That vexed problem of choice

Sean Wallis
Survey of English Usage, UCL
s.wallis@ucl.ac.uk

Abstract

A key challenge in corpus linguistics concerns the difficulty of operationalising linguistic questions in terms of *choices* made by speakers or writers. Whereas lab researchers design an experiment around a choice, comparable corpus research implies the inference of counterfactual alternates. This non-trivial requirement leads many to rely on a per million word baseline, meaning that variation separately due to *opportunity* and *choice* cannot be distinguished.

We formalise definitions of *mutual substitution* and the *true rate of alternation* as useful idealisations, recognising they may not always hold. Analysing data from a new volume on the verb phrase,[1] we demonstrate how a focus on choices available to speakers allows researchers to factor out the effect of changing opportunities to draw conclusions about choices.

We discuss research strategies where alternates may not be easily identified, including refining baselines by eliminating forms and surveying change against multiple baselines. Finally we address three objections that have been made to this framework, that alternates are not reliably identifiable, baselines are arbitrary, and differing ecological pressures apply to different terms. Throughout we motivate our responses by evidence from current research, demonstrating that whereas the problem of identifying choices may be 'vexed', it represents a highly fruitful paradigm for corpus linguistics.

Keywords: experimental design, research methodology, alternation, variationism, baseline, choice

1. Introduction

Many of the research questions we typically wish a corpus to answer can be formulated in terms of variables representing a *linguistic choice* made by speakers or writers, sometimes called *onomasiology*. A complementary approach (*semiasology*) examines the range of meanings associated with a particular expression. The essential idea is simple: in forming utterances, speakers and writers make a series of conscious and unconscious choices.

Numerous studies have examined how frequencies of words, lexical sequences and grammatical constructions vary under the pressure of changing external sociolinguistic conditions. Papers have been published on differences between speech and writing, the impact of change over time, and so forth. The point of this paper is to reemphasise that at the heart of these studies, correctly conceived, is necessarily *a model of choice*.

This statement has become axiomatic in sociolinguistics (Labov 1972; Lavandera 1978) and cognitive linguistics research, but in this paper we wish to emphasise that corpus linguists of all stripes cannot avoid this question. *If a speaker had no choice about the words or constructions they used, then language would be invariant.*[2] Bauer (1994: 19) comments that "change is impossible without some variation". Logically, therefore, studies of language variation and change should be primarily conceived as questions of choice.

[1] This paper was written in part as a result of numerous discussions over the course of editing Aarts, Close, Leech and Wallis (2012), which explains why a number of papers from this volume have been cited.

[2] Choices can also be said to *interact* in two important distinct senses. First, they *enable or exclude* other choices in a binary fashion (excepting the odd innovation), primarily through the operation of grammatical rules. Second, they *influence* other choices, by changing the likelihood that a second choice will be made in a particular direction given the first decision, for example through priming or templating. It should be obvious that the precondition of investigating the interaction of choices is the reliable retrieval of choice conditions in the first place.

It follows that rather than simply evaluate changes in normalised frequencies of individual forms, we need to try to frame experiments to investigate changing use within a group of alternative forms. Likewise, semiasological studies can only show variation of semantic types where alternatives to those types are taken into account.

Herein lies the difficulty. In laboratory experiments it is straightforward to constrain choices in advance: present subjects with a stimulus and ask them to press button A or B in response. The experimenter designs in the choice. However corpus research is performed on unconstrained responses. Researchers carry out *ex post facto* analysis of data. Variationist 'linguistic choice' research therefore requires the inference of the *counterfactual*, i.e. alongside what subjects wrote or said, we need to infer what they could have written or spoken instead.

Unfortunately, it is frequently non-trivial to identify counterfactual 'alternates' (e.g. 'non-progressive but progressivisable VPs'), and this fact has produced a number of practical objections from linguists. In this paper we explore three principal objections, what they imply, and how they can be overcome to the limits of our data. We also show that even where they cannot be wholly overcome by identifying a definitive alternate pattern in every case, the *perspective* of linguistic choice experiments remains optimum for obtaining linguistically meaningful results. Even if our data does not match this theoretical ideal, we can still approximate towards it. Recognising the limitations of experiments is central to responsible scientific reporting.[3]

This is not to make a case for *only* focusing on strict choice. Whereas lab experiments cue choices, corpora can provide much better estimates of the overall likelihood of encountering a form 'in the wild'. Empirically obtaining a rate of exposure per million words ('pmw') may help us rank forms by frequency to guide dictionary construction, design language teaching syllabi, or simply to provide valuable background information regarding which forms are more dominant. However, absolute frequency of exposure is not the same as the preference for a form given a choice of alternates. This paper explores a range of approaches that allow us to distinguish variation in the *opportunity* to use a form (affected by many factors including context) and variation in the *choice* of a particular form when that opportunity arises.

As a contribution to a methodological debate, this paper is not intended to supplant linguistic concerns – far from it – but rather to discuss ways in which linguistic hypotheses may be tested against corpus data in a manner maximally commensurable with those of other types of linguistic research (see Schönefeld 2011).

As an aside, it may be worth noting that many sciences employ what we might term ecological models of choice, i.e., choices made by organisms in a naturalistic context. Examples include

> Market research: researchers are tasked with finding out in what circumstances might shoppers buy product A rather than product B. This choice is the focus of the research (the dependent variable). Other variables, such as whether they purchased other items at the same time, locations of products in the store, etc., may be considered as independent predictor variables.
>
> Plant morphology: consider the choice of a rose to grow a flower from a node: not all nodes, where leaves appear, contain a flower. Different environmental factors and species cause the numbers of flowers produced to vary. The meaningful baseline for flower growth would be the number of nodes capable of producing a flower.

The second example illustrates the sense we employ the term 'choice' in this paper. Although we have referred to living organisms, mathematically the principle can be extended to *the unavoidable*

[3] The process of obtaining an optimum experimental design is occasionally referred to as one of 'formulation and operationalisation of hypotheses' (Gries 2009: 176), where *formulation* refers to the definition of variables and cases, and *operationalisation* the definition of corpus queries. Our argument in this paper, developed from Nelson *et al* (2002: 258-262) is that this process of formulation frequently does not take into account the importance of relevant baselines.

process of selection from any set of alternative outcomes arising in a process. The capacity for conscious decision-making is not a precondition for this selection to take place.

These examples present similar challenges to the linguist attempting to extract choices from naturalistic data. If you wish to study why shoppers or roses do what they do, and what outcomes they produce, you should pose the research question in terms of a logical model of choice. Shoppers and plant nodes incapable of making the selection are discounted. More complex models may include the impact of co-occurring outcomes – repression of new growth when a plant has flowers on other stems, for example – but such models are best built on the basic choice model.

Note that we may need an explicit *theory* predicting the counterfactual, e.g. where a flower failed to appear. Different plants will grow flowers at different points in their structure – in the case of the rose, at the apex of the stem rather than the side, so the flower represents the end point of growth, and only the growing tip is capable of making the 'choice'. As we shall see in section 5.2, studies of choice can include evaluating an effect downstream of the selection.

2. Some preliminaries

Aarts, Close and Wallis (2013) describe a simple choice experiment. They discuss a study where a single sociolinguistic independent variable (IV, time), affects a particular linguistic choice (modal *will* vs. *shall*) made by a participant in the spoken *Diachronic Corpus of Present-day Spoken English* (DCPSE) corpus. The dependent variable consists of a choice available to speakers.

2.1 Mutual substitution

Alternation depends on *mutual substitution*:

> *Given a corpus, identify all events of type A that alternate with events of type B, such that A is mutually replaceable by B.*

This means that we aim to identify all points in the text where a participant could plausibly substitute B for A, and vice versa, limited by grammatical constraints. A stricter variation of mutual substitution we will term *independent* mutual substitution, where meaning is held to be constant across the choice:

> *Given a corpus, identify all events of type A that alternate with events of type B, such that A is mutually replaceable by B without altering the meaning of the text.*

Note that if the choice does not alter the broader meaning of the utterance, then it is easier to see it as essentially self-contained: it is unlikely to be semantically affected by its context (see the 'ecological objection', 6.3 below), and the choice has no other effect on the sentence.

We will return to the question of constant meaning in section 2.3 below. First we will consider an experiment which holds to this ideal.

Suppose A represents the modal verb *shall* and B *will*. Whereas almost all cases of *shall* may legitimately be replaced by *will* (a mere eight are formulaic in DCPSE) the reverse does not apply. To ensure mutual substitution is possible, Aarts *et al.* limited cases of *will* and *shall* to the first person. They further eliminated interrogative and negative forms because different constraints are likely to apply. Even apparently 'lexical' choices may require grammatical restrictions for their correct analysis.

The choice of A vs. B (where only one or other is possible) is exclusive, so observed change may be characterised as *replacement*. The proportions of cases of Type A and B are inversely related, and when one increases the other must decline. With more than two alternates, Type A still 'replaces' B_1, B_2 etc. but we cannot be sure which particular counterfactual alternate is being replaced.

However, consider what happens if we violate the principle of mutual substitution by including a non-substitutable form within our dataset, which we will call Type C. The most common way such cases are introduced into a corpus experiment is through word-based baselines (per million words, etc). In this case Type C represents every other word in the corpus *except* first person declarative *will* and *shall*. If we do not characterise the problem in terms of variation in the choice of A or B, we cannot distinguish variation due to *use* (A vs. B) or *opportunity* (A+B vs. C). As we shall see in Figure 2, the normalised frequency of employing *will* or *shall* in the first place is not evenly distributed across DCPSE subcorpora. Variation in Type C cases is confounding noise which can obscure the relationship between A and B. If such cases cannot be eliminated altogether, they should be minimised.

Finally, if cases of A and B are mutually substitutable, every choice point is *free to vary*, i.e. a genuine choice exists and all cases could theoretically be of one type or the other. The true rate (see below) of either condition A or B therefore ranges from 0 to 1.[4]

By isolating first person positive declarative *will / shall*, Aarts *et al.* could focus on freely-varying instances. It was then possible to consider further specialising, to delineate subsets by modal semantics, and generalising, to include the clitic *'ll* and semi-modal *BE going to*.

This principle of mutual substitution is extensible beyond a single binary choice. To take the example of *shall* / *will* / *'ll* / *BE going to*, provided *shall* may be replaced by *will* / *'ll* or *BE going to*, and vice versa, it is reasonable to consider each case against the remainder. Given that *'ll* is a contraction of *will*, and that *BE going to* is a separate modal class, we may wish to hypothesise that these choices are ordered, and form the decision tree {{*shall*, {*will*, *'ll*}}, *BE going to*}. The authors then examined three distinct and potentially independent trends: the ratios modal vs. semi-modal, *shall* vs. *will* / *'ll*, and the rate of contraction of *will*. Alternative decision paths might be conceivable, but these were chosen on morphosyntactic grounds. Note that four alternates entails three 'degrees of freedom', i.e. the data can be explained by three patterns of binary alternation.

2.2. True rate of alternation

Given a pair of alternates, A and B, we can define the *true rate* of A as the fraction of cases that are A out of a given choice A+B. This is the conditional probability

$$p(\mathrm{A} \mid \{\mathrm{A}, \mathrm{B}\}) = \frac{F(\mathrm{A})}{F(\mathrm{A}) \quad F(\mathrm{B})},$$

where $F(\mathrm{A})$ is the total number of cases (unnormalised frequency) of Type A, etc.

Having formalised the true rate in this way, this definition permits some simple statistical methods. We can compute a *confidence interval* on this rate (Figure 1a, b). The concept of a true rate is also implied by *contingency correlation tests* (log-likelihood, ^{2}, etc.). The following contingency table can be easily constructed:

IV DV:	**A**	**B**	**Total**
condition 1	$f_1(\mathrm{A})$	$f_1(\mathrm{B})$	$f_1(\mathrm{A})+f_1(\mathrm{B})$
condition 2	$f_2(\mathrm{A})$	$f_2(\mathrm{B})$	$f_2(\mathrm{A})+f_2(\mathrm{B})$
TOTAL	$F(\mathrm{A})$	$F(\mathrm{B})$	$F(\mathrm{A})+F(\mathrm{B})$

Table 1: A sketch 2 2 table for a ^{2} test to determine if the true rate of A, out of A + B, varies between two values of an independent variable IV. The notation f_1 and f_2 refers to subset frequencies.

[4] The assumption that values are free to vary underpins contingency tests (Wallis 2013) but is frequently overlooked. Statistical tests are pretty robust, but that is no reason to assume that they give the correct results in conditions they were not designed for. See http://corplingstats.wordpress.com/2012/09/30/free-to-vary.

The null hypothesis of this test is that the true rate of Type A events out of the set {A, B} is constant over conditions 1 and 2. The sum $F(A)+F(B)$ is the *baseline* for this evaluation.

Aarts *et al.* initially applied 2×2 ^{2} contingency tests over a 'time' variable, comparing rates between two subcorpora in DCPSE.

	shall	*will*	**Total**	*p*(*shall*)	^{2}(*shall*)	^{2}(*will*)	**Summary**
LLC (1960s)	110	78	188	0.59	1.32	1.45	$d^{\%}$ = -30.24% 20.84%
ICE-GB (1990s)	40	58	98	0.41	2.53	2.79	= 0.17
TOTAL	150	136	286		**3.85**	**4.24**	^{2} = **8.09**

Table 2: 2 2 ^{2} for *shall* and *will* between ICE-GB and LLC (spoken, first person subject, declarative), excluding *'ll* and negative cases (after Aarts *et al.* 2013). The rate *p*(*shall*) is relative to the opportunity, i.e. *p*(*shall* | {*shall*, *will*}).

Note that once we know that the true rate is meaningful and represents a freely varying choice in the data, other measures may also be usefully cited. Table 2 includes, in the 'Summary' column:

The percentage difference between true rates, $d^{\%}$ = –30% decrease on LLC (20%, at a 95% level of confidence), or a fall between –10% to –50%.

Cramér's 2 2 measure of association (ranging between 0 and 1 where 0 = independent, i.e. no effect, and 1 = dependent), a different measure of the size of the effect.

Results of goodness of fit chi-square tests (bottom row, bold) and the 2×2 test for homegeneity.[5]

Note that this principle applies over any number of conditions, including time series. Aarts *et al.* plotted *p*(*shall* | {*shall*, *will*}) over aggregated points in time, inferring a logistic S-curve where *will* can be seen to replace *shall* (Figure 1a).

The authors then examined the impact of including clitic *'ll* within the alternation. Note how the introduction of *'ll* (a shortened form of *will*) in Figure 1b alters the logistic curve and what we might infer from it.

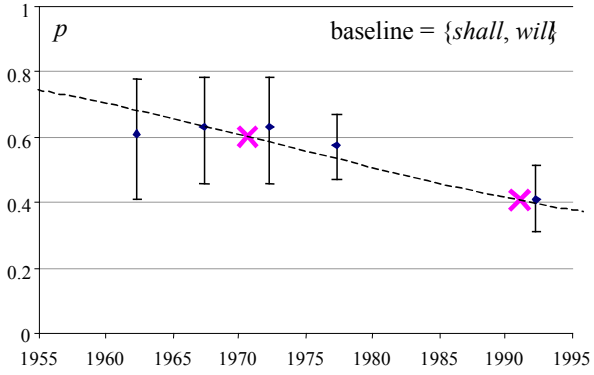

Figure 1a: Declining use of *shall* as a proportion *p* of the set {*shall, will*} with first person subjects, 5-year data from DCPSE (after Aarts *et al.* 2013).

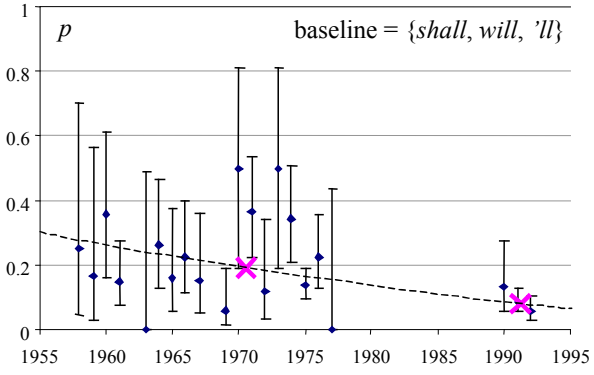

Figure 1b: The effect on the logistic model for *shall* of expanding the baseline to include clitic *'ll* (annual data).

Figure 1a implies that a shift in majority usage from *shall* to *will* occurred in the mid '70s. However, once we include the cliticised form into our calculations in Figure 1b, it seems that *will / 'll* became dominant much earlier, i.e. at some point in the first half of the 20th Century. Note that *both observations may be correct* – the replacement of *shall* by explicit *will* may indeed be a delayed reaction to the first change. *Changing the baseline alters the hypothesis.*

[5] This paper is not concerned with statistical measures but with principles of experimental design. For more information on what these measures mean see Wallis (2013) and http://corplingstats.wordpress.com/2012/04/01/crib-sheet.

Another way of combining these trends is the three-way alternation (*shall / will / 'll*) in Figure 1c. All probabilities are proportions out of the baseline, the total probability is 1, and there are two degrees of freedom. Note how this sketch shows that the proportion of *shall* falls against the baseline over time (cf. Figure 1b) but also as a proportion of *will+shall* (Figure 1a). However it also shows that*will* also falls 'in real terms', i.e. there is a tendency towards cliticisation.

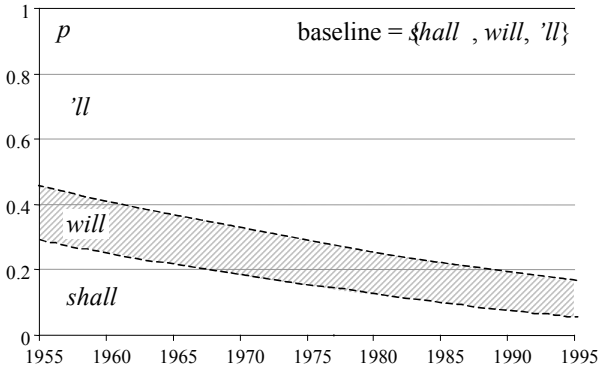

Figure 1c: Plotting *shall+will* against *'ll*, same data. For all points $p(shall) + p(will) + p('ll) = 1$.

Figure 1c also reveals another problem of conflating opportunity and use. Suppose that, by mistake, we included a number of 'Type C' cases that did not alternate with both *shall* and *will*. Imagine that *'ll* in Figure 1c could not substitute for *shall* or *will*. These cases would mislead us in our attempts to fit $p(shall)$ to a logistic curve, because $p(shall)$ could never rise above the dotted line above *will*.

S-curves level out at probabilistic extremes, 0 and 1, but if the variation includes a non-alternating form, we are forced to work within a changing 'envelope' independent from *shall* vs. *will*. This mathematical problem matters most when the proportion of non-alternating Type C cases is high. Simply increasing the amount of data will not solve this problem. The solution is to narrowing the baseline to the freely-varying choice.

2.3. Must meaning be constant?

In section 2.1 we defined independent mutual substitution in terms of a requirement for meaning to be constant across the choice. When choosing between *will* and *shall*, for instance, if the speaker has to change the meaning of the sentence, then the decision is not simply a local act of lexical preference. This restriction is preferred, for the reasons noted previously, because the choice can be treated as independent from its context. Requiring meaning to be constant means reviewing every case for alternation and eliminating non-alternating cases. For example it is impossible to replace *shall* with *will* in the formulaic *ye shall be saved* (DCPSE DL-J01 #49) without changing the purpose of the utterance. However, there are circumstances where such a restriction may prove unfeasible or unnecessarily restrictive, and conclusions must take account of this.

From a sociolinguistic perspective, Lavandera (1978) discusses a cline from phonological variations such as stress patterns which have no ostensible meaning, through grammatical and lexical variation where referential meaning is constant but alternants may differ in integral meaning (e.g. level of formality), to a 'dangerous hypothesis' where even referential meaning may be allowed to change. She echoes the requirement for checking cases for genuine alternation, calling for a "strategy of setting aside more and more contexts where both alternants occur but do not say exactly the same thing" (1978: 178).

The argument in this paper is pluralistic. In designing an experiment where speakers choose between two or more forms, we need to decide as far as possible whether we can exclude semantic explanations for variation. If the choice has no effect on the meaning of the sentence we may

eliminate semantic explanations from our conclusions. However where this is not possible we will need to consider this question more carefully.

A choice may modify the meaning but still be a free choice available to the speaker, at least as a plausible null hypothesis. The key requirement is that every instance included in an experimental dataset must not be fixed (e.g. quotations, or idiomatic cases of *shall*) and should be free to vary in a like manner to all other instances (there should be the same alternative forms for all cases).

Logically and mathematically, alternation studies do not require meaning to be identical in both cases for the choice to be available. In section 5, we consider choices such as the decision to premodify an NP, where the additional attributive adjective phrase changes the meaning of the head noun. But if the two meanings are clearly incompatible – *I* vs. *you*, modal *can* vs. *will*, for example – then the choice will almost certainly be predetermined by the speaker's intention, topic and context. Even if an experiment that evaluated such 'alternations' made sense, we would have to eliminate contextual explanations before considering others.

This leads us to one final comment: if a choice is 'unconscious', in that a speaker is unaware that they are making it at the time, then conscious explanations of that choice are more easily set aside. Such choices will probably be routine, involve unmarked forms, have little semantic impact on the surrounding utterance, and, ideally, be found in unplanned spoken data. We can still observe variation of conscious choices, but if all we are doing is observing writers 'crafting' non-spontaneously, cognitive explanations will be difficult to defend.

3. Refining baselines and the ratio principle

3.1. Word-based baselines

We have seen an example of how one might study the ratio of *shall* to a baseline of *shall+will* in conditions where *will* is mutually substitutable with *shall*. It is not always straightforward to identify a baseline like this (see also 6.1 below). A common response to the difficulty of identifying counterfactual 'Type Bs' has been to investigate changes in "normalised frequency counts" of the object under study (Type A), i.e. *frequencies standardised against a word baseline* (per thousand words, per million words, etc.). However, then 'Type B' is *all words other than A*!

This baseline is rarely optimum. Such a model implies that every time any word is uttered, a speaker has the *option* of uttering a Type A expression instead. Further, as we saw in Figure 1c, it requires that, at minimum, the prior likelihood of doing so is constant over the entire corpus (the maximum should be flat), or at least between subcorpora over which the comparison is being made.

There are some experiments where a word baseline may be valid. For example, we could design an experiment to find variables which predict when speakers pause, self-correct or stop sentences abruptly. With no prior knowledge of the structure of English the default assumption would be that this could take place at any word. Hence *any* variation from this is considered remarkable and worth examining.

However, most corpus studies in the literature focus on elements which are unlikely to be evenly distributed in the corpus, but whose expression depends on multiple prior factors including grammatical restrictions on use. Employing a word baseline means that we are likely to find changes in the ratio of Type A to the number of words which are due to changing *opportunity* to use A rather than changing *use* of A. As a rule, "over-general" baselines conflate opportunity and use.

To continue the example of incomplete sentences, suppose that following an initial investigation we refine our hypothesis to focus on incomplete clauses. As soon as we do this however, clause frequency will be a more revealing baseline than word frequency.

3.2. Refining baselines

One way we can see the effect of different baselines is simply to plot their ratio across the contrast under scrutiny. Figure 2 plots the proportions of possible baseline forms – words, tensed VPs, all modals and the pair {*will*, *shall*} – in DCPSE, across the time contrast 1960s:1990s (LLC:ICE-GB).

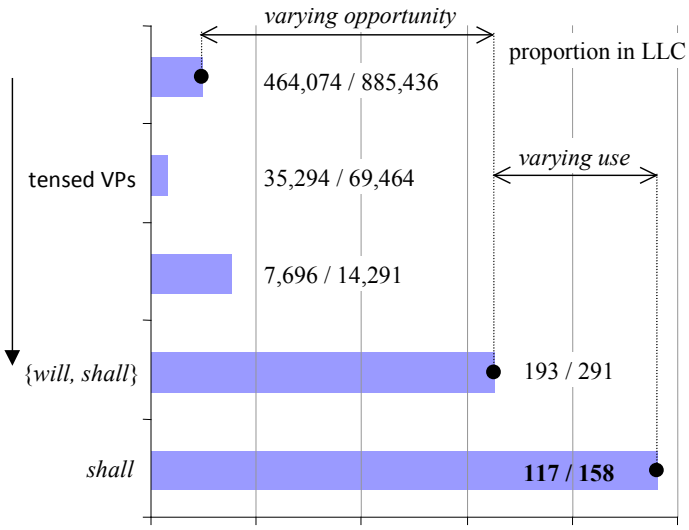

Figure 2: Bar chart representing the degree to which potential baseline forms for examining change in uses of modal *shall* are found more frequently in the earlier (LLC, 1960s) component of the DCPSE corpus. If the choice of baseline did not matter, these proportions would be identical. Variation between the optimum {*will*, *shall*} baseline and the number of words default baseline constitutes variation of opportunity, which is a distracting factor if we wish to study when speakers say *shall* rather than *will*.

Around 52% of words are found in the earlier subcorpus, whereas tensed VPs are more evenly distributed at nearly exactly 1:1. The proportion of all modals found in the 1960s data is 54%. However, the proportion of first person declarative *will* or *shall* combined in the LLC data is 66% of the total (2:1). Finally, 74% of all cases of first person declarative *shall* are found in the 1960s subcorpus.

The LLC subcorpus has a higher proportion of *will / shall* first person declarative forms than the set of all modals would lead one to expect. Perhaps the earlier data contains more frequent expressions of obligation or prediction. Perhaps modal use is changing over time in other ways. Irrespective, only when we alight on the set of first person declarative *will / shall* cases do we identify all cases of the opportunity to express *shall*. Employing the nearest baseline to a change allows us to factor out variation of opportunity.

3.3. Variation and reproducibility

A second problem with employing over-general baselines concerns reproducibility between differently-sampled corpora. If different types of text provide different opportunities to employ particular forms, then comparability between corpora can be guaranteed to improve by factoring out this variation.

Bowie, Wallis and Aarts (2013) studied the impact of spoken text categories on changing modal use over time in DCPSE, noting that modals could only be employed in tensed verb phrases (tVPs). Figure 3 plots the per million word frequency of tensed VPs in DCPSE. Across the entire corpus, the ratio of tVPs to words is almost constant (Total column, see also Figure 2). However, once the data is broken down by text category, this statement no longer holds true, as other columns illustrate. Indeed, we find two types of variation. First, we find significant synchronic variation between text categories – some categories have more tensed VPs than others (they present more

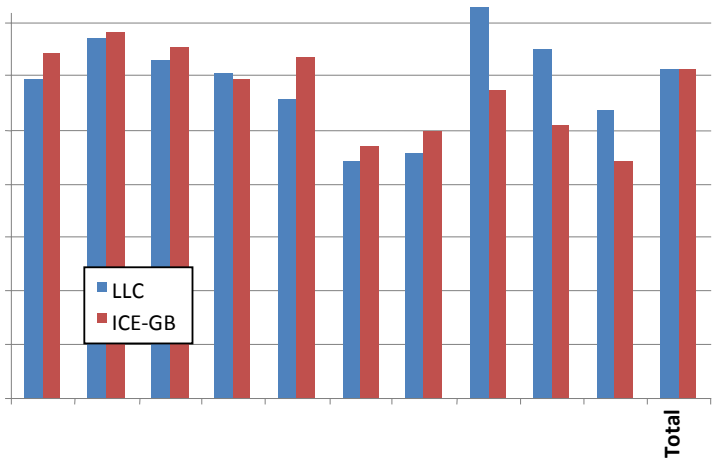

Figure 3: Frequency of tensed VPs per million words, by text category (*x* axis), compared across the two 'time' subcorpora of DCPSE, LLC (1960s) and ICE-GB (1990s), after Bowie *et al.* (2013).

opportunities for modal use). Second, these tensed VPs, and therefore opportunities, increase over time in some categories, whereas in others they decrease.[6] This variation must confound any study which examines the ratio between core modals and words.

Algebraically, the relationship between words, tensed VPs and modals can be expressed as a series of ratios:

$$\frac{F(\text{modal})}{F(\text{word})} \quad \frac{F(\text{modal})}{F(\text{tVP})} \times \frac{F(\text{tVP})}{F(\text{word})}$$

This formula is an example of the *ratio principle*: the ratio of the first and last term a:d in a sequence a:b, b:c, c:d is equal to the product of that sequence. Since each ratio is subject to variation, it follows that, as far as possible, *we should study each ratio separately* to isolate these different levels of variation. This way we obtain more information about differences within the corpus across a contrast than simply studying a:d alone.

If we know that the ratio of tensed VPs to words (what we might call 'tensed VP density') can vary, then the ratio *F*(modal) / *F*(word) includes this variation, and therefore *a modal:word ratio is suboptimal in* explaining *change in the core modals*. Fortunately, we can simply factor out tVP density variation out by changing the baseline and simply considering the first ratio above, i.e. *F*(modal) / *F*(tVP).

It follows that tensed VPs must be a more reliable baseline for studying changes in core modals than words. This is not to argue that every tensed VP could plausibly substitute for a given core modal, nor that the set of core modals represents an alternate set. We have not eliminated non-alternating cases. Rather, we simply note that it is frequently possible to *improve* upon word baselines in a way that eliminates variation in tVP density and reduces experimental noise overall.[7]

Other researchers have noted how changing baselines can lead to different conclusions. Smitterberg (2005) changed the baseline for investigating the incidence of progressive constructions from words to 'progressivisable VPs'. He identifies a number of 'Type C' contexts in which verb phrases

[6] This variation is not due to tense: a plot of all VPs obtains a similar pattern.
[7] An advantage of parsed corpora (such as ICE-GB and DCPSE) are that they permit the reliable identification of numerous grammatical categories, allowing us to experiment with grammatical baselines with relative ease.

cannot be progressivised, including imperatives, non-finite VPs, and the *BE going to* future construction. He found that if he ranked text types by the writers' tendency to employ the progressive construction, and compared the results obtained with a per-million word baseline with those using his 'progressivisable VP' baseline (S-coefficient), he obtained a different ordering in some cases. Diachronic conclusions in his data, on the other hand, were broadly similar.

We would hypothesise two reasons for these results, which are likely to apply to numerous questions of clausal and phrasal density.

1. Language variation between text category. Not all types of text exhibit the same ratio of VP to word. For example, in DCPSE, conversation has a higher number of verb phrases per word than commentary (Figure 2). Focusing on a VP baseline can therefore affect the ranking of texts. This source of variation also applies to comparing results from different corpora.

2. Sampling noise. Corpora like DCPSE are sampled in a 'balanced' way as far as possible, but within specific text categories, or across particular contrasts, this is harder to guarantee. With fewer texts per category, and a smaller number of participants and topics, sampling variation is introduced into each category. While it is true that sampling noise affects all levels in a decision tree (VP progressivisable VP progressive VP), focusing on baselines keeps noise to a minimum.

In conclusion, the ideal of strict alternation may not always be possible. In some cases it may only be possible to identify a set of potential elements that could *plausibly alternate* with Type A, and conclusions must be phrased appropriately. It is not possible to claim that a decrease in modal use out of tensed VPs represents the replacement of modal forms with non-modal ones, because non-alternating VP forms may have increased for other reasons over the same contrast. Nonetheless, tensed VPs is a more meaningful baseline than words.

Finally, there is one further mathematical problem with employing over-generous baselines, separate from the problem of introducing unwanted variation. This concerns the mathematics of the Binomial model and its approximations (log-likelihood, ^{2}, etc).[8] Suffice it to say that if large numbers of Type C cases are introduced into a baseline, this class of significance test becomes excessively conservative, causing us to reject results which would be statistically significant if the experiment was formulated more precisely.

4. Surveying 'absolute' and 'relative' variation

The concept of 'normalisation' remains an important one: to make results comparable between different corpora. However as we have seen, a word-based baseline is rarely optimal for this purpose.

In many studies it may be useful to employ more than one baseline provided, of course, that one makes clear which baseline is employed at each point. Considering patterns of change within the core modals, Bowie *et al.* employ two baselines (Figure 4):
A) tensed VPs as a baseline for global ('absolute') change,
B) the modal set, i.e. modal tensed VPs, for within-set ('relative') change.

Note that both terms, 'absolute' and 'relative' are in fact *relative to something*, i.e. the baseline, but they convey an important distinction. Relative change of a given core modal, e.g. *must*, refers to changes in the proportion, or share, of the core modal set occupied by *must*. On the other hand absolute change of modal *must* against tensed VPs represents the fluctuation of *must* as a proportion of all tensed VPs.

[8] See http://corplingstats.wordpress.com/2012/09/30/free-to-vary.

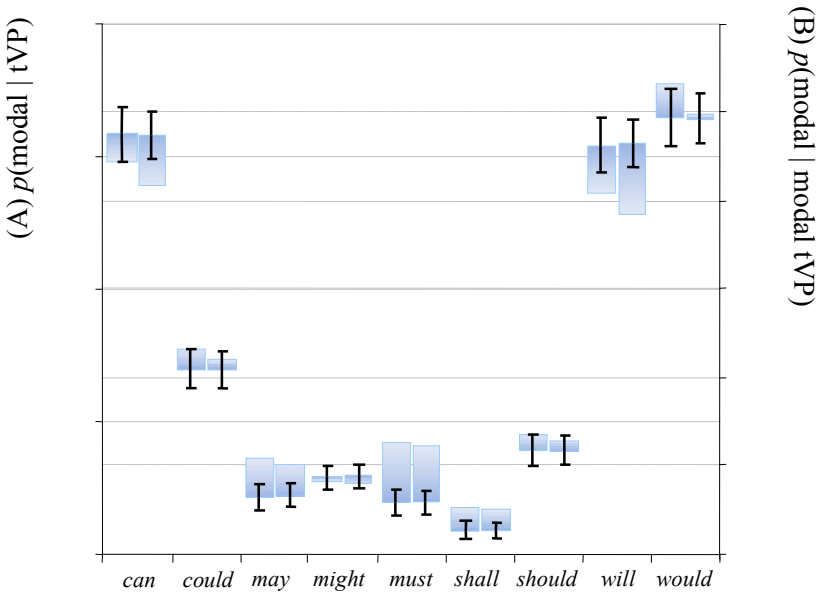

Figure 4: Changes in individual core modals over time using DCPSE, evaluated against (A) tensed VPs, left and (B) modal set, right, after Bowie *et al.* (2012). The floating columns represent simple differences in probability (start = light, end = dark) with confidence intervals placed at the end point.

Neither baseline represents an alternation set. For example, *must* does not alternate with the other core modals but with semi-modal *have [got] to*, which is excluded from this 'relative' baseline but included in the 'absolute' one. This type of study is therefore not an alternation study but what we might term a *survey of patterns of change*. Such a survey allows us to identify a map of potentially interesting change which may then be subject to a deeper, and more time-consuming alternation investigation. It also permits us to identify where we may have insufficient data for such research and results are only indicative.

Employing a global and a within-set baseline in a study also permits us to combine them to distinguish 'typical' and 'atypical' change. Examining change within the modal set (B) allows us to distinguish between modals which change consistent with the modal set (typical change), or change in a statistically different manner to the set (atypical change). Clearly, atypical change may be in either direction: the item may be changing faster or slower than the set.

Figure 4 exposes the fact that, e.g. *would* is a typical modal (it does not change significantly compared to the set: the right hand difference column is smaller than the confidence interval), whereas *may* is atypical and ahead of the overall downward change ('leading the way'). Finally, *can* is atypical in moving in the opposite direction ('bucking the trend').

5. From alternation to choice

So far we have tended to consider examples of alternation within grammatical categories. However it seems apposite to give examples of some possibly more unusual 'alternation' research questions that the method outlined supports. In order to do this, we will need to relax the constraint that meaning is held to be constant (see section 2.3), and our conclusions must be adapted accordingly.

5.1. Simple grammatical interaction

Nelson *et al.* (2002: Chapter 9) and Wallis (2003) summarise simple grammatical interaction experiments.[9] Thus far the experimental condition (the independent variable) was a sociolinguistic category (time, genre, etc), whereas the outcome (dependent variable) was a lexical or grammatical choice. However, there is no reason why the independent variable cannot also be grammatical, with one proviso. It must be possible for each variable to be applied to the same location (i.e. the same

[9] See also www.ucl.ac.uk/english-usage/resources/ftfs/experiment3.htm

case) in the corpus. The main way to ensure this is to construct queries to capture patterns for each cell in the table independently.

Nelson, Wallis and Aarts (2002) provide several examples of this type of experiment. Comparing the interaction of clausal mood and transitivity (2002: 274), a contingency table of clause frequency is subdivided so that each individual cell will have both features (Table 3) and clauses are classified by both variables independently. This approach can be extended to more complex structures (2002: 278) using a *Fuzzy Tree Fragment* template to identify structures with phrasal features in different locations.

This independent variable does not consist of mutually substitutable alternates. Change of mood (from interrogative to exclamative, say) radically alters the semantics of the sentence. Note however that alternation experiments require that *the dependent variable* is free to vary.

CL		**dependent variable (transitivity)**			
		DV = *m*	DV = *d*	DV = 0	TOTAL
independent variable (mood)	IV = *e*	`CL(exclam, montr)`	`CL(exclam, ditr)`		`CL(exclam)`
	IV = *i*	`CL(inter, montr)`	`CL(inter, ditr)`	-	`CL(inter)`
	IV = 0	-	-	-	-
	TOTAL	`CL(montr)`	`CL(intr)`	-	`CL`

Table 3: Enumerating a contingency table for mood transitivity, from Nelson *et al*. (2002).

Similarly, complementation patterns tend not to be mutually substitutable! We can either accept that this particular study is a general survey of change (cf. section 4) or limit data to a subset of cases where alternation of transitivity would be plausible, such as the decision to add an optional object.

5.2. To add or not to add

So far we have suggested that dependent outcomes consist of different explicitly-expressed elements: *who* vs. *whom*, monotransitive vs. ditransitive clauses, etc. However a legitimate outcome may also be the *absence* of an element (cf. Nelson *et al*. 2002: 280). Like the rose growing a flower, the decision to add a construction is an alternation decision.

Wallis (forthcoming) examines the problem of *sequential addition of constructions*, that is processes of phrase or clause formation where speakers have the repeated choice of adding a construction or not. Each choice point presents two alternate options: to add a construction ('Type A'), in which case the choice becomes available again, or to stop ('Type B').

For example, when speakers form noun phrases with common noun heads they are faced with a series of decisions to add an attributive adjective phrase before the noun, thus: *the ship*; *the tall ship*; *the old tall ship*, etc. Employing an alternation model allows us to recognise that we may usefully study the probability of adding an adjective phrase at each stage. The null hypothesis is that this probability is constant, irrespective of the number of adjective phrases previously added.

Table 4 summarises the frequency distribution of common noun-headed NPs containing x attributive adjective phrases ('at least x' F). We can then analyse what happens to the probability of each successive choice. We work out the probability of adding the x-th adjective phrase by simple division, i.e., $p(x) = F(x) / F(x - 1)$. There are almost 156 thousand NPs with common noun heads, of which 36 thousand (23 percent) have at least one attributive adjective phrase, and so forth.

***x* adjective phrases**	**0**	**1**	**2**	**3**	**4**
'at least *x*' *F*	155,961	35,986	2,892	151	6
probability *p*		0.2307	0.0804	0.0522	0.0397
upper bound p^+			0.0832	0.0603	0.0709
significance ($p > p^+$)			s-	s-	ns-

Table 4: Sequential analysis of probability of adding an adjective phrase to an NP with a common noun head in ICE-GB, from Wallis (forthcoming).

The key observation is that the probability *falls systematically and sequentially*, i.e. every time an adjective phrase is added, it becomes more difficult to add the next.

The decision to add an adjective phrase necessarily alters the meaning of the utterance, even if "the thing itself" referred to remains constant. Far from avoiding research questions where decisions impact on each other, this relatively unusual experimental design attempts to measure the effect of such sequential decisions. Possible reasons for this systematic decline are discussed in the paper,[10] which also demonstrates that this phenomenon is *not* simply a universal tendency of grammatical constructions.

This concept of "choice experiment" can be extended to refer to any question of lexical or grammatical choice provided that Type A and B conditions can be reasonably reliably identified.

5.3. Grammatically diverse alternates

A different kind of choice is suggested by Biber and Gray (2013) who look at a tendency for verbal expressions to be replaced by nominalisations ('derived nouns' and 'converted verbs'). Their study plots both tendencies separately against a word baseline. They find that the frequency of nouns increases from 250 per million words ('pmw') in C19th academic writing texts to 330 pmw in C20th texts, whereas the verbs decrease from 90 to 75 pmw. However they note that such a finding does not allow them to conclude that verbs are being replaced by nouns. For one thing, the growth in nouns is not matched by a comparable decline in verbs.

Logically, if the hypothesis is that a decision is being made to substitute one grammatical construction for the other, then as far as is possible, *this hypothesis should be tested*. We cannot infer from variation per million words that replacement is taking place. We could simply be seeing changes in opportunity, or several different changes happening at once. The task is to 'tease out' the tendency we are interested in.

This type of experimental question would not be limited to alternates within a grammatical category. This kind of alternation is almost certainly *stylistic*, choices being made at a planning stage of the sentence (or indeed, text). The idea is that the grammar operationalises the choice as a nominal or verbal structure, so there are knock-on effects on the grammar. If a writer wishes to reformulate a nominalisation as a verbal structure (or vice versa), this is often not a matter of simple substitution of one form for another in the same 'slot', but will often involve a series of further changes to the text which will vary in nature depending on the larger structure within which the form in question is embedded. Take the following example:

(1) *Freud's analysis of the three dreams embedded in the story* is masterly. [W2A-002 #64]

[10] In the case of attributive adjectives the explanation for this decline is most likely semantic (including semantic ordering, c.f. *tall old ship*, and logical restriction, c.f. *tall short ship*). An alternative explanation, that this fall reflects a tendency for communicative brevity, seems less likely, as the probability continues to decline at each step.

Here the nominal structure is functioning as subject of the clause. We could not simply replace it with the corresponding finite clause *Freud analysed the three dreams embedded in the story*. One possible reformulation would involve a head noun of rather general meaning postmodified by a finite relative clause, e.g. *the way Freud analysed the three dreams embedded in the story*.

The obvious point is simply that this stylistic choice is also likely to be part of a network of linked choices. Our aim is to capture a dataset of 'choice points' to examine how the choice varies over time, across texts, or indeed, due to other choices made in the text.

One method would be to take frequencies of head nouns and verbal equivalents (*analysis / analyse*; *prediction / predict*; *increase* (noun) / *increase* (verb)) where alternate forms exist. In this case Type A could represent the nominal form and B the verbal form. Focusing on head nouns avoids premodifiers such as *analysis planning stage*, which would be interesting in themselves, but seem to constitute a different case.

The initial assumption might be that all such forms alternate in each case, but these should then be checked to see if the choice in that context is plausible. For example, nominal forms occurring in set phrases and special terms (e.g. *frame* ***analysis***) might not freely alternate with a verbal form, so it might be preferable to exclude them from the frequency comparison The 'rate of nominalisation' of these forms would be the relative frequency or proportion $F(\mathrm{A}) / (F(\mathrm{A})+F(\mathrm{B}))$, as before.

The drawback of this approach is that it assumes that we can enumerate a finite list. Some forms may potentially alternate but be unlikely to do so, or an acceptable nominalisation or verbalisation may be in the process of developing (*diary / diarise*) or disappearing (*abomination / abominate*). When writers employ nouns or verbs they may also select a different head word, e.g. *detest* instead of *abominate*; *plan* instead of *diarise*. This strategy would apply for example, where no current generally acceptable morphologically derived form exists.

A complementary approach might be to cast the net wider and narrower at the same time. One option (easier in a parsed corpus than a lexical corpus) would be to examine particular grammatical structures which potentially alternate. For example, we could compare rates of general 'low content' verbs like *use*, *perform* etc. in 'V+nominalisation' constructions against an equivalent verbal construction (*perform an analysis* / *analyse*); or we might consider simple alternation of nominal structure and *-ing*-participial clause as complement of a preposition (*through analysis of the key themes / through analysing the key themes*). The strategy would be to 'divide and conquer' by focusing on the specific structures that permit alternation to take place.

In moving our conceptualisation of linguistic choice experiments from simple lexical replacement to wholesale stylistic decisions we risk Lavandera's (1978) objection that studying grammatical alternation "requires quite an ingenious dismissal of possible differences in referential meaning" (see also section 2.3). An important question in operationalising this broader notion of choice is the degree to which the structure of utterances can vary while referential meaning is preserved.

Our proposal is that provided that the referenced object *itself* is constant, it is meaningful to employ an experimental model of choice regarding how it is referred. Moreover, as we saw, such models are mathematically provably superior to alternatives. On the other hand, the explanations of results must consider the degree to which the speaker or writer is aware of their choices.

6. Objections

So far we have concentrated on the benefits of the linguistic choice approach. These include *psycholinguistic plausibility* (choices are genuinely available to speakers or writers at each point), *research focus* (use is differentiated from opportunity) and *empirical reliability* (minimising invariant 'Type C' terms).

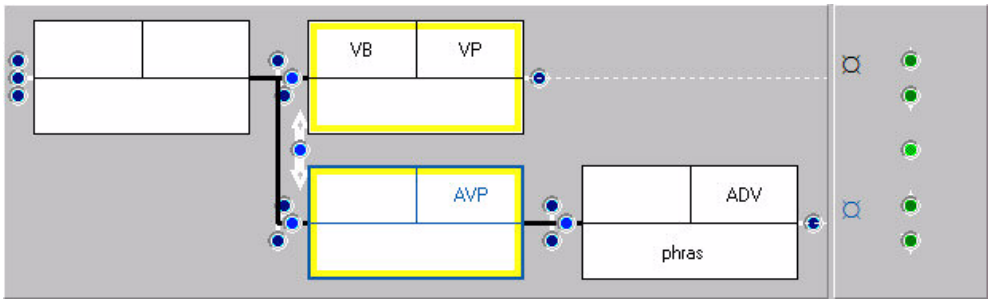

Figure 5: FTF for VPs followed or preceded by phrasal adverbs.

However there are three main lines of objection that have been made to this approach. These are (1) alternates are not reliably identifiable, (2) baselines are arbitrarily chosen by the researcher, and (3) different ecological pressures apply to different terms. These objections are not insurmountable, but require careful consideration.

6.1. Alternates are not reliably identifiable

In the case of nominalisations, we saw how the task of identifying alternates in a corpus may not be simple. However, nominalisations at least allow us to identify converted verbs or derived nouns. Consider the problem of identifying alternates of phrasal verbs, defined loosely as constructions consisting of a verb plus a phrasal adverb, which we will approach in two directions.

Previous studies such as Biber *et al.* (1999) and Gardner and Davies (2007) have relied on per million word frequencies to support claims of change over time. We observed that VP density is a potentially confounding factor in any study, and therefore the minimum position would be to employ a verb baseline. The question is whether we can improve on this.

6.1.1 Bottom up – serially identifying alternates

Phrasal verbs commonly have multiple meanings. The OED lists 15 core meanings of *put up* (excluding the literal 'place something at a high point'), but these core meanings are further differentiated. Latinate alternates of *put up* are likely to differ for each meaning, and may themselves have other phrasal alternates. One would need a thesaurus to enumerate these common meanings from first principles, and indeed a number of studies have employed WordNet for the semantic analysis of items.[11] A corpus allows us to focus on common phrasal verbs by compiling word-lists ordered by frequency of occurrence. We will briefly illustrate how this can be done with *put up* in ICE-GB. A lexical search identifies ten cases, all with *up* tagged as a phrasal adverb, e.g.

(2) Just because you're tolerant doesn't mean you have to like it to *put up* with it but you don't necessarily have to [S1A-037 #1]

We can generalise the lexical query into a *Fuzzy Tree Fragment* (Figure 5) by inspection of the tree structure for this sentence.[12] This FTF obtains some 7,000 cases in ICE-GB and can be treated as a template for further restriction. We can restrict the main verb to be a member of the set {*put*, *puts*, *putting*} and the phrasal adverb to *up*. We find 20 cases of *put* [X] *up* in ICE-GB, which may be manually reviewed in terms of thesaural alternatives.

In fact this frequency distribution is slightly exaggerated by repetition of topic: two cases of *put your feet up* and *put a motion up* appear in close proximity to another.

[11] See http://wordnet.princeton.edu. Note that we cannot solve this problem by simply examining the probability of adding a phrasal adverb to *put* (*up*, *down*, etc). This would give us information about meaning distribution (semiasology) rather than linguistic choice constrained by meaning (onomasiology).

[12] See Nelson *et al* (2002) on Fuzzy Tree Fragments. The authors strongly encourage researchers to examine corpus trees and experiment in order to obtain the most general and reliable queries.

put up	tolerate, suffer	4	*put up* with it [S1A-037 #1]
	?position	4	*put* your feet *up* [S1A-032 #21]
	?build, make	3	shacks *put up* without any planning [S2B-022 #118]
	display, project	2	*put up* two… trees [on the screen] [S1B-002 #157]
	sell	2	*put* the plant *up* for sale [W2C-015 #8]
	propose	2	*put* anything [a motion] *up* [S1B-077 #127]
	increase	1	*put up* the poll tax [W2C-009 #3]
	accommodate, house	1	we could've *put* the children *up* [S1A-073 #197]
	supply (finance)	1	*put up* the money [W2F-007 #36]

Note that although *put up* is considered a phrasal verb in ICE-GB for all its meanings, some linguists might wish to exclude literal meanings ('position' and possibly 'build'), or deal with them separately. A careful (and time-consuming) bottom-up approach would allow an alternation study to be conducted by searching for verbs which could replace *put up* and then checking cases for potential alternation. It would also be necessary to decide the degree to which the structure of the clause could legitimately change and still be considered an 'alternate' form. Consider:

put the plant *up* for sale, vs.
offer the plant for sale, vs.
put the plant *on the market*.

Ultimately it should be clear that the effort involved in classifying each phrasal verb and finding alternates is very high, and therefore a bottom-up approach is most suited for narrow-focus studies, e.g. verbs expressing a particular sense or set of senses, examining particular phrasal verbs, or examining phrasal verbs appearing in a limited subset such as those found in academic prose. To examine phrasal verbs as a broad phenomenon we have to approach the problem differently.

6.1.2 Top down – improving a verb baseline

An alternative top-down method would be to attempt to refine the baseline of all verbs by excluding 'Type C' forms which either *cannot* alternate with phrasal verbs, or *which alternate in specific ways* which should be separately addressed. This creates a three-way sorting problem: identifying verb forms where potential alternation may be classified as 'yes', 'no' or 'maybe' – the latter requiring us to examine specific corpus examples.

The idea would be to identify candidates for exclusion from the set of all verbs to form the set of 'phrasable verbs'. Candidates would be most likely drawn from copular and stative verbs. Greenbaum (1996:152) notes that phrasal adverbs tend to have "spatial meaning, literal or metaphorical." This implies that numerous verbs of movement, action and transition (dynamic verbs) are potential alternates for dynamic phrasal verbs.

Copular phrasal verbs are rare: ICE-GB has *turn out* (e.g. *which has turned out extremely fruitful*), and Huddleston and Pullum (2002: 288) cite some 8 examples including *end up* (*she ended up as captain*). Copular verbs could simply be identified and excluded.

Suppose we were to hypothesise that the set of dynamic verbs represents an *upper bound* of the set of possible alternates. This superset may include verbs that have no present phrasal alternate, but since the set of phrasal verbs is open, although we may risk underestimating the true rate, it still represents a credible baseline for comparison purposes, e.g. by genre or time period. There are some common examples of stative phrasal verbs including *stay in* and *stay on* ('remain'). Potentially borderline stative cases include *fit in [with]*, *put up [with]* (meaning 'tolerate', see above), *keep on/carry on/go on* ('continue') and *put down [to]* ('attribute'). Like copular verbs, these are finite, and so they could be enumerated and treated as special cases.

To form an estimate of a high frequency term in the corpus which requires manual discrimination we can examine a subsample and extrapolate from that subsample to the corpus.

To demonstrate the principle, we extracted 106,770 examples of main verbs (excluding copular forms) from ICE-GB. A 0.25% random subsample of this contained 263 main verbs of which 151 (57%) were manually identified as dynamic. A 95% Wilson confidence interval (Wallis 2013) gives us a dynamic rate of between approximately 135 and 167 out of 263. Scaling back up, this range is between 54,856 and 67,519 out of 106,770. In practice, we need to perform this task separately for each value of the independent (predicting) variable, and combine this with a statistical test. The details are too complex to discuss in this paper, but are documented online.[13]

One advantage of this method is that if the result is non-significant, we can increase the sensitivity of the test by expanding the subsample and reviewing more cases. But if we have a significant result we may simply stop.

To a fair degree of accuracy, many stative verbs can be identified lexically, and therefore we can enumerate an exclusion set. Indeed the optimum approach may be to combine top-down and bottom-up approaches, extrapolating from a subsample for just those verbs (e.g. *think*, *be*, *have*) which can be both stative and dynamic, and using corpus queries to identify the remainder.

In conclusion, although we have deliberately chosen a difficult research question, we have tried to show that researchers can attack it from both directions: top-down, to arrive at a plausible baseline for overall behaviour, and bottom-up, to identify alternates on a verb-by-verb basis.

6.2. Baselines are arbitrary

Many readers may accept the general argument summarised in this paper but retain a nagging doubt. The option to employ different baselines implies that there is no single 'objective' baseline (such as per million words) to compare a term against. Does this mean that the experimenter has a free hand in choosing a baseline? Indeed, there is a risk of an experimenter selecting a baseline in order to present results as significant.

To recap, our argument has been that baselines act as a *control*: in order to evaluate the rate or distribution of a choice, 'Type A', we must first attempt to measure the rate or distribution of the opportunity. The baseline for A will therefore typically depend on both:

1) the conceptualisation of the choice A vs. B (and thus the contrast 'Type B'), and
2) the ability to reliably retrieve the baseline in practice.

This formulation is far from arbitrary. Whereas the distribution of any given term may be contrasted with another, the point of an experiment is to test a plausible linguistic hypothesis. One could contrast nominalisations with nouns or with potentially 'nominalisable' verb phrases. But if we have a hypothesis that nominalisation is a *process* of replacing a verb phrase with a nominal equivalent, we need to test this hypothesis.

An awareness of the importance of baselines requires an increased care in experimental design and reporting conclusions, but it does not make the question of baselines arbitrary. Indeed, where baselines still included numbers of non-alternating 'Type C' cases, they would be open to further refinement. This flexibility in selecting baselines increases the range of potentially valid experiments that might be conducted. We have already discussed a familiar process of comparing multiple terms against the *same* baseline, as in Figure 3. It is also valid to compare the same term against *multiple* baselines.

[13] See http://corplingstats.wordpress.com/2014/04/10/imperfect-data.

present	LLC	ICE-GB	Total	present perfect goodness of fit
present non-perfect	33,131	32,114	65,245	$d^{\%}$ = -4.45 ± 5.13%
present perfect	2,696	2,488	5,184	' = 0.0227
TOTAL	35,827	34,602	70,429	^{2} = 2.68 ns
past				
other TPM VPs	18,201	14,293	32,494	$d^{\%}$ = +14.92 ± 5.47%
present perfect	2,696	2,488	5,184	' = 0.0694
TOTAL	20,897	16,781	37,678	^{2} = **25.06 s**

Table 5. Comparing present perfect cases against (upper) tensed, present-marked VPs, (lower) tensed, past-marked VPs (after Bowie *et al.* 2013).

Bowie *et al*. (2013) investigate the incidence of present perfect constructions (such as *I've done a bit of writing before*) over time. They show that these constructions appear to correlate with present-marked constructions (cf. *I am writing again*) more closely than with those marked for past tense (*I was writing before*). Note that these examples are not mutually substitutable alternates: the present perfect has a particular temporal aspect that is not easily expressed any other way.

Consequently, whereas the use of present perfect could correlate with either present or past tensed VPs, Bowie *et al*. found that the correlation was greatest with present tensed VPs. The simplest explanation of this data, therefore, was that the present perfect changed in proportion to what we might call the 'presentness of the text', i.e. how frequently the text referred to the present.[14] Given this correlation there was no need to hypothesise that we were witnessing a growth in the use of the present perfect as an alternative to other past tense forms.

Table 5 measures correlation using three factors:
the term itself (present perfect),
the baseline (present or past tensed VPs), and
the sociolinguistic contrast (LLC vs. ICE-GB).

This means that as well as comparing baselines against a term it is possible to evaluate terms and baselines over different contrasts.[15]

Far from being a weakness, recognition of the possibility of employing alternative baselines allows for their comparison and the identification of novel results. It also permits experimental improvement either by refining baselines or exploring the effect of different contrasts.

6.3. Multiple ecological pressures apply

The third and final objection concerns the fact that there may be multiple pressures on a particular term being used. Speakers make numerous choices, each influenced by multiple pressures, so to talk about a single 'choice' is misleading. As Smith and Leech (2013) put it: "it is commonplace in linguistics that there is no such thing as free variation" and that indeed multiple differing constraints apply to each term. On the basis of this observation they propose an 'ecological' approach, although this is not clearly defined. I want to argue that this objection may conflate two distinct arguments.

[14] A sharp-eyed reader may note the unconventional use of ^{2}. The goodness of fit ^{2} test is a contingency correlation test, where a small value (derived from ^{2} and scaled by the number of cases) indicates a near-flat distribution, i.e. a close match between term and baseline.

[15] The author has repeated this experiment over the following DCPSE contrasts: 2 time categories, 10 sampling categories of spoken data ('genre'), 280 individual texts, and ~460 subtexts. In each case the present perfect correlates more closely with present tensed VPs. See http://corplingstats.wordpress.com/2012/03/31/gof-measures

6.3.1. Multiplicity of lexical meaning

Firstly, words often have multiple meanings and associations, so exact semantic alternates represent an idealisation. With content or 'lexical' words, the idea of reliably identifying semantics-preserving alternates may be a pipe-dream where we have little choice but to employ more general baselines and survey results (witness our struggle in 6.1, and see also 2.3). On the other hand, grammatical words and constructions tend to present a relatively closed choice.

This argument is clearly correct. It is why in this paper we have discussed a *range* of approaches, from refining the baseline to limiting the choice to alternates. The example of phrasal verbs also demonstrates that a choice of an alternate form may have other implications for the speaker/writer (or entail further lexical-grammatical constraints).

Holding meaning constant in an absolute sense is rarely possible, although some grammatical alternations may operate as simple replacements in many cases (e.g. *the people living in France* vs. *the people who were living in France*). On the other hand, holding referential meaning constant, so that the choice consists of different ways of expressing the same concept, is frequently possible.

6.3.2. Multiplicity of pressures on choices

A second, related argument goes something like this: multiple pressures apply to every choice. Therefore the kind of narrow experimental research perspective we have outlined oversimplifies patterns found in rich, 'ecological' corpora.

This argument has more problematic implications.

In analysing natural language we expect that speakers have personal preferences, may adopt particular uses with genre and register, be affected by context and audience, etc. We do not require that at every single choice point the exact same influences, biases and constraints apply in the mind of the speaker. We cannot completely eliminate these constraints in each and every case.

However, *we are not attempting to explain why, precisely, a speaker chose to perform a particular utterance at a given point.* Rather, we are attempting to generalise across the entire set of such choices to identify statistically sound patterns, correlations and trends.[16] The critical question for a researcher may be formulated differently:

*Does one or more of these multiple constraints represent a **systematic** bias on the rate?*

If the answer to this question is yes, then it implies that these constraints should be detectable by experiment – provided we have sufficient data and pose the question correctly. (Consider the preceding discussion on how we might investigate the consequences of employing different baselines.) However if the answer is no, then these constraints do not pass a significance test and are considered background variation or 'noise'.

This problem is not unique to linguistics. Indeed it is true for all post-hoc analysis of real-world data, and it is precisely this issue that statistical estimates of error and significance are intended to address. Conversely, experimental 'lab' data which avoids such variation risks being insufficiently 'real world'! So if the ecological objection is understood in this second sense, it is really an objection to *all* experimental research. Moreover, using a word baseline would make matters worse, because variation of opportunity is also affected by this surrounding ecology (see 3.3).

[16] We may be being misled by terminology. The concept of 'free variation' we employ in this paper concerns grammatical and semantic **possibility** – the choice becomes available to the speaker (so they could hypothetically replace *who* with *whom* in every case). We may distinguish this concept from the mere possibility of utterance (speakers could, in theory, repeat the word *whom ad nauseam*) but we do not expect them plausibly to do so.

The fact that variation in outcomes may be due to multiple interacting factors in the environment does not invalidate models of choice. Incidentally, scientific ecology is full of such models. The familiar logistic S-curve, first developed in ecology, predicts the outcome over time of a process of selection from alternative outcomes (classically, natural selection of species in a niche), which, as we saw in section 1, is the most general definition of a choice model.

Given sufficient data, we may employ further distinctions in data to tease out a particular factor.[17] For example, Wallis (forthcoming) notes that successive decline in the probability of adding attributive adjective phrases to NPs (see 5.2), could be due to one or more of the following explanations: logical constraints (*tall* ¬*short*), semantic ordering constraints (size > colour >…) or principles of communicative economy. Further experiments may help distinguish these tentative explanations.[18] Gries (2011) dramatically improves classification accuracy in a syntactic priming experiment by allowing information about verb lemmas to be incorporated in his statistical model. By semantically classifying modals, Aarts *et al*. (2013) shows that the decline of first person *shall* against *will* is due to the decline of Epistemic *shall*.

One way the ecological objection might be addressed with multiple alternates is by refining the experiment linguistically and examining patterns of alternation in different contexts separately. This may also lead to the introduction of specific alternant forms. Thus the same semantic classification of modals in Aarts *et al.*could be applied to *BE going to* and *'ll*.

This objection does not undermine our argument. Rather, as we noted in section 2, it simply means that the condition of mutual substitution must be upheld for us to conclude that *replacement* is taking place. If alternate forms cannot be isolated, we are left with our broader point, namely that some baselines are more optimal than others, and that as researchers we have a duty to employ the best baseline we can to obtain the most reliable and robust results.

7. Conclusions

This paper is a call for researchers to pay attention to questions of choice and baselines. It is not necessary to assume that an observed change is due to a single source, even as we attempt to reduce a multitude of competing influences to a single testable alternation.

Whether we are aware of doing so or not, in corpus research we employ a baseline for the purposes of a *control*, i.e. that the null hypothesis is that the term we are interested in ('Type A') does not change in proportion to this baseline. Consequently, it is incumbent on researchers to attempt to minimise this variation. This is the minimum condition employed in the case of surveys of the type outlined in section 4.

In alternation research we employ an additional theoretical assumption, namely that this baseline represents the *opportunity* for the speaker to choose Type A. We wish to identify those occasions at which the speaker was 'on the threshold of choosing' A or B.

We have seen that the notion of alternation can be interpreted *strictly*, i.e. where mutual substitution must occur between all cases of Types A and B, or *generously*, where we risk allowing a small number of additional non-alternating 'Type C' terms into a baseline. A strict interpretation would

[17] Scientists talk about the 'art' of experimental design and the fact that it is easier to *retrospectively* agree that a particular experiment was a convincing demonstration of a now-agreed principle than it is to identify the optimum future experiment.

[18] For example the 'communicative economy' hypothesis indicated above implies that subsequent reference to the same referent in a text is likely to be shorter than the first, and that this explains the tendency for NP growth to fall. Theoretically at least, this might be tested by eliminating all subsequent references, although this will be time-consuming. However, this hypothesis would also tend to predict that in most cases subsequent references would be pronouns and not counted in any case!

require that the researcher check each case for the plausibility of their alternation. Statistical methods are relatively robust, but this does not mean we should not try to eliminate non-alternating terms as far as possible, to minimise the impact of unwanted variation and to improve the accuracy of significance testing. It should also be clear by now that this also means relegating word-based baselines to the sidelines.

We differ from Lavandera, who argues for ensuring that referential meaning remains constant across an alternation. We agree that this is desirable, but it is not always possible, and that constancy of referential meaning can be a relatively elusive concept in lexis and grammar. Thus we argue instead for the minimum condition that the choice is *possible*.

So both Lavandera and Aarts *et al*. would eliminate non-alternating formulaic cases of *shall* (e.g. *ye shall be saved*), but we would also accept that an investigation of choices may need to include cases where the chosen form could formally alter the meaning of the utterance. A cautious relaxation of a constant meaning constraint permits us to investigate variation in forms in the absence of a simple alternate (e.g. the present perfect) or the compound effect of sequential decisions to add attributive adjective phrases. If meaning is allowed to change, however, then such a change may need to factor in our experimental conclusions.

The ability to evaluate numerous different contrasts allows multiple baselines to be considered and permits novel experiments such as those described in section 5. The option to experiment with different baselines should be liberating, with the caveat that these baselines be *linguistically meaningful*. It follows that, in reporting experiments, alternates and baselines must be clearly stated, to enable reproducibility and comparability of results.

In order to differentiate putative explanations for a phenomenon it is perfectly reasonable to 'experiment with experimental design', i.e. to reformulate an experiment, redefine variables and queries (as with *shall / will*), and so forth. To this we can now add the possibility of refining baselines and comparing fitness measures over a given contrast with those obtained over individual texts. We hope that we have shown that focusing on alternation as an ideal represents an important starting point for innovation in corpus research methodology, allowing us to ask new questions of our data. We also believe that these methods allow corpus linguists to obtain evidence that is more commensurable with those obtained in other fields of linguistics (Schönefeld 2011), such as those concerned with the process of language production.

There is no such thing as an 'assumption free' science. The key question for any researcher posed by this paper is *how may we define an optimum baseline for a given research question?* This question is in the domain of linguistic argument – and here statisticians must defer.

Acknowledgements

This paper has been the result of many discussions with fellow linguists, not all of whom would agree with my arguments! Nonetheless, I offer it as a small contribution to what I believe to be an important debate.

My colleague Jill Bowie provided a great deal of insightful criticism, and the paper is *far* better as a result. The author would also like to thank Bas Aarts, Joanne Close, Gunther Kaltenböck, Geoff Leech, Gerald Nelson, Gabriel Ozón and Nicholas Smith. Research on the Verb Phrase in English was supported by AHRC AH/E006299/1. Additional statistical notes can be found on the author's blog **corp.ling.stats**.

References

ACLW: Aarts, B., J. Close, G. Leech and S.A. Wallis (eds.) 2013. *The Verb Phrase in English: Investigating recent language change with corpora*. Cambridge: CUP. Preview at www.ucl.ac.uk/english-usage/projects/verb-phrase/book.

Aarts, B., J. Close and S.A. Wallis, 2013. Choices over time: methodological issues in investigating current change. ACLW Chapter 2.

Bauer, L. 1994. *Watching English Change: An Introduction to the Study of Linguistic Change in Standard Englishes in the Twentieth Century*. London: Longman.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan, 1999. *The Longman grammar of spoken and written English*. London: Longman.

Biber, D. and B. Gray, 2013. Nominalizing the verb phrase in academic science writing. ACLW Chapter 5.

Bowie, J., S.A. Wallis and B. Aarts, 2013. The perfect in spoken English. ACLW Chapter 13.

Bowie, J., S.A. Wallis and B. Aarts, 2014. Contemporary change in modal usage in spoken British English: mapping the impact of 'genre'. In Marín Arrese, J.I. and J. Van der Auwer (eds.). *Current issues on Evidentiality and Modality in English*. Berlin: Mouton de Gruyter.

Gardner, D. and M. Davies, 2007. Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly*, 41(2), 339-360.

Gries, S. Th. 2009. *Quantitative Corpus Linguistics with R*. New York: Routledge.

Gries, S. Th. 2011. Studying syntactic priming in corpora. In Schönefeld, D. (ed.) *Converging Evidence*. Amsterdam: John Benjamins.

Huddleston, R. and G. Pullum (eds.) 2002. *The Cambridge Grammar of the English Language*. Cambridge: CUP.

Labov, W. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Lavandera, B.R. 1978. Where does the sociolinguistic variable stop? *Language in Society* **7**: 171-182.

Nelson, G., S.A. Wallis and B. Aarts, 2002. *Exploring Natural Language*. Amsterdam: John Benjamins.

Schönefeld, D. (ed.) 2011. *Converging Evidence. Methodological and theoretical issues for linguistic research*. Amsterdam: John Benjamins.

Smith, N. and G. Leech, 2013. Verb structures in twentieth century British English. ACLW Chapter 4.

Smitterberg, E. 2005. *The progressive in 19th Century English: a process of integration.* Amsterdam: Rodopi.

Wallis, S.A. 2003. Scientific experiments in parsed corpora: an overview. In Granger S. & Petch-Tyson, S. (ed.) *Extending the scope of corpus-based research: new applications, new challenges*, Language and Computers 48. Rodopi: Amsterdam. 12-23.

Wallis, S.A. 2013. z-*squared: the origin and use of* ^{2}. *Journal of Quantitative Linguistics*. 20:4, 350-378.

Wallis, S.A. forthcoming. *Capturing linguistic interaction in a grammar: a method for empirically evaluating the grammar of a parsed corpus*. prepublished: www.ucl.ac.uk/english-usage/staff/sean/resources/analysing-grammatical-interaction.pdf