	20	RADIANT/ RADIATE	129	147	338	323	48	134
snort	60	72	122	115	239	263								INHALE/ INHALATION	60	80	38	50	67	178
snore	50	60	104	79	110	87	snail	20	14	80	16	33	29	NASAL/ NASALITY	20	23	31	16	22	30
sneeze	10	9	66	29	41	55	snare	259	198	115	84	37	35	EXHALE/ EXHALATION	70	83	42	44	49	166
snout	30	3	31	6	34	26	snipe	0	6	14	12	15	16	beak	60	60	70	38	65	44
snivel	10	9	21	8	12	5								CAVIL	20	17	21	10	2	2
														EXPAND/ EXPANSION	349	170	139	122	128	187
spray	60	138	136	160	189	223	spruce	10	37	17	51	66	40	wet	438	664	794	553	763	834
sprinkle	139	124	84	105	57	79								scatter	538	563	386	425	256	241
							spleen	100	95	14	18	10	10							
														DIVERGE	30	43	35	18	3	5
crash	179	106	240	338	494	407	ash	269	609	212	292	345	248	slap	30	66	150	168	345	496
splash	30	32	296	188	282	242								COLLIDE/ COLLISION	0	26	101	81	90	86
slash	10	20	49	58	101	131								IMMERSE/ IMMERSION	40	92	42	30	51	41
mash	10	6	24	19	33	33	hash	10	14	10	22	18	26	pulp	0	0	10	13	27	59
thrash	40	37	56	61	88	53	sash	50	55	73	86	49	39	SEVER	40	284	56	78	44	43
curl	378	348	540	483	623	703								curve	40	55	428	277	391	373
														SPIRAL	40	11	38	42	102	100
whirl	129	267	282	437	189	178								VORTEX	10	14	28	19	29	11
twirl	60	23	49	38	58	111	hurl	259	259	101	247	166	93							

dump	20	11	59	70	258	338								cluster	249	190	164	204	138	170
slump	20	3	3	10	215	187	pump	40	55	87	179	199	305	knot	269	218	279	253	248	302
clump	70	37	97	112	108	97								ditch	159	170	178	127	151	137
hump	40	17	38	35	53	54								SUBSIDE	438	342	136	94	138	76
rump	20	3	0	4	29	29								backside	20	6	0	0	78	59
Grand totals	**4333**	**3092**	**4237**	**4177**	**4848**	**4568**		**1355**	**1454**	**815**	**1124**	**1047**	**968**		**4313**	**5454**	**4948**	**4699**	**4105**	**4464**

Proportional results: P+S

P lemma	1800 UK	1800 US	1900 UK	1900 US	2000 UK	2000 US	S lemma	1800 UK	1800 US	1900 UK	1900 US	2000 UK	2000 US
glow	1429	955	732	809	587	522							
glare	565	252	432	406	725	550	shine	1071	1745	1365	1735	727	518
gleam	1267	669	762	626	399	211	scowl	69	101	121	175	223	229
glisten	173	192	186	213	111	120	LUSTER	323	646	87	19	22	14
							RADIANT/RADIATE	150	171	368	364	54	148
snort	69	84	133	130	267	291	INHALE/INHALATION	69	94	42	56	75	197
snore	58	71	114	89	123	97	NASAL/NASALITY	23	27	34	18	24	33
sneeze	12	10	72	33	46	60	EXHALE/EXHALATION	81	98	45	49	54	184
snout	35	3	34	7	38	29	beak	69	71	76	43	72	49
snivel	12	10	23	9	13	6	CAVIL	23	20	23	11	3	3
							EXPAND/EXPANSION	403	198	152	138	143	208
spray	69	161	148	180	211	247	wet	507	777	864	624	853	924
sprinkle	161	145	91	118	64	87	scatter	622	659	421	479	286	267
							DIVERGE	35	50	38	21	4	5
crash	207	124	262	381	552	451	slap	35	77	163	189	386	549
splash	35	37	322	211	315	268	COLLIDE/COLLISION	0	30	110	92	101	95
slash	12	24	53	65	113	145	IMMERSE/IMMERSION	46	108	45	34	57	45
mash	12	7	27	21	37	37	pulp	0	0	11	15	30	66
thrash	46	44	61	68	98	59	SEVER	46	333	61	87	49	48

curl	438	407	588	544	696	779	curve	46	64	466	312	437	413
							SPIRAL	46	13	42	48	114	111
whirl	150	313	307	493	211	197	VORTEX	12	17	30	22	32	13
twirl	69	27	53	42	65	123							
dump	23	13	64	79	289	375	cluster	288	222	178	230	154	188
slump	23	3	4	11	241	207	knot	311	256	303	285	277	334
clump	81	44	106	126	120	107	ditch	184	198	193	143	168	152
hump	46	20	42	39	59	60	SUBSIDE	507	400	148	106	154	85
rump	23	3	0	5	33	32	backside	23	7	0	0	87	66
All	**5012**	**3618**	**4613**	**4706**	**5415**	**5058**		**4988**	**6382**	**5387**	**5294**	**4585**	**4942**
							S_R/(P+S)	1763	2206	1224	1065	886	1187
							S_G/(P+S)	3226	4176	4162	4229	3699	3755
P/(P+S_G)	**6084**	**4642**	**5257**	**5267**	**5941**	**5739**		**3916**	**5358**	**4743**	**4733**	**4059**	**4261**

Detective fiction: raw results

The detective fiction set includes results for the whole set and for the first 1000 words of each work, and a third column shows the ratio of the first-1000-words subset to the whole set. Proportional results for the first-1000-words subset are the proportion just of that subset.

P lemma	all	first 1000	ratio	C lemma	all	first 1000	ratio	S lemma	all	first 1000	ratio
glow	198	11	5.56%								
glare	183	6	3.28%	glue	18	1	5.56%	shine	308	10	3.25%
gleam	278	4	1.44%					scowl	28	0	0.00%
glisten	52	3	5.77%					LUSTER	35	7	20.00%
				glade	7	1	14.29%	RADIANT/RADIATE	88	3	3.41%
snort	34	1	2.94%					INHALE/INHALATION	17	0	0.00%
snore	19	1	5.26%	snail	8	0	0.00%	NASAL/NASALITY	9	0	0.00%
sneeze	13	0	0.00%	snare	60	1	1.67%	EXHALE/EXHALATION	18	1	5.56%

snout	0	0	0.00%	snipe	1	0	0.00%	beak	22	1	4.55%
snivel	5	0	0.00%					CAVIL	3	1	33.33%
								EXPAND/EXPANSION	21	0	0.00%
spray	17	0	0.00%	spruce	10	0	0.00%	wet	185	2	1.08%
sprinkle	32	2	6.25%					scatter	196	8	4.08%
				spleen	1	0	0.00%				
								DIVERGE	5	0	0.00%
crash	142	3	2.11%	ash	127	3	2.36%	slap	69	1	1.45%
splash	63	3	4.76%					COLLIDE/COLLISION	27	0	0.00%
slash	15	1	6.67%					IMMERSE/IMMERSION	23	1	4.35%
mash	1	0	0.00%	hash	3	0	0.00%	pulp	3	0	0.00%
thrash	14	0	0.00%	sash	46	0	0.00%	SEVER	34	0	0.00%
curl	117	2	1.71%					curve	156	4	2.56%
								SPIRAL	12	2	16.67%
whirl	131	3	2.29%					VORTEX	5	1	20.00%
twirl	21	0	0.00%	hurl	76	4	5.26%				
dump	18	2	11.11%					cluster	54	7	12.96%
slump	4	0	0.00%	pump	45	0	0.00%	knot	161	3	1.86%
clump	46	0	0.00%					ditch	60	3	5.00%
hump	9	2	22.22%					SUBSIDE	42	2	4.76%
rump	2	0	0.00%					backside	0	0	0.00%
Grand totals	**1414**	**44**	**3.11%**		**402**	**10**	**2.49%**		**1581**	**57**	**3.61%**
All types	**3397**	**111**	**3.27%**								

Detective fiction: proportional results: all sets

P lemma	all	first 1000	ratio	C lemma	all	first 1000	ratio	S lemma	all	first 1000	ratio
glow	583	991	170.02%								
glare	539	541	100.34%	glue	53	90	170.02%	shine	907	901	99.36%
gleam	818	360	44.03%					scowl	82	0	0.00%

glisten	153	270	176.56%				LUSTER	103	631	612.07%
				glade	21	90	437.19%	RADIANT/RADIATE	259	270	104.33%
snort	100	90	90.01%					INHALE/INHALATION	50	0	0.00%
snore	56	90	161.07%	snail	24	0	0.00%	NASAL/NASALITY	26	0	0.00%
sneeze	38	0	0.00%	snare	177	90	51.01%	EXHALE/EXHALATION	53	90	170.02%
snout	0	0	0.00%	snipe	3	0	0.00%	beak	65	90	139.11%
snivel	15	0	0.00%					CAVIL	9	90	1020.12%
								EXPAND/EXPANSION	62	0	0.00%
spray	50	0	0.00%	spruce	29	0	0.00%	wet	545	180	33.08%
sprinkle	94	180	191.27%					scatter	577	721	124.91%
				spleen	3	0	0.00%				
								DIVERGE	15	0	0.00%
crash	418	270	64.66%	ash	374	270	72.29%	slap	203	90	44.35%
splash	185	270	145.73%					COLLIDE/COLLISION	79	0	0.00%
slash	44	90	204.02%					IMMERSE/IMMERSION	68	90	133.06%
mash	3	0	0.00%	hash	9	0	0.00%	pulp	9	0	0.00%
thrash	41	0	0.00%	sash	135	0	0.00%	SEVER	100	0	0.00%
curl	344	180	52.31%					curve	459	360	78.47%
								SPIRAL	35	180	510.06%
whirl	386	270	70.08%					VORTEX	15	90	612.07%
twirl	62	0	0.00%	hurl	224	360	161.07%				
dump	53	180	340.04%					cluster	159	631	396.71%
slump	12	0	0.00%	pump	132	0	0.00%	knot	474	270	57.03%
clump	135	0	0.00%					ditch	177	270	153.02%
hump	26	180	680.08%					SUBSIDE	124	180	145.73%
rump	6	0	0.00%					backside	0	0	0.00%
Grand totals	**4162**	**3964**	**95.23%**		**1183**	**901**	**76.13%**		**4654**	**5135**	**110.34%**

Detective fiction: proportional results: P+S

P lemma	all	first 1000	ratio	S lemma	all	first 1000	ratio
glow	661	1089	164.74%				
glare	611	594	97.22%	shine	1028	990	96.28%
gleam	928	396	42.67%	scowl	93	0	0.00%
glisten	174	297	171.08%	LUSTER	117	693	593.07%
				RADIANT/RADIATE	294	297	101.09%
snort	114	99	87.22%	INHALE/INHALATION	57	0	0.00%
snore	63	99	156.07%	NASAL/NASALITY	30	0	0.00%
sneeze	43	0	0.00%	EXHALE/EXHALATION	60	99	164.74%
snout	0	0	0.00%	beak	73	99	134.79%
snivel	17	0	0.00%	CAVIL	10	99	988.45%
				EXPAND/EXPANSION	70	0	0.00%
spray	57	0	0.00%	wet	618	198	32.06%
sprinkle	107	198	185.33%	scatter	654	792	121.03%
				DIVERGE	17	0	0.00%
crash	474	297	62.65%	slap	230	99	42.98%
splash	210	297	141.21%	COLLIDE/COLLISION	90	0	0.00%
slash	50	99	197.69%	IMMERSE/IMMERSION	77	99	128.93%
mash	3	0	0.00%	pulp	10	0	0.00%
thrash	47	0	0.00%	SEVER	114	0	0.00%
curl	391	198	50.69%	curve	521	396	76.03%
				SPIRAL	40	198	494.22%
whirl	437	297	67.91%	VORTEX	17	99	593.07%
twirl	70	0	0.00%				
dump	60	198	329.48%	cluster	180	693	384.40%
slump	13	0	0.00%	knot	538	297	55.25%
clump	154	0	0.00%	ditch	200	297	148.27%
hump	30	198	658.97%	SUBSIDE	140	198	141.21%

rump	7	0	0.00%	backside	0	0	0.00%
All	**4721**	**4356**	**92.27%**		**5279**	**5644**	**106.91%**
				S_R/(P+S)	1132	1782	157.45%
				S_G/(P+S)	4147	3861	93.11%
P/(P+S_G)	**5324**	**5301**	**99.58%**		**4676**	**4699**	**100.48%**

Drama

Raw results

P lemma	1800 UK	2000 US	C lemma	1800 UK	2000 US	S lemma	1800 UK	2000 US
glow	23	640						
glare	9	650	glue	1	121	shine	31	501
gleam	18	91				scowl	1	155
glisten	3	100				LUSTER	18	4
			glade	2	17	RADIANT/RADIATE	13	76
snort	0	157				INHALE/INHALATION	4	122
snore	4	98	snail	3	25	NASAL/NASALITY	0	32
sneeze	2	56	snare	7	24	EXHALE/EXHALATION	4	170
snout	0	42	snipe	0	7	beak	31	44
snivel	2	3				CAVIL	1	0
						EXPAND/EXPANSION	5	104
spray	3	584	spruce	1	7	wet	8	743
sprinkle	3	61				scatter	13	390
			spleen	14	17			
						DIVERGE	0	3
crash	2	1394	ash	9	811	slap	3	897
splash	0	365				COLLIDE/COLLISION	0	202
slash	0	300				IMMERSE/IMMERSION	1	29
mash	0	43	hash	2	33	pulp	0	35

thrash	1	143	sash	1	37	SEVER	6	65
curl	10	329				curve	4	157
						SPIRAL	0	120
whirl	8	313				VORTEX	3	58
twirl	2	113	hurl	11	378			
dump	0	562				cluster	1	192
slump	0	397	pump	3	551	knot	13	192
clump	2	52				ditch	5	182
hump	2	79				SUBSIDE	2	72
rump	0	15				backside	0	37
Grand totals	**94**	**6587**		**54**	**2028**		**167**	**4582**
All types	**315**	**13197**						

Proportional results: all sets

P lemma	1800 UK	2000 US	C lemma	1800 UK	2000 US	S lemma	1800 UK	2000 US
glow	730	485						
glare	286	493	glue	32	92	shine	984	380
gleam	571	69				scowl	32	117
glisten	95	76				LUSTER	571	3
			glade	63	13	RADIANT/RADIATE	413	58
snort	0	119				INHALE/INHALATION	127	92
snore	127	74	snail	95	19	NASAL/NASALITY	0	24
sneeze	63	42	snare	222	18	EXHALE/EXHALATION	127	129
snout	0	32	snipe	0	5	beak	984	33
snivel	63	2				CAVIL	32	0
						EXPAND/EXPANSION	159	79
spray	95	443	spruce	32	5	wet	254	563
sprinkle	95	46				scatter	413	296
			spleen	444	13			

						DIVERGE	0	2
crash	63	1056	ash	286	615	slap	95	680
splash	0	277				COLLIDE/COLLISION	0	153
slash	0	227				IMMERSE/IMMERSION	32	22
mash	0	33	hash	63	25	pulp	0	27
thrash	32	108	sash	32	28	SEVER	190	49
curl	317	249				curve	127	119
						SPIRAL	0	91
whirl	254	237				VORTEX	95	44
twirl	63	86	hurl	349	286			
dump	0	426				cluster	32	145
slump	0	301	pump	95	418	knot	413	145
clump	63	39				ditch	159	138
hump	63	60				SUBSIDE	63	55
rump	0	11				backside	0	28
Grand totals	**2984**	**4991**		**1714**	**1537**		**5302**	**3472**

Proportional results: P+S

P lemma	1800 UK	2000 US	S lemma	1800 UK	2000 US
glow	881	573			
glare	345	582	shine	1188	449
gleam	690	81	scowl	38	139
glisten	115	90	LUSTER	690	4
			RADIANT/RADIATE	498	68
snort	0	141	INHALE/INHALATION	153	109
snore	153	88	NASAL/NASALITY	0	29
sneeze	77	50	EXHALE/EXHALATION	153	152
snout	0	38	beak	1188	39
snivel	77	3	CAVIL	38	0

			EXPAND/EXPANSION	192	93
spray	115	523	wet	307	665
sprinkle	115	55	scatter	498	349
			DIVERGE	0	3
crash	77	1248	slap	115	803
splash	0	327	COLLIDE/COLLISION	0	181
slash	0	269	IMMERSE/IMMERSION	38	26
mash	0	38	pulp	0	31
thrash	38	128	SEVER	230	58
curl	383	295	curve	153	141
			SPIRAL	0	107
whirl	307	280	VORTEX	115	52
twirl	77	101			
dump	0	503	cluster	38	172
slump	0	355	knot	498	172
clump	77	47	ditch	192	163
hump	77	71	SUBSIDE	77	64
rump	0	13	backside	0	33
All	**3602**	**5898**		**6398**	**4102**
			S_R/(P+S)	2184	946
			S_G/(P+S)	4215	3156
P/(P+S_G)	**4608**	**6514**		**5392**	**3486**

Magazines

Raw results

P lemma	1800 US	1900 US	2000 US	C lemma	1800 US	1900 US	2000 US	S lemma	1800 US	1900 US	2000 US
glow	42	160	527								

glare	11	87	197	glue	0	17	436	shine	45	409	629
gleam	12	100	52					scowl	1	9	41
glisten	5	40	72					LUSTER	27	39	65
				glade	2	21	33	RADIANT/RADIATE	3	136	260
snort	0	23	78					INHALE/INHALATION	7	6	175
snore	2	27	80	snail	2	21	61	NASAL/NASALITY	1	10	40
sneeze	0	8	63	snare	1	33	49	EXHALE/EXHALATION	10	23	120
snout	0	9	36	snipe	0	3	45	beak	3	38	58
snivel	0	1	2					CAVIL	2	20	2
								EXPAND/EXPANSION	68	530	1668
spray	9	44	686	spruce	6	67	122	wet	24	170	656
sprinkle	7	75	546					scatter	76	363	349
				spleen	6	4	30				
								DIVERGE	5	36	36
crash	6	113	937	ash	18	116	212	slap	1	27	235
splash	0	85	221					COLLIDE/COLLISION	21	126	317
slash	0	21	239					IMMERSE/IMMERSION	4	40	157
mash	0	14	83	hash	2	4	61	pulp	1	14	110
thrash	0	15	63	sash	1	29	22	SEVER	19	67	97
curl	10	82	360					curve	14	150	448
								SPIRAL	3	33	300
whirl	14	204	126					VORTEX	9	33	27
twirl	1	6	55	hurl	6	86	103				
dump	0	44	512					cluster	15	118	730
slump	0	8	189	pump	4	112	757	knot	7	174	264
clump	0	30	151					ditch	3	145	175
hump	0	13	68					SUBSIDE	18	61	106
rump	1	8	21					backside	1	0	36
Grand totals	**120**	**1217**	**5364**		**48**	**513**	**1931**		**388**	**2777**	**7101**
All types	**556**	**4507**	**14396**								

Proportional results: all sets

P lemma	1800 US	1900 US	2000 US	C lemma	1800 US	1900 US	2000 US	S lemma	1800 US	1900 US	2000 US
glow	755	355	366								
glare	198	193	137	glue	0	38	303	shine	809	907	437
gleam	216	222	36					scowl	18	20	28
glisten	90	89	50					LUSTER	486	87	45
				glade	36	47	23	RADIANT/RADIATE	54	302	181
snort	0	51	54					INHALE/INHALATION	126	13	122
snore	36	60	56	snail	36	47	42	NASAL/NASALITY	18	22	28
sneeze	0	18	44	snare	18	73	34	EXHALE/EXHALATION	180	51	83
snout	0	20	25	snipe	0	7	31	beak	54	84	40
snivel	0	2	1					CAVIL	36	44	1
								EXPAND/EXPANSION	1223	1176	1159
spray	162	98	477	spruce	108	149	85	wet	432	377	456
sprinkle	126	166	379					scatter	1367	805	242
				spleen	108	9	21				
								diverge	90	80	25
crash	108	251	651	ash	324	257	147	slap	18	60	163
splash	0	189	154					COLLIDE/COLLISION	378	280	220
slash	0	47	166					IMMERSE/IMMERSION	72	89	109
mash	0	31	58	hash	36	9	42	pulp	18	31	76
thrash	0	33	44	sash	18	64	15	SEVER	342	149	67
curl	180	182	250					curve	252	333	311
								SPIRAL	54	73	208
whirl	252	453	88					VORTEX	162	73	19
twirl	18	13	38	hurl	108	191	72				
dump	0	98	356					cluster	270	262	507
slump	0	18	131	pump	72	249	526	knot	126	386	183
clump	0	67	105					ditch	54	322	122
hump	0	29	47					SUBSIDE	324	135	74

rump	18	18	15				backside	18	0	25
Grand totals	**2158**	**2700**	**3726**		**863**	**1138**	**1341**	**6978**	**6162**	**4933**

Proportional results: P+S

P lemma	1800 US	1900 US	2000 US	S lemma	1800 US	1900 US	2000 US
glow	827	401	423				
glare	217	218	158	shine	886	1024	505
gleam	236	250	42	scowl	20	23	33
glisten	98	100	58	LUSTER	531	98	52
				RADIANT/RADIATE	59	341	209
snort	0	58	63	INHALE/INHALATION	138	15	140
snore	39	68	64	NASAL/NASALITY	20	25	32
sneeze	0	20	51	EXHALE/EXHALATION	197	58	96
snout	0	23	29	beak	59	95	47
snivel	0	3	2	CAVIL	39	50	2
				EXPAND/EXPANSION	1339	1327	1338
spray	177	110	550	wet	472	426	526
sprinkle	138	188	438	scatter	1496	909	280
				DIVERGE	98	90	29
crash	118	283	752	slap	20	68	189
splash	0	213	177	COLLIDE/COLLISION	413	315	254
slash	0	53	192	IMMERSE/IMMERSION	79	100	126
mash	0	35	67	pulp	20	35	88
thrash	0	38	51	SEVER	374	168	78
curl	197	205	289	curve	276	376	359
				SPIRAL	59	83	241
whirl	276	511	101	VORTEX	177	83	22
twirl	20	15	44				

dump	0	110	411	cluster	295	295	586
slump	0	20	152	knot	138	436	212
clump	0	75	121	ditch	59	363	140
hump	0	33	55	SUBSIDE	354	153	85
rump	20	20	17	backside	20	0	29
All	**2362**	**3047**	**4303**		**7638**	**6953**	**5697**
				S_R/(P+S)	3878	2904	2704
				S_G/(P+S)	3760	4049	2993
P/(P+S_G)	**3859**	**4294**	**5898**		**6141**	**5706**	**4102**

Newspapers

Raw results

P lemma	1800 UK	1900 US	2000 UK	2000 US	C lemma	1800 UK	1900 US	2000 UK	2000 US	S lemma	1800 UK	1900 US	2000 UK	2000 US
glow	1	8	58	951										
glare	3	13	41	430	glue	0	5	85	722	shine	35	25	191	1491
gleam	1	12	12	166						scowl	0	2	8	143
glisten	0	2	6	143						LUSTER	13	2	25	210
					glade	0	0	5	138	RADIANT/RADIATE	0	11	32	395
snort	0	6	8	153						INHALE/INHALATION	0	5	53	401
snore	0	9	13	85	snail	0	6	44	252	NASAL/NASALITY	0	2	12	138
sneeze	0	2	23	139	snare	4	6	23	248	EXHALE/EXHALATION	0	1	7	167
snout	0	1	4	66	snipe	0	5	38	180	beak	0	3	18	105
snivel	0	0	1	7						CAVIL	2	3	0	14
										EXPAND/EXPANSION	4	82	650	11091
spray	0	7	135	1836	spruce	3	14	38	337	wet	25	62	283	1544
sprinkle	3	12	41	1748						scatter	9	69	97	1293
					spleen	12	0	12	76					
										DIVERGE	0	2	3	92

crash	0	139	1668	5322	ash	24	51	243	1061	slap	3	13	114	1240
splash	2	15	175	990						COLLIDE/COLLISION	0	183	351	1372
slash	0	17	284	1364						IMMERSE/IMMERSION	2	7	27	643
mash	0	0	29	409	hash	1	3	16	337	pulp	1	13	18	507
thrash	0	8	110	232	sash	6	5	11	88	SEVER	1	25	63	368
curl	6	14	73	658						curve	1	52	57	1189
										SPIRAL	0	2	73	846
whirl	0	20	16	211						VORTEX	0	0	4	106
twirl	1	1	11	178	hurl	1	61	154	455					
dump	0	27	410	2659						cluster	4	12	58	1892
slump	0	18	471	1482	pump	16	66	259	3174	knot	11	85	78	702
clump	0	6	24	260						ditch	24	23	140	730
hump	0	0	62	203						SUBSIDE	5	25	26	360
rump	3	1	19	83						backside	4	0	34	147
Grand totals	**20**	**338**	**3694**	**19775**		**67**	**222**	**928**	**7068**		**144**	**709**	**2422**	**27186**
All types	**231**	**1269**	**7044**	**54029**										

Proportional results: all sets

P lemma	1800 UK	1900 US	2000 UK	2000 US	C lemma	1800 UK	1900 US	2000 UK	2000 US	S lemma	1800 UK	1900 US	2000 UK	2000 US
glow	43	63	82	176										
glare	130	102	58	80	glue	0	39	121	134	shine	1515	197	271	276
gleam	43	95	17	31						scowl	0	16	11	26
glisten	0	16	9	26						LUSTER	563	16	35	39
					glade	0	0	7	26	RADIANT/RADIATE	0	87	45	73
snort	0	47	11	28						INHALE/INHALATION	0	39	75	74
snore	0	71	18	16	snail	0	47	62	47	NASAL/NASALITY	0	16	17	26
sneeze	0	16	33	26	snare	173	47	33	46	EXHALE/EXHALATION	0	8	10	31
snout	0	8	6	12	snipe	0	39	54	33	beak	0	24	26	19
snivel	0	0	1	1						CAVIL	87	24	0	3

									EXPAND/EXPANSION	173	646	923	2053
spray	0	55	192	340	spruce	130	110	54	62	wet	1082	489	402	286
sprinkle	130	95	58	324						scatter	390	544	138	239
					spleen	519	0	17	14					
										DIVERGE	0	16	4	17
crash	0	1095	2368	985	ash	1039	402	345	196	slap	130	102	162	230
splash	87	118	248	183						COLLIDE/COLLISION	0	1442	498	254
slash	0	134	403	252						IMMERSE/IMMERSION	87	55	38	119
mash	0	0	41	76	hash	43	24	23	62	pulp	43	102	26	94
thrash	0	63	156	43	sash	260	39	16	16	SEVER	43	197	89	68
curl	260	110	104	122						curve	43	410	81	220
										SPIRAL	0	16	104	157
whirl	0	158	23	39						VORTEX	0	0	6	20
twirl	43	8	16	33	hurl	43	481	219	84					
dump	0	213	582	492						cluster	173	95	82	350
slump	0	142	669	274	pump	693	520	368	587	knot	476	670	111	130
clump	0	47	34	48						ditch	1039	181	199	135
hump	0	0	88	38						SUBSIDE	216	197	37	67
rump	130	8	27	15						backside	173	0	48	27
Grand totals	**866**	**2664**	**5244**	**3660**		**2900**	**1749**	**1317**	**1308**		**6234**	**5587**	**3438**	**5032**

Proportional results: P+S

P lemma	1800 UK	1900 US	2000 UK	2000 US	S lemma	1800 UK	1900 US	2000 UK	2000 US
glow	61	76	95	203					
glare	183	124	67	92	shine	2134	239	312	317
gleam	61	115	20	35	scowl	0	19	13	30
glisten	0	19	10	30	LUSTER	793	19	41	45
					RADIANT/RADIATE	0	105	52	84
snort	0	57	13	33	INHALE/INHALATION	0	48	87	85

snore	0	86	21	18	NASAL/NASALITY	0	19	20	29
sneeze	0	19	38	30	EXHALE/EXHALATION	0	10	11	36
snout	0	10	7	14	beak	0	29	29	22
snivel	0	0	2	1	CAVIL	122	29	0	3
					EXPAND/EXPANSION	244	783	1063	2362
spray	0	67	221	391	wet	1524	592	463	329
sprinkle	183	115	67	372	scatter	549	659	159	275
					DIVERGE	0	19	5	20
crash	0	1328	2727	1133	slap	183	124	186	264
splash	122	143	286	211	COLLIDE/COLLISION	0	1748	574	292
slash	0	162	464	290	IMMERSE/IMMERSION	122	67	44	137
mash	0	0	47	87	pulp	61	124	29	108
thrash	0	76	180	49	SEVER	61	239	103	78
curl	366	134	119	140	curve	61	497	93	253
					SPIRAL	0	19	119	180
whirl	0	191	26	45	VORTEX	0	0	7	23
twirl	61	10	18	38					
dump	0	258	670	566	cluster	244	115	95	403
slump	0	172	770	316	knot	671	812	128	149
clump	0	57	39	55	ditch	1463	220	229	155
hump	0	0	101	43	SUBSIDE	305	239	43	77
rump	183	10	31	18	backside	244	0	56	31
All	**1220**	**3228**	**6040**	**4211**		**8780**	**6772**	**3960**	**5789**
					S_R/(P+S)	1646	3343	2168	3450
					S_G/(P+S)	7134	3429	1792	2339
P/(P+S_G)	**1460**	**4849**	**7712**	**6429**		**8540**	**5151**	**2288**	**3571**

Tabloid comparison: raw results

P lemma	US all	UK all	UK tabloid	UK all-tabloid	C lemma	US all	UK all	UK tabloid	UK all-tabloid	S lemma	US all	UK all	UK tabloid	UK all-tabloid
glow	951	58	7	51	glove					burn				
glare	430	41	3	38	glue	722	85	3	82	shine	1491	191	22	169
gleam	166	12	1	11	gland					scowl	143	8	0	8
glisten	143	6	0	6	glucose					LUSTER	210	25	0	25
glower					glade	138	5	0	5	RADIANT/RADIATE	395	32	3	29
snort	153	8	0	8	snack					INHALE/INHALATION	401	53	2	51
snore	85	13	5	8	snail	252	44	2	42	NASAL/NASALITY	138	12	2	10
sneeze	139	23	2	21	snare	248	23	1	22	EXHALE/EXHALATION	167	7	0	7
snout	66	4	0	4	snipe	180	38	3	35	beak	105	18	2	16
snivel	7	1	0	1	snooker					CAVIL	14	0	0	0
spread					splendid/splendor					EXPAND/ EXPANSION	11091	650	7	643
spray	1836	135	20	115	spruce	337	38	2	36	wet	1544	283	23	260
sprinkle	1748	41	3	38	sprig					scatter	1293	97	5	92
splash					spleen	76	12	0	12	dampen				
splay					splanchnic					DIVERGE	92	3	0	3
crash	5322	1668	178	1490	ash	1061	243	24	219	slap	1240	114	13	101
splash	990	175	43	132	rash					COLLIDE/COLLISION	1372	351	16	335
slash	1364	284	44	240	stash					IMMERSE/ IMMERSION	643	27	2	25
mash	409	29	0	29	hash	337	16	1	15	pulp	507	18	1	17
thrash	232	110	11	99	sash	88	11	1	10	SEVER	368	63	3	60
curl	658	73	7	66	pearl					curve	1189	57	1	56
swirl					earl					SPIRAL	846	73	1	72
whirl	211	16	1	15	squirrel					VORTEX	106	4	0	4
twirl	178	11	3	8	hurl	455	154	18	136	gyre/gyrate				
dump	2659	410	34	376	jump					cluster	1892	58	7	51
slump	1482	471	85	386	pump	3174	259	9	250	knot	702	78	3	75

clump	260	24	5	19	trump					ditch	730	140	9	131
hump	203	62	1	61	chump					SUBSIDE	360	26	0	26
rump	83	19	0	19	sump					backside	147	34	9	25
Grand totals	**19775**	**3694**	**453**	**3241**		**7068**	**928**	**64**	**864**		**27186**	**2422**	**131**	**2291**
All types	**54029**	**7044**	**648**	**6396**										

Tabloid comparison: proportional results: all sets

P lemma	US all	UK all	UK tabloid	UK all-tabloid	C lemma	US all	UK all	UK tabloid	UK all-tabloid	S lemma	US all	UK all	UK tabloid	UK all-tabloid
glow	176	82	108	80	glove					burn				
glare	80	58	46	59	glue	134	121	46	128	shine	276	271	340	264
gleam	31	17	15	17	gland					scowl	26	11	0	13
glisten	26	9	0	9	glucose					LUSTER	39	35	0	39
glower					glade	26	7	0	8	RADIANT/RADIATE	73	45	46	45
snort	28	11	0	13	snack					INHALE/ INHALATION	74	75	31	80
snore	16	18	77	13	snail	47	62	31	66	NASAL/NASALITY	26	17	31	16
sneeze	26	33	31	33	snare	46	33	15	34	EXHALE/ EXHALATION	31	10	0	11
snout	12	6	0	6	snipe	33	54	46	55	beak	19	26	31	25
snivel	1	1	0	2	snooker					CAVIL	3	0	0	0
spread					splendid/ splendor					EXPAND/ EXPANSION	2053	923	108	1005
spray	340	192	309	180	spruce	62	54	31	56	wet	286	402	355	407
sprinkle	324	58	46	59	sprig					scatter	239	138	77	144
splash					spleen	14	17	0	19	dampen				
splay					splanchnic					DIVERGE	17	4	0	5
crash	985	2368	2747	2330	ash	196	345	370	342	slap	230	162	201	158
splash	183	248	664	206	rash					COLLIDE/COLLISION	254	498	247	524
slash	252	403	679	375	stash					IMMERSE/	119	38	31	39

									IMMERSION					
mash	76	41	0	45	hash	62	23	15	23	pulp	94	26	15	27
thrash	43	156	170	155	sash	16	16	15	16	SEVER	68	89	46	94
curl	122	104	108	103	pearl					curve	220	81	15	88
swirl					earl					SPIRAL	157	104	15	113
whirl	39	23	15	23	squirrel					VORTEX	20	6	0	6
twirl	33	16	46	13	hurl	84	219	278	213	gyre/gyrate				
dump	492	582	525	588	jump					cluster	350	82	108	80
slump	274	669	1312	604	pump	587	368	139	391	knot	130	111	46	117
clump	48	34	77	30	trump					ditch	135	199	139	205
hump	38	88	15	95	chump					SUBSIDE	67	37	0	41
rump	15	27	0	30	sump					backside	27	48	139	39
Grand totals	**3660**	**5244**	**6991**	**5067**		**1308**	**1317**	**988**	**1351**		**5032**	**3438**	**2022**	**3582**

Tabloid comparison: proportional results: P+S

P lemma	US all	UK all	UK tabloid	UK all–tabloid	S lemma	US all	UK all	UK tabloid	UK all–tabloid
glow	203	95	120	92	burn				
glare	92	67	51	69	shine	317	312	377	305
gleam	35	20	17	20	scowl	30	13	0	14
glisten	30	10	0	11	LUSTER	45	41	0	45
glower					RADIANT/RADIATE	84	52	51	52
snort	33	13	0	14	INHALE/INHALATION	85	87	34	92
snore	18	21	86	14	NASAL/NASALITY	29	20	34	18
sneeze	30	38	34	38	EXHALE/EXHALATION	36	11	0	13
snout	14	7	0	7	beak	22	29	34	29
snivel	1	2	0	2	CAVIL	3	0	0	0
spread					EXPAND/EXPANSION	2362	1063	120	1162
spray	391	221	342	208	wet	329	463	394	470
sprinkle	372	67	51	69	scatter	275	159	86	166

splash					dampen				
splay					DIVERGE	20	5	0	5
crash	1133	2727	3048	2693	slap	264	186	223	183
splash	211	286	736	239	COLLIDE/COLLISION	292	574	274	606
slash	290	464	753	434	IMMERSE/IMMERSION	137	44	34	45
mash	87	47	0	52	pulp	108	29	17	31
thrash	49	180	188	179	SEVER	78	103	51	108
curl	140	119	120	119	curve	253	93	17	101
swirl					SPIRAL	180	119	17	130
whirl	45	26	17	27	VORTEX	23	7	0	7
twirl	38	18	51	14	gyre/gyrate				
dump	566	670	582	680	cluster	403	95	120	92
slump	316	770	1455	698	knot	149	128	51	136
clump	55	39	86	34	ditch	155	229	154	237
hump	43	101	17	110	SUBSIDE	77	43	0	47
rump	18	31	0	34	backside	31	56	154	45
All	**4211**	**6040**	**7757**	**5859**		**5789**	**3960**	**2243**	**4141**
					S_R/(P+S)	3450	2168	616	2332
					S_G/(P+S)	2339	1792	1627	1809
P/(P+S_G)	**6429**	**7712**	**8266**	**7640**		**3571**	**2288**	**1734**	**2360**

Academic articles in the humanities

Raw results

P lemma	1900 UK	2000 UK	2000 US	C lemma	1900 UK	2000 UK	2000 US	S lemma	1900 UK	2000 UK	2000 US
glow	17	10	118								
glare	12	7	55	glue	1	12	136	shine	24	20	257
gleam	3	8	40					scowl	0	0	12
glisten	0	0	22					LUSTER	0	4	55

				glade	1	1	7	RADIANT/RADIATE	4	23	170
snort	0	0	7					INHALE/INHALATION	1	3	82
snore	0	1	5	snail	3	0	46	NASAL/NASALITY	0	4	110
sneeze	0	1	15	snare	7	4	62	EXHALE/EXHALATION	0	2	47
snout	0	4	12	snipe	1	2	13	beak	0	14	49
snivel	0	0	2					CAVIL	0	2	4
								EXPAND/EXPANSION	59	386	4805
spray	2	6	198	spruce	0	3	24	wet	8	64	265
sprinkle	1	3	74					scatter	19	53	311
				spleen	0	5	8				
								DIVERGE	3	20	220
crash	6	21	355	ash	3	33	309	slap	2	8	131
splash	0	7	64					COLLIDE/COLLISION	12	33	281
slash	3	3	144					IMMERSE/IMMERSION	0	20	533
mash	0	0	6	hash	1	0	21	pulp	0	1	79
thrash	1	2	10	sash	0	1	37	SEVER	1	22	236
curl	1	7	59					curve	6	120	457
								SPIRAL	0	44	392
whirl	2	4	36					VORTEX	4	3	31
twirl	1	0	26	hurl	1	6	83				
dump	3	19	221					cluster	2	65	773
slump	0	21	76	pump	1	12	185	knot	2	40	136
clump	0	2	32					ditch	1	72	79
hump	0	7	19					SUBSIDE	3	11	86
rump	0	4	19					backside	0	2	12
Grand totals	**52**	**137**	**1615**		**19**	**79**	**931**		**151**	**1036**	**9613**
All types	**222**	**1252**	**12159**								

Proportional results: all sets

P lemma	1900 UK	2000 UK	2000 US	C lemma	1900 UK	2000 UK	2000 US	S lemma	1900 UK	2000 UK	2000 US
glow	766	80	97								
glare	541	56	45	glue	45	96	112	shine	1081	160	211
gleam	135	64	33					scowl	0	0	10
glisten	0	0	18					LUSTER	0	32	45
				glade	45	8	6	RADIANT/RADIATE	180	184	140
snort	0	0	6					INHALE/INHALATION	45	24	67
snore	0	8	4	snail	135	0	38	NASAL/NASALITY	0	32	90
sneeze	0	8	12	snare	315	32	51	EXHALE/EXHALATION	0	16	39
snout	0	32	10	snipe	45	16	11	beak	0	112	40
snivel	0	0	2					CAVIL	0	16	3
								EXPAND/EXPANSION	2658	3083	3952
spray	90	48	163	spruce	0	24	20	wet	360	511	218
sprinkle	45	24	61					scatter	856	423	256
				spleen	0	40	7				
								DIVERGE	135	160	181
crash	270	168	292	ash	135	264	254	slap	90	64	108
splash	0	56	53					COLLIDE/COLLISION	541	264	231
slash	135	24	118					IMMERSE/IMMERSION	0	160	438
mash	0	0	5	hash	45	0	17	pulp	0	8	65
thrash	45	16	8	sash	0	8	30	SEVER	45	176	194
curl	45	56	49					curve	270	958	376
								SPIRAL	0	351	322
whirl	90	32	30					VORTEX	180	24	25
twirl	45	0	21	hurl	45	48	68				
dump	135	152	182					cluster	90	519	636
slump	0	168	63	pump	45	96	152	knot	90	319	112
clump	0	16	26					ditch	45	575	65
hump	0	56	16					SUBSIDE	135	88	71

rump	0	32	16				backside	0	16	10
Grand totals	**2342**	**1094**	**1328**		**856**	**631**	**766**	**6802**	**8275**	**7906**

Proportional results: P+S

P lemma	1900 UK	2000 UK	2000 US	S lemma	1900 UK	2000 UK	2000 US
glow	837	85	105				
glare	591	60	49	shine	1182	171	229
gleam	148	68	36	scowl	0	0	11
glisten	0	0	20	LUSTER	0	34	49
				RADIANT/RADIATE	197	196	151
snort	0	0	6	INHALE/INHALATION	49	26	73
snore	0	9	4	NASAL/NASALITY	0	34	98
sneeze	0	9	13	EXHALE/EXHALATION	0	17	42
snout	0	34	11	beak	0	119	44
snivel	0	0	2	CAVIL	0	17	4
				EXPAND/EXPANSION	2906	3291	4279
spray	99	51	176	wet	394	546	236
sprinkle	49	26	66	scatter	936	452	277
				DIVERGE	148	171	196
crash	296	179	316	slap	99	68	117
splash	0	60	57	COLLIDE/COLLISION	591	281	250
slash	148	26	128	IMMERSE/IMMERSION	0	171	475
mash	0	0	5	pulp	0	9	70
thrash	49	17	9	SEVER	49	188	210
curl	49	60	53	curve	296	1023	407
				SPIRAL	0	375	349
whirl	99	34	32	VORTEX	197	26	28
twirl	49	0	23				

dump	148	162	197	cluster	99	554	688
slump	0	179	68	knot	99	341	121
clump	0	17	29	ditch	49	614	70
hump	0	60	17	SUBSIDE	148	94	77
rump	0	34	17	backside	0	17	11
All	**2562**	**1168**	**1438**		**7438**	**8832**	**8562**
				S_R/(P+S)	4286	4919	6281
				S_G/(P+S)	3153	3913	2281
P/(P+S_G)	**4483**	**2299**	**3867**		**5517**	**7701**	**6133**

Hansard

Raw results

P lemma	1800 UK	1900 UK	2000 UK	C lemma	1800 UK	1900 UK	2000 UK	S lemma	1800 UK	1900 UK	2000 UK
glow	11	31	91								
glare	10	45	147	glue	0	8	163	shine	13	60	353
gleam	0	37	63					scowl	0	5	13
glisten	1	0	2					LUSTER	42	21	27
				glade	0	0	10	RADIANT/RADIATE	0	18	44
snort	1	3	21					INHALE/INHALATION	2	7	144
snore	1	15	9	snail	3	25	62	NASAL/NASALITY	0	1	20
sneeze	0	4	41	snare	19	87	147	EXHALE/EXHALATION	5	6	6
snout	0	0	61	snipe	1	17	175	beak	0	10	15
snivel	0	2	6					CAVIL	26	85	100
								EXPAND/EXPANSION	20	1107	12241
spray	0	54	370	spruce	0	15	50	wet	12	487	610
sprinkle	0	5	29					scatter	47	468	432
				spleen	13	7	35				
								DIVERGE	6	43	200

crash	3	16	1444	ash	12	280	433	slap	1	25	391
splash	2	11	79					COLLIDE/COLLISION	14	546	511
slash	1	6	715					IMMERSE/IMMERSION	5	43	135
mash	8	27	14	hash	4	9	49	pulp	0	28	112
thrash	0	185	219	sash	2	13	11	SEVER	22	146	200
curl	0	5	30					curve	0	43	390
								SPIRAL	1	6	588
whirl	1	11	19					VORTEX	7	19	26
twirl	0	0	10	hurl	12	125	124				
dump	0	347	1970					cluster	3	18	546
slump	0	23	666	pump	4	160	1113	knot	7	624	189
clump	0	5	14					ditch	6	112	469
hump	1	3	240					SUBSIDE	61	44	105
rump	0	7	219					backside	0	0	139
Grand totals	**40**	**842**	**6479**		**70**	**746**	**2372**		**300**	**3972**	**18006**
All types	**410**	**5560**	**26857**								

Proportional results: all sets

P lemma	1800 UK	1900 UK	2000 UK	C lemma	1800 UK	1900 UK	2000 UK	S lemma	1800 UK	1900 UK	2000 UK
glow	268	56	34								
glare	244	81	55	glue	0	14	61	shine	317	108	131
gleam	0	67	23					scowl	0	9	5
glisten	24	0	1					LUSTER	1024	38	10
				glade	0	0	4	RADIANT/RADIATE	0	32	16
snort	24	5	8					INHALE/INHALATION	49	13	54
snore	24	27	3	snail	73	45	23	NASAL/NASALITY	0	2	7
sneeze	0	7	15	snare	463	156	55	EXHALE/EXHALATION	122	11	2
snout	0	0	23	snipe	24	31	65	beak	0	18	6
snivel	0	4	2					CAVIL	634	153	37

							EXPAND/EXPANSION	488	1991	4558
spray	0	97	138	spruce	0	27	19	wet	293	876	227
sprinkle	0	9	11					scatter	1146	842	161
				spleen	317	13	13				
								DIVERGE	146	77	74
crash	73	29	538	ash	293	504	161	slap	24	45	146
splash	49	20	29					COLLIDE/COLLISION	341	982	190
slash	24	11	266					IMMERSE/IMMERSION	122	77	50
mash	195	49	5	hash	98	16	18	pulp	0	50	42
thrash	0	333	82	sash	49	23	4	SEVER	537	263	74
curl	0	9	11					curve	0	77	145
								SPIRAL	24	11	219
whirl	24	20	7					VORTEX	171	34	10
twirl	0	0	4	hurl	293	225	46				
dump	0	624	734					cluster	73	32	203
slump	0	41	248	pump	98	288	414	knot	171	1122	70
clump	0	9	5					ditch	146	201	175
hump	24	5	89					SUBSIDE	1488	79	39
rump	0	13	82					backside	0	0	52
Grand totals	**976**	**1514**	**2412**		**1707**	**1342**	**883**		**7317**	**7144**	**6704**

Proportional results: P+S

P lemma	1800 UK	1900 UK	2000 UK	S lemma	1800 UK	1900 UK	2000 UK
glow	324	64	37				
glare	294	93	60	shine	382	125	144
gleam	0	77	26	scowl	0	10	5
glisten	29	0	1	LUSTER	1235	44	11
				RADIANT/RADIATE	0	37	18
snort	29	6	9	INHALE/INHALATION	59	15	59

snore	29	31	4	NASAL/NASALITY	0	2	8
sneeze	0	8	17	EXHALE/EXHALATION	147	12	2
snout	0	0	25	beak	0	21	6
snivel	0	4	2	CAVIL	765	177	41
				EXPAND/EXPANSION	588	2300	4999
spray	0	112	151	wet	353	1012	249
sprinkle	0	10	12	scatter	1382	972	176
				DIVERGE	176	89	82
crash	88	33	590	slap	29	52	160
splash	59	23	32	COLLIDE/COLLISION	412	1134	209
slash	29	12	292	IMMERSE/IMMERSION	147	89	55
mash	235	56	6	pulp	0	58	46
thrash	0	384	89	SEVER	647	303	82
curl	0	10	12	curve	0	89	159
				SPIRAL	29	12	240
whirl	29	23	8	VORTEX	206	39	11
twirl	0	0	4				
dump	0	721	805	cluster	88	37	223
slump	0	48	272	knot	206	1296	77
clump	0	10	6	ditch	176	233	192
hump	29	6	98	SUBSIDE	1794	91	43
rump	0	15	89	backside	0	0	57
All	**1176**	**1749**	**2646**		**8824**	**8251**	**7354**
				$S_R/(P+S)$	6206	4346	5860
				$S_G/(P+S)$	2618	3905	1494
$P/(P+S_G)$	**3101**	**3093**	**6391**		**6899**	**6907**	**3609**