Dominic Stewart

The Language Iceberg

Dictionaries and the Frequency of Inflectional Forms

Abstract

One of the many important lessons imparted by corpus linguistics is that the information supplied in dictionaries and grammars represents no more than the tip of the iceberg. In lexicography, the object of description is fundamentally the lemma. When more specific details are not supplied, the assumption on the part of the average dictionary user is in all probability that the various forms of any given lemma show not only similar frequency and lexical environment, but also similar meaning. It would appear that inflectional forms are under-represented in dictionaries.

In the literature very little attention has been devoted to inflectional forms, and in particular to their raw frequency. The frequency rates of the forms of a single lemma can not only differ markedly from each other, but can also prove to be far higher or lower than the average for single inflectional forms, something which has important implications for language learners and which could therefore claim more emphasis in language-learning materials.

The main focus in this paper will be on the frequency of inflectional verb forms as represented in dictionaries, including a final case study of the frequency of such forms within some idiomatic expressions. The question lying at the heart of this work is whether the raw frequency of inflectional forms should occupy a more important position in language-learning materials. However this may be, the counting of inflectional forms can reveal new and surprising insights into the vast and mostly untamed wilderness of language.

Keywords: *learner's dictionaries, lexicography, headwords, lemmas, frequency of use, inflectional verb forms*

1. Introduction

Over the last thirty years or so, British learner's dictionaries have progressed in leaps and bounds, proving to be of immeasurable benefit to students of English not only in Europe but also around the world. If it was once true that the sun never set on the British Empire, the same could now be said of British lexicography.

However, one of the many important lessons imparted by corpus linguistics is that the information supplied in dictionaries (and grammars) represents no more than the tip of the iceberg. This is perhaps inevitable, but the tip which is visible to the user can be almost too seductive, a kind of enforced *trompe l'oeil*. Just as an attractive, well-organised department store will be contingent upon backrooms or basements overflowing with a disarrangement of products and tools, so too an appealing, well-organised dictionary must spring from the maverick waywardness of language.

One of the most telling examples of the submerged part of this iceberg is the fact that inflectional forms are under-represented in dictionaries. In lexicography, the basic unit of description is the lemma. When more specific details are not supplied, the unconscious assumption on the part of the average dictionary user is probably that the various inflectional forms of, for instance, the lemma *SEARCH*[1] (*search, searches, searching, searched*) show not only similar frequency and lexical environment, but also, by consequence or by extension, similar meaning. As Sinclair, Jones and Daley (2004, 4) point out:

> The hypothesis implicit in dictionary and thesaurus organization is that a changing grammatical role does not affect the semantic value of a word, except where this is explicitly stated [...] Looked at lexically, in terms of the statistics of word occurrence, this hypothesis cannot be substantiated on present evidence, which instead suggests that there is such variation in the homogeneity of grammatical variants that some more complex hypothesis will have to be put forward when fuller information is available.

Along similar lines, Tognini-Bonelli (2001, 92) writes:

> The fact that lemma and inflected forms are bound to share the same meaning and differ only in their grammatical profile—the lexical profile is not usually considered relevant—is one of those apparently inoffensive assumptions on which most of our reference works are based: we look up a verb in a dictionary under the base form, or an Italian adjective under the masculine singular. An association that has a certain value in terms of convenience (a dictionary entry, for example) should not be taken for granted and left totally unquestioned.

In the same way, Knowles and Don (2004, 71) observe that in corpus linguistics "it has become apparent that individual members of the lemma can behave independently and develop their own meanings and collocations."

The focus of these observations is on meaning and lexical profiles, but the *frequencies* of inflectional forms can be just as important. The frequency rates of the forms of a given lemma can not only differ markedly from each other, but can also prove to be far higher or lower than the average for individual inflectional forms in general. The elaborate language analyses enabled by corpus linguistics allow a sharper focus on the frequency and function of inflectional

[1] Hereafter capital letters will be used when the reference is to the lemma.

forms, something which is of course important for language learners and which could therefore claim more emphasis in learner's dictionaries.

I shall begin with a review of works dealing with the inflectional forms of nouns and verbs, and then turn my attention to raw frequency counts of such forms and their relationship with function. The primary focus will be on verbs, in particular the differing functions of inflectional verb forms as represented in dictionaries, with some observations concerning the way dictionaries prioritise certain functions and ignore others. The final section provides a case study of the frequency of inflectional verb forms within idiomatic expressions.

The question lying at the heart of this paper is whether the raw frequency of inflectional forms should occupy a more important position in language-learning materials. However this may be, the counting of inflectional forms can reveal new and surprising insights into the vast and mostly untamed wilderness of language.

Unless otherwise stated, the corpus adopted here, both for reference and for examples, is the *British Web 2007*, also known as *ukWaC*, a web-derived corpus containing over 1 billion 300 million words from websites within the .uk domain. Consulted using *The Sketch Engine* (Kilgarriff et al. 2014), it is a general-purpose corpus with a broad range of text types.

2. Inflectional variation: nouns

Sinclair has observed that "in English *enormous* can be used of both pleasant and unpleasant things," but that "*enormity* is restricted to crimes, scandals and heavy burdens" (Sinclair and Carter 2004, 150). Here the differing discourse prosodies are striking, but since *enormous* and *enormity* belong to different grammatical classes (adjective and noun) they can be—and indeed are—assigned separate entries in all the major monolingual dictionaries, with the result that the differing prosodies and lexical environments are captured reasonably well. Consider the respective entries in the *Longman Dictionary of Contemporary English*:

> *enormous*
> very big in size or in amount SYN huge
> an enormous bunch of flowers
> an enormous amount of money
> The team made an enormous effort.
>
> *enormity*
> 1 [singular] the great size, seriousness, or difficulty of a situation, problem, event etc.
> enormity of
> Even now, the full enormity of his crimes has not been exposed.
> the enormity of the task
> 2 [countable usually plural] formal a very evil and cruel act SYN atrocity

However, when the focus is on the inflectional forms of nouns, it is understandably the singular-plural dichotomy that has a monopoly. It is common knowledge, for example, that certain nouns occur prevalently in the singular—most obviously uncountable nouns such as *fun, progress, fame* and *mirth*—whereas others mostly occur in the plural, such as *proceedings* and *regulations*. It is also common knowledge that on a lexical and semantic level some nouns operate very differently from singular to plural, e.g., *damage* vs *damages*, *hair* vs *hairs*. This is particularly true of things that come in pairs, for example parts of the body such as *arms, legs, ears, eyes*. Sinclair writes (Sinclair and Carter 2004, 30-31):

> This pairing cuts across the regular relationship of singular and plural in nouns. Normally we can expect the plural of a noun to refer to more than one of whatever the singular refers to, but with pairs the singular is not as often required as the plural. It is therefore available for other functions.

Many of these "other functions" of the singular form connect with fixed expressions—a good percentage of which are figurative—such as *keep an eye on, have a good ear for* (*music, languages*), *on the back foot, get off on the wrong foot*. Of course the respective plural forms can be part of figurative expressions too (*be up to one's ears, find one's feet, get cold feet*), but according to Sinclair's investigations the plural forms denote the actual part of the body far more frequently than the singular. For example, the adjectives *blue* and *brown* collocate with *eyes* rather than *eye*, whereas *eye* habitually occurs in expressions to do with visualising and evaluating. Having said that, the singular-plural dichotomy in the above instances is reasonably well-catered for in dictionaries, whether implicitly or explicitly.

Other singular-plural oppositions, however, are handled in more implicit fashion by lexicographers. Zhang (2013, 39-42) notes that the respective lexical environments of both *disadvantage* vs *disadvantages* and of *opportunity* vs *opportunities* show some significant differences. We are informed that (i) *outweigh* and *possible* recur significantly with *disadvantages* but much less so with *disadvantage*, and that (ii) *opportunity* has significant recurrence with *golden* and *great* (connecting with "the evaluation of an opportunity"), whereas *opportunities* is associated much more frequently with job/career prospects. These differences are confirmed in the *British Web 2007* but only implied in dictionaries. Under the entry *disadvantage* in the main British dictionaries there is no suggestion of differing lexical environments between singular and plural, while under *opportunity* the *Oxford Advanced Learner's Dictionary* devotes a separate line to 'career/employment/job opportunities,' though these come across as mere examples of *opportunity*, i.e., it is not clarified that the presence of the plural in these collocations is much more recurrent than the presence of the singular. See also Doyle (cited in Hoey, 2005, 8) on *frequency* vs *frequencies*, Esser (2000, 97) on *tree* vs *trees*, and Sinclair, Jones and Daley (2004, 188-193) on *word* vs *words* and *year* vs *years*.

3. Inflectional variation: verbs

Esser (2000, 98) observes that verbs are of primary interest in terms of the behaviour and meaning of inflectional forms because "tenses and aspects may have a restricting influence on the verb senses." It is also true that verbs tend to have a greater number of inflectional forms than nouns and adjectives, with the result that the potential for semantic and lexicogrammatical variation is greater.

Esser examines the verb *SPEED*, claiming that whereas the forms *speed* and *speeding* can both signify (i) 'move quickly' and (ii) 'go too fast,' the forms *speeds*, *speeded* and *sped* convey only the meaning 'move quickly.' Stubbs (2009, 120) analyses the inflectional forms of the verb *SEEK*, discovering in the 20-million word corpus adopted that they can have markedly different collocates. Focusing upon the twenty most frequent collocates of each form, he finds that *seek, seeking* and *sought* all share the collocates *asylum, court, government, help, political, support*, whereas the forms *seeks* and *seek* share only one collocate, namely *professional.* Stubbs also stresses the absence of shared collocates between the pairs *seeks*/*sought* and *seeks*/*seeking*, mostly because *seeks* is frequent in lonely hearts ads, where its most recurrent collocates include *attractive, black, caring, female, guy, lady, male, man, professional*. Similarly, Stubbs (1996, 172-173) also considers the verb *EDUCATE*, highlighting that the form *educate* collocates primarily with semi-synonyms such as *enlighten, entertain, help, inform, train*, while *educated* recurs with *at*—most frequently in the phrase *he was educated at*—followed by *school, university* and *college*, in addition to a range of prestigious institutional names, including Cambridge, Charterhouse, Eton, Harrow, Harvard, Oxford, Yale.

Research conducted by O'Halloran is in much the same vein (2007), but puts the spotlight on the differing levels of concreteness/figurativeness across inflectional forms. The author's main example is a comparison, in the hard news register, of *eruption*/*s*, *erupt*/*s* and *erupted*. His corpus investigations reveal that *eruption* is much more likely to carry meanings associated with volcanoes, while *erupt*/*s* and especially *erupted* have "a semantic preference for human phenomena, rather than for volcanoes, and carry a negative register prosody" (O'Halloran 2007, Section 5). Along the same lines, Zhang (2013, 36-39) underlines that in the *Bank of English* corpus the inflectional form *flaring* tends to be associated with concrete situations (*flaring nostrils/eyes/gas*), whereas *flared* is usually more abstract and characterised by an unfavourable discourse prosody, its top co-occurrences being *trouble, violence* and *tempers*. See also the analogous observations made by Tognini-Bonelli (2001, 95) concerning the forms *facing* and *faced*. The former has a greater propensity for association with words relating to physical position (*stood, sat, sitting*), whereas *faced* is more abstract, collocating chiefly with words denoting problems or difficulties.

These observations, like those relating to nouns in the previous section, focus on various aspects, such as collocates, denotational meaning, discourse prosody and a cline of concrete to figurative, helping to expose the shortcomings of the incautious supposition that inflectional forms of a single lemma share the same characteristics.

3.1 Raw frequencies of inflectional verb forms

Just as the unsuspecting dictionary user might assume—if further information is not forthcoming—that inflectional forms have similar lexical environments, similar ratios of concrete vs abstract uses, and similar meanings, it might also be assumed that the various forms have similar frequencies, whether it be their raw frequencies, or the recurrence of the tense/aspect/voice that the forms represent.

I shall begin with some examples of raw frequency counts of inflectional verb forms, based on random concordance samples from the *British Web 2007*.[2] However, first of all it is helpful to be aware of the average frequency rates for inflectional verb forms in English, or at least in British English. The statistics provided in the table below are based on Leech, Rayson and Wilson (2014), who provide the frequency counts of inflectional forms extracted from the *British National Corpus*. In order to form an idea of the overall average rates of recurrence of inflectional verb forms, I selected all the verbs beginning with the letter 's' listed by the authors (though their list does not include relatively infrequent verbs), and then calculated the respective percentages for each verb. The averages supplied in the table below are thus partial and approximate, but they provide some parameters with which to assess the outcomes supplied during the course of this article. Clearly it is controversial to establish an average recurrence of forms in the *-ed* category, since while most verbs have the same realisation for simple past tense and past participle (*managed, sent, swung*), many others show distinct forms (*swore, sworn*; *sang, sung*). This means that for the sake of convenience, past tenses and past participles, even

[2] If the research proposed here had been limited to raw frequency alone, then the calculations could have been based simply on the overall percentages in the corpus of each inflectional form, but since I also pause to consider the function of the different forms (in sections 3.3-3.5) it seemed better to have manageable samples for close analysis.

when their morphology is different (*spoke, spoken*), are both categorised as *-ed* forms. Other verbs whose paradigms seriously hamper frequency counts because they have so few inflectional forms, such as *set* and *shut*, are excluded.

base form	**form with *-ing***	**form with *-s/-es***	**form with *-ed***
33%	14%	5%	48%

Tab. 1: Percentages of inflectional verb forms

Now that we have an approximate idea of the average frequencies of inflectional verb forms, I shall calculate as an initial example the raw frequencies of the inflectional forms of the verbs *ACCLAIM* (1.2 occurrences per million tokens in the corpus) and *PERMEATE* (2.1 occurrences per million tokens). The choice of these two verbs is fairly random, but for the sake of a balanced comparison it was important that their overall frequency in the corpus did not differ too radically (by way of comparison, the verb *REPLY* has a strike rate of 35.1 per million, and the verb *BUY* 207.3 per million), and that both verbs had identical forms for past tense and past participle. Here are the percentages:

acclaim	*acclaiming*	*acclaims*	*acclaimed*
15%	3%	2%	80%
permeate	*permeating*	*permeates*	*permeated*
28%	11%	29%	32%

Tab. 2: Percentages of inflectional forms of the verbs *acclaim* and *permeate*

A glance at these figures suffices to highlight that there are enormous differences between the two verbs: *ACCLAIM* is dominated by the form ending in *-ed*, with the result that (i) the base form has a lower rate than average and (ii) that there are few *-ing* or *-s*/*-es* forms (hereafter simply *-s* forms), whereas the percentages of *PERMEATE* are more evenly distributed. Particularly worthy of note, however, is the frequency of *permeates*, since as stated above, the average percentage of *-s* forms for verbs is approximately 5%. Naturally such variations in raw frequency are explainable up to a point in functional terms; *ACCLAIM* has a much higher percentage than average of occurrences in the passive, with the result that *-ed* is attested very frequently, and *PERMEATE*, since it is relatively unlikely to be governed by animate grammatical subjects (and therefore first- and second-person grammatical subjects), is predominantly adopted with a third-person subject, whichever the tense or aspect, and this would in part account for the unusually high rate of *permeates*. Having said that, other verbs without this animate/inanimate constraint can also feature abnormally high recurrence for the form ending in *-s*, for example *resembles* occupies 32% of *RESEMBLE*, and is thus substantially higher than the 5% average for *-s* forms in general:

resemble	*resembling*	*resembles*	*resembled*
37%	19%	32%	12%

Tab. 3: Percentages of inflectional forms of the verb *resemble*

Also conspicuous here is the very low rate of *-ed* forms, in part due to the fact that *resemble* is very rarely adopted in passive structures.

Although these raw figures are assuredly of interest to the linguist, their real significance is questionable, and this is because inflectional forms have multiple functions. The fact that *acclaimed* accounts for something like 80% of occurrences of the verb *ACCLAIM* is in itself of no great assistance to learners; far more revealing is the substantial percentage of passive structures. In the same way, higher than average occurrences of the base form of a verb (for example *guess* accounts for over two-thirds of the lexeme *GUESS*) are meaningful only to a degree, in that the base form has several functions, i.e., present tense aside from the third-person singular; the imperative; part of the *will*-future and conditional (*will/would like*); part of the interrogative and negative of the simple past and present (*did she like?, she didn't like, does she like?, she doesn't like*)—see Halliday and James (1993, 48). Having said that, some raw frequency counts can be eye-opening, as will be exemplified in the next section.

3.2 Regulations *vs* resembles

For the purposes of comparison I shall return momentarily to the inflectional forms of a noun. In the *Oxford Advanced Learner's Dictionary*, the reader is informed that the noun *regulation* is 'usually plural.' This indication must in the first instance stem from a raw frequency count of the inflectional forms *regulation* and *regulations*. In the *British Web 2007*, the form *regulation* has a recurrence of 42.38 per million, while the form *regulations* has a recurrence of 78.56 per million, and is therefore almost twice as frequent. For the corpus analyst, this search is refreshingly uncomplicated, firstly because both *regulation* and *regulations* are always nouns, so there is no possibility of tagging problems in the corpus related to overlaps with other parts of speech (see *position* and *positions*, which both function as noun and verb). Secondly, the form without the *-s* ending—*regulation*—is always singular, and the form with the *-s* ending—*regulations*—is always plural (as is well-known, this is not always the case: nouns ending in *-s* can be singular, e.g., *the news is good; an important means of communication*, and nouns not ending in *-s* can be plural, e.g., *she weighs 8 stone; two geese; a hundred sheep*).

For the lexicographer, therefore, the situation is in this instance uncomplicated, enabling a simple conversion from form to function: *regulation* is always a noun and is always singular, and *regulations* is always a noun and is always plural. Further, the corpus statistics speak for themselves: the singular vs plural ratio here is approximately 1:2, while the average ratio of this dichotomy in English has been calculated at 8:1, i.e, the singular is as a rule 8 times more frequent than the plural (see Halliday and James, 1993). For all these reasons, the lexicographical flag 'usually plural' assigned to the lemma *REGULATION* would appear to be unchallengeable.

Let us now turn to the inflectional verb form *resembles*, which in the *British Web 2007* has a recurrence of 3.3 per million tokens. As noted above, *resembles* occupies almost a third of the total number of occurrences of the lemma *RESEMBLE*. This strike rate is abnormally high, since—again as mentioned above—the average percentage of *-s* forms for verbs is approximately 5%. Therefore *resembles* is proportionally over six times more frequent than the average for *-s* forms. Nevertheless, this is not flagged in any of the major learner's dictionaries.

Why is this? After all, as in the case of *regulation/s*, the figures are striking and seem incontrovertible. And as in the case of *regulation/s*, the form *resembles* corresponds exclusively to a single function, in this instance the third-person singular of the present simple. Of course one might wish to argue that verbal *-s* forms are functionally less circumscribed than they seem, inasmuch as they can refer to future time ('Anne's flight leaves tomorrow at 12'), or even past time, above all in humorous anecdotes ('so my brother goes into this shop in Barcelona and, incredibly, the assistant recognises him' [my examples]), but this does not seem very relevant. It would be like arguing that *regulations* is functionally less circumscribed than it seems because it can refer to two, three, ten or a hundred regulations. However this may be, the fact remains that a flag such as (i) 'often with *-s*' or (ii) 'often present-tense third-person singular' is never provided for any verb. Would it be so difficult for lexicographers to include a flag of this nature?

Certainly with regard to a formal (rather than functional) flag of the type 'often with *-s,*' the answer is yes, it would indeed be difficult. The form *resembles* has the advantage of being virtually a mirror image of the function third-person singular present simple, but more generally what could a lexicographer usefully convey to the user by highlighting the raw frequency of other inflectional forms? By highlighting that, for example, *ACCLAIM* is ‘often with -ed,’ or that *TEEM* is ‘often with -ing,’ or that *GUESS* is ‘often base form’? As pointed out in 3.1 above, these inflectional forms have a host of different functions, with the result that in general this type of formal indication would be virtually meaningless.

Having acknowledged this, there could still be a case for arguing that a functional flag such as ‘often present simple third-person singular,’ perhaps abbreviated to ‘oft pres simp 3rd sing,’ would be useful with respect to the lemma *RESEMBLE*, along the lines of ‘usually plural’ for the lemma *REGULATION*. Certainly this information would be of benefit to the learner, but the risk is clearly that of opening the floodgates. The proposed indication, though it seems innocuous enough, actually entails four different functions: present (tense), simple (aspect), third (person) and singular (number). Now whereas it seems feasible for lexicographers to underline that a given verb is adopted, for instance, primarily in the past tense, or primarily in the third person, and whereas dictionaries do already include—as will be discussed later in 3.3 and 3.4 below—information regarding unusually high or low rates of aspectual features, there is clearly a danger of overcomplicating the issue, and of overloading the user with information, if there are repeated elaborate flags of the kind ‘oft past simple first sing pass.’

Further, the fact that inflectional verb forms can have a multiplicity of functions makes the relevant corpus data harder to collect for the lexicographer. For example, the form *guessed* can connect to the past simple, present perfect, past perfect, future perfect, active, passive, first person, second person, third person, singular, plural, with the result that the extraction of the data necessary for the types of flag discussed above is a mammoth task. For the lexicographer, inflectional forms of verbs bring an extra level of complication by comparison with inflectional forms of nouns and other word classes (though this is by no means to suggest that the transition of nouns from form to function is obstacle-free; one need only reflect upon the complex issue of countable and uncountable nouns, in particular the prodigious number of singular nouns with both countable and uncountable function (e.g., *controversy*), or with a singular form which is not countable (*that was a big help* but not **those were two big helps*). The systematic introduction of elaborate functional formula in the dictionary would require infinite corpus research, transporting lexicographers into the stormy seas of multifunctionality.

Having said that, it is well-known that dictionaries do supply information of a functional nature, but such information appears to be restricted to only a handful of specific functions. This point will be taken up in the next section.

3.3 Differing functions of inflectional verb forms as represented in dictionaries

Perhaps the most immediate question arising from the previous paragraph is the following: given that dictionaries do not flag anomalously high occurrences of inflectional verb forms, even when they correspond unequivocally to a single function (*resembles*), then do they at least flag high frequencies of specific verb functions? Interestingly, aside from the dichotomy transitive vs intransitive, it is predominantly the passive and the progressive functions which are prioritised (see Stewart 2020, 4-10; 2021, 69-84). For example, the verbs *entitle* and *acclaim* are marked as usually/often passive in the major learner's dictionaries, while *kid* and *slim* are generally marked as usually/often progressive. Yet hardly any other verb functions are labelled, and in any case on a much smaller scale, for example the imperative. Also conspicuous is that the passive and progressive are occasionally cited when they are not a feature of the verb in question. Once again with reference to the lemma *RESEMBLE*, consider the following:

Labels 'no passive' and 'no progressive' assigned to *RESEMBLE* in four British learner's dictionaries

Oxford Advanced Learner's	No passive	No progressive
Macmillan Dictionary	No passive	-
Collins Reverso	-	No progressive
Longman Dic of Contemporary English	-	No progressive

In the case in point, this is undoubtedly valuable information for learners of English, but it is not clear why generally speaking it is the passive and the progressive functions that steal the limelight at the expense of other worthy candidates. For instance, the verb *TEND* shows an atypically high rate of frequency for the simple present tense, *EXCLAIM* for the simple past tense, *PLEDGE* for perfective forms, *HOPE* for the first-person singular (all tenses but particularly the present simple) and *SOUND* for the third-person singular (all tenses), not to mention future and conditional forms. Yet none of these gets a mention in dictionaries.

3.4 Prioritisation of the passive and progressive in dictionaries

Why is it that passive/not passive and progressive/not progressive are routinely marked up by lexicographers, whereas other verb functions are for the most part absent? It seems odd that lexicographers should take pains to inform the user that *RESEMBLE* is adopted sparingly in the passive and progressive (though the flag 'not passive' is often absent when the data would suggest that it should be present, for example the verb *DODGE*—see Section 4 below—is hardly ever passive but this is not highlighted in dictionaries), but omit to point out not only that the present simple third-person singular of *RESEMBLE* is extraordinarily frequent, but also that this verb's third-person forms in general (singular/plural, all tenses) have practically a monopoly, even if one excludes *-ing* forms that introduce relative clauses ("these companies send out letters resembling invoices") from the calculations.

Could it be argued that there is something more cardinal, more elemental, about the categories of voice and aspect than the categories of person and tense? It seems doubtful, especially as the perfective aspect is another category that is never labelled: for example, as well as *PLEDGE,* the verbs *ELUDE* and *CAMPAIGN* show notably high percentages for perfective forms, whereas *DEPEND, DESERVE* and *CONTAIN* have low percentages (see Stewart 2021, 57-59), but dictionaries make no allusion to such frequencies.

Or is there something less intuitively obvious, and therefore more useful to the learner, about the frequency of passive and progressive by comparison with other verb functions? Is the fact that *ACCLAIM* is often passive less accessible to the learner's introspection than (i) the fact that *ELUDE* has an unusually high recurrence of perfective forms, or than (ii) the fact that *PERMEATE* has practically a monopoly of third-person grammatical subjects? Again this seems improbable. Even if the native speaker of English might be able to work out introspectively that *PERMEATE* is less likely to be governed by first- or second-person subjects, this may not be apparent to learners at all. And regarding perfectives, I have studied and taken a keen interest in Italian for around 40 years, but until 12 months ago it had never occurred to me that perfective forms occupy a staggering 56% of the total number of occurrences of the verb *DEMERITARE* (as far as I can make out, Italian verbs have an average strike rate of between 5% and 10% for perfective forms). In my view, the hypothesis that the frequency of passive and progressive is less accessible to introspection—and therefore deserves more attention—with respect to other verb categories is decidedly weak.

Further, it is both curious and ironic that priority is afforded in dictionaries to two verb functions—passive and progressive—that (i) elude univocal definitions, and (ii) are harder to identify in a corpus than many other verb functions. Stewart (2020, 5-11) discusses the difficulty inherent in pinpointing exactly which structures qualify as passive and progressive. I shall not reproduce here the argumentation supplied; suffice it to say that, in part because definitions of these two categories differ from one dictionary to another, dictionary users may not have a crystal-clear idea about what 'usually passive' or 'usually progressive' actually means, and therefore about what type of language use these labels encompass. Further, even if the individual lexicographer succeeds in establishing firm ground-rules about which structures can be classified as passives/progressives and which cannot, in a corpus it can prove laborious to identify them, partly because, as alluded to in section 3.2, both the *-ing* and *-ed* forms have multiple functions, but also because the automatic tagging of a corpus can leave a lot to be desired. For example, the form *relaxing* in the collocation *relaxing music* is clearly adjectival, but in one third of the 158 occurrences of this collocation in the *British Web 2007*, *relaxing* is tagged as a verb, as are almost all the occurrences of the form *relaxed*, again manifestly adjectival, in the collocation *relaxed atmosphere*.

For all these reasons, it seems an arduous task to assign passive/progressive labels to verb lemmas. In terms of blood, sweat and tears, there is simply no contest with the verbal *-s* form, which has a single function, which is different from any other form, and therefore whose rate of frequency can be determined in a flash. In short, one struggles to find intuitive reasons to explain the lexicographer's prioritisation of passive and progressive over other verb functions.

3.5 Different senses of a headword

A further complication for the lexicographer in regard of the discussion above is that the different senses of a headword are bound to vary—sometimes significantly—in terms of the frequency, distribution and lexical environment of inflectional forms. An obvious example is the noun *DAMAGE*, whose singular and plural have distinct meanings, for example the *Oxford Advanced Learner's Dictionary* defines the singular in terms of physical or psychological harm, and the plural as "an amount of money that a court decides should be paid to somebody by the person, company, etc. that has caused them harm or injury." And this is reflected in the top collocates of the singular and plural:

- *damage* as noun: cause, loss, brain, suffered, criminal, repair, accidental, done, property, prevent, environmental
- *damages* as noun: consequential, liquidated, punitive, liable, losses, recover, arising, awarded, incidental, claim

If we turn again to verbs, and in particular to the progressive and passive, we naturally find comparable examples. The progressive forms of the verb *DIE* are frequently associated with the meaning 'long for' ("Any books out there you are dying to get your hands on?"), whereas the non-progressive forms much less so (Sinclair 1966:419), and the verb *PICTURE* occurs far more often in the passive when its meaning corresponds to 'show in a photograph' ("This lake is pictured on so many calendars, chocolate boxes and jigsaw puzzles that it's a familiar sight to many long before they actually see it for themselves") than when its meaning corresponds either to 'describe' ("John is picturing for us the relationship of the Church to Christ") or to 'form a mental image of' ("and while I was busy picturing his past life to myself, he had bowed me out of the room").

Dictionaries cater for such instances up to a point, i.e., singular/plural/passive/progressive with reference to the different meanings of a headword, but once again it is manifest that special attention is reserved for these functions at the expense of most others. Yet it might be important to emphasise, for instance, that *PICTURE* as verb occurs only sporadically in the first-person singular when its meaning corresponds to 'show in a photograph' ("I'm pictured here with my host family at the Rockerfeller Center, NY"), and that it occurs much more recurrently in the first-person singular when the meaning is 'form a mental image of' ("I can still picture him kicking conkers down a country lane").

4. Frequency of inflectional verb forms within idiomatic expressions

As a rule, dictionaries barely cater for the frequency and function of inflectional forms within idiomatic expressions. As an illustration of this, let us firstly consider the frequency of the forms of the verb *DODGE* (3341 occurrences at a rate of 2.16 per million) in the *British Web 2007*:

dodge	*dodging*	*dodges*	*dodged*
39%	40%	5%	16%

Tab. 4: Percentages of inflectional forms of the verb *dodge*

The 5% strike rate for *dodges* is normal for verbal *-s* forms, but it is noticeable that the percentage is much higher than average for *-ing*, and much lower than average for *-ed*. It is hard to account for this. Intuitively one might predict a recurrence of structures in the corpus involving *dodging* tagged as verb but in reality with noun function (*tax dodging, fare dodging* etc.), but their presence in the corpus is negligible. If we now enter the query '*DODGE* + *BULLET*' within a span of 5 (thus capturing *we dodged a bullet, dodging bullets, dodge their bullets, a bullet was dodged* etc., most of them with metaphorical meaning), which retrieves 139 occurrences at a strike rate of 0.09 per million, there is an interesting difference:

dodge+bullet	*dodging+bullet*	*dodges+bullet*	*dodged+bullet*
39%	39%	< 1%	21%

Tab. 5: Percentages of inflectional forms of the verb *dodge* + *bullet* within a span of 5

Whereas the two forms *dodge* and *dodging* remain the most frequent, showing very similar rates of frequency not only to each other but also by comparison with *DODGE* as a whole, *dodges* is barely attested. Indeed of the total number of 139 occurrences of '*DODGE + BULLET*' within a span of 5, only 1 corresponds to *dodges* and thus the percentage is only marginally above zero. This is a telling difference by comparison with the 5% for *dodges* shown in the previous table. Of course it might be objected that in the case of this latter search, the corpus is too small for meaningful results, but in the massive *English Web 2020* corpus, '*dodges* + *BULLET*' is 20 times less recurrent than *dodges* in all contexts, so there is definitely a recognisable pattern. Aside from transitivity/intransitivity, no frequency flags are supplied in dictionaries for *DODGE* or for '*DODGE* + *BULLET*.' However, in the *Oxford Lexico UK* online, it is worth noting the curious fact that 9 of the 11 examples provided (thus over 80%) of '*DODGE* + *BULLET*' include the form *dodged* which, as recorded above, occupies a much lower-than-average 21% of the inflectional forms of this collocation (This type of discrepancy is not uncommon in dictionaries, for example the verb *TEEM* has a strike rate for *teeming* which approaches 75%, but no *-ing* form shows up in the five examples supplied under the entry *TEEM* in the *Longman Dictionary of Contemporary English*).

As a further example, let us examine forms of the verb *STICK* (almost 67,000 occurrences in the *British Web 2007* at a rate of 43.15 per million), first of all focusing upon their recurrence unspecified for co-text, and subsequently upon their recurrence in idioms.

stick	*sticking*	*sticks*	*stuck*
35%	19%	6%	40%

Tab. 6: Percentages of inflectional forms of the verb *stick*

These figures are are fairly unremarkable in the sense that they correspond more or less to the average percentages for inflectional forms of verbs. If we now consider the search '*STICK to * guns*,' where the asterisk stands for 'any word' (387 results at a rate of 0.25 per million, including structures of the type *she stuck to her guns, they're sticking to their guns*, though the query does not capture a handful of relevant corpus instances such as "I admire you for sticking to your ideological guns," because the asterisk in the query retrieves just one word), we discover outcomes not dissimilar to *STICK* as verb unspecified for context, though there is a higher rate of *-s* forms:

*stick to * guns*	*sticking to * guns*	*sticks to * guns*	*stuck to * guns*
31%	18%	9%	42%

Tab. 7: Percentages of inflectional forms of the verb phrase *stick to * guns*

However, turning to the scores for '*STICK * NECK out*' (306 results at a rate of 0.2 per million: outcomes include structures of the type *they stuck their necks out, I'm going to stick my neck out*), we find:

*stick * NECK out*	*sticking * NECK out*	*sticks * NECK out*	*stuck * NECK out*
60%	13%	4%	23%

Tab. 8: Percentages of inflectional forms of the verb phrase *stick * neck out*

Here the primary difference by comparison with *STICK* as a whole is that the base form *stick* is attested far more often, mostly at the expense of *stuck*, and this appears to be due primarily to the prolific use of this idiom with future tenses and modal expressions (*will, going to, might, could, be willing to*) to express intention or future possibility, whereas the rate for '*sticks * neck out*' is close to the average for *-s* forms in general.

Let us now turn to scores for *STICK* co-occurring with '*sore THUMB*' within a span of 5, which produces 126 occurrences in the corpus. 98% of the outcomes of this query embrace the sequences *out like a sore thumb* or *out like sore thumbs* (e.g., "George and Madeline stick out like sore thumbs in this typically southern setting"):

stick+sore THUMB	*sticking+sore THUMB*	*sticks+sore THUMB*	*stuck+sore THUMB*
41%	13%	32%	14%

Tab. 9: Percentages of inflectional forms of the verb *stick* + *sore thumb* within a span of 5

Here the figures for *stick* and *sticking* do not vary significantly from those for *stick* and *sticking* unspecified for context, but what is striking is the extraordinarily high percentage of *sticks* compared both to the average of 6% of the *-s* form for *STICK*, '*STICK* to * *guns*,' '*STICK * NECK out*,' and to the average for *-s* forms in general. This is mostly at the expense of *stuck*, which shows a low 14% compared with *stuck* unspecified for context (40%). Accounting for this distribution seems arduous, but learners of English should ideally be made aware of the anomalous frequency of the *-s* form here.

As a final example, consider the idiom '*STICK in * MIND*' (786 occurrences at a rate of 0.51 per million, capturing usage such as *stick in my mind, sticking in everyone's minds, sticks in his mind, stuck in the minds*):

*stick in * MIND*	*sticking in * MIND*	*sticks in * MIND*	*stuck in * MIND*
32%	< 1%	39%	29%

Tab. 10: Percentages of inflectional forms of the verb phrase *stick in * mind*

Here, by comparison with *STICK* overall, the base form *stick* again constitutes a third of the total, while *stuck* maintains a healthy if lower percentage. However, the most conspicuous outcomes are those of (i) '*sticking in * MIND*,' which is close to zero, and (ii) '*sticks in * MIND*,' which reaches 39%, an astounding score when compared both to the average of 6% of the *-s* form for *STICK*, '*STICK* to * *guns*,' '*STICK * NECK out*,' and to the average for *-s* forms in general. These variations are again not easy to account for, though one notes the preference for non-progressive structures. Thus both '*sticks+sore thumb*' with span 5 and '*sticks in * MIND*' show exceptionally high scores.

Note in passing that '*STICK in * * MIND*'—with two missing words—produces 20 occurrences in the corpus (including "designs that stick in your customers' minds," "the point remains stuck in my own mind," as well as one irrelevant occurrence which is "stuck in London never mind") and the following distribution:

*stick in * * MIND*	*sticking in * * MIND*	*sticks in * * MIND*	*stuck in * * MIND*
50%	0%	15%	35%

Tab. 11: Percentages of inflectional forms of the verb phrase *stick in * * mind*

By comparison with the previous table, the most salient figure here is that the *-s* form is 2.5 times less frequent. This statistic might seem unremarkable by virtue of the limited number of occurrences in the corpus, but it becomes considerably more interesting when we ascertain that searches in other large corpora (*English Web 2013, English Web 2018, English broadsheet newspapers 1993-2013*) generate approximately the same outcome: '*sticks in * MIND*' is consistently 2 or 2.5 times more frequent than '*sticks in * * MIND*.' Now while there is clearly the risk of overload if this type of more subtle information is repeatedly communicated to learners with reference to all sorts of idiomatic expressions, this frequency discrepancy provides further evidence that existing language descriptions leave a lot, so to speak, on the cutting-room floor.

Naturally the queries outlined above for *STICK* could be extended to phrasal verbs such as *stick by, stick it out, stick up for* etc.

5. Conclusions

In language research very little attention has been devoted to the frequency of inflectional forms, though some interest has been shown in their lexical profiles and their varying degrees of metaphorical usage. Dictionaries, on the other hand, with regard to nouns, focus for the most part on the singular and plural, in part because the correlation between form (noun form without *-s* vs noun form with *-s*) and function (singular noun vs plural noun) is fairly manageable, i.e., the flag 'usually singular' generally corresponds to 'usually without *-s*/*-es*,' and 'usually plural' generally corresponds to 'usually with *-s*/*-es*.' The correlation between the form and function of verbs is much less straightforward, and as a result it would be largely unhelpful for dictionaries to supply formal labels of the type 'mostly base form' or 'mostly with *-ed*,' though indications are supplied in functional terms, even if these are almost always confined to the passive voice ('usually passive' or 'no passive'), to the progressive aspect ('often continuous,' 'no progressive' etc.) and, very occasionally, to the imperative.

Why the passive and progressive are prioritised rather than other worthy candidates (present tense, past tense, present perfect, future forms, first/second/third person, singular vs plural subject) is not entirely clear, but a particularly worthy pretender is the third-person singular of the simple present, whose morphology is almost always instantly distinguishable from that of other verb functions, something which makes its identification in a corpus considerably less complicated. Just as it is useful for the learner to be apprised of the fact that *REGULATION* is 'usually plural' (the plural form is twice as frequent as the singular and proportionally around six times more frequent than plural forms in general), it would also be useful for the learner to know that *RESEMBLE* and *PERMEATE* are 'often simple present third-person singular' (again around six times more frequent than the average). Clearly learners of English are not primed to recognise or produce features of this nature.

During the last part of this article, the huge variation of the frequencies of inflectional verb forms was illustrated with reference to the verbs *DODGE* and *STICK* (i) unspecified for context and (ii) within specific idioms. Particular attention was again devoted to *-s* forms, and it was noted firstly that *dodges* unspecified for context is over five times more frequent than '*dodges + BULLET*' within a span of 5, and secondly that '*sticks in * MIND*' and '*sticks + sore THUMB*' are prodigiously more frequent than *sticks* unspecified for context, than '*sticks to * guns*' and than '*sticks * NECK out*.' One could certainly debate the reasons for these discrepancies, but native or near-native speakers of English are primed to recognise and reproduce idioms in accordance with such frequency variations, whereas learners are not, and the jury is out as to whether learners should be made aware of variations of this type, or whether there is the risk of information overload.

Finally, it was noted that in a range of large corpora, the *-s* form of '*STICK in * MIND*' ("One such summer night sticks in my mind") is consistently over twice as common as the *-s* form of '*STICK in * * MIND*' ("It sticks in most people's minds because of the theme tune"). Inclusion of very subtle details such as this in language resources would probably either sap learners' morale or blow their mental circuit once and for all, but if nothing else it would provide further proof of how partial our current knowledge of the language iceberg really is.

Dominic Stewart *teaches English Language and Linguistics and Italian-English Translation at the University of Trento. He has published primarily in the field of corpus linguistics and translation. His publications include* Semantic Prosody: A Critical Analysis *(2010)*, Italian to English Translation with Sketch Engine: A Guide to the Translation of Tourist Texts *(2018), and* Frequency in the Dictionary: A Corpus-Assisted Contrastive Analysis of English and Italian *(2021).*

Works cited

Esser, Jürgen. "Corpus Linguistics and the Linguistic Sign." *Corpus Linguistics and Linguistic Theory*: *Papers from ICAME 20*. Edited by Christian Mair and Marianne Hundt. Amsterdam: Rodopi, 2000. 91-101.

Halliday, Michael and Zoe James. "A Quantitative Study of Polarity and Primary Tense in the English Finite Clause." *Techniques of Description: Spoken and Written Discourse*. Edited by John Sinclair, Michael Hoey and Gwyneth Fox. London: Routledge, 1993. 32-66.

Hoey, Michael. *Lexical Priming: A New Theory of Words and Language*. London: Routledge, 2005.

Kilgarriff, Adam, et al. "The Sketch Engine: Ten Years On." *Lexicography ASIALEX* 1 (2014): 7-36.

Knowles, Gerry and Zuraidah Mohd Don. "The Notion of a 'Lemma.'" *International Journal of Corpus Linguistics* 9.1 (2004): 69-81.

Leech, Geoffrey, Paul Rayson and Andrew Wilson. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Routledge, 2014.

O'Halloran, Kieran. "Critical Discourse Analysis and the Corpus-Informed Interpretation of Metaphor at the Register Level." *Applied Linguistics* 28.1 (2007): 1-24.

Sinclair, John. "Beginning the Study of Lexis." *In Memory of J.R. Firth*. Edited by Charles E. Bazell, et al. London: Longman, 1966. 410-430.

Sinclair, John and Ronald Carter. *Trust the Text: Language, Corpus and Discourse*. London: Routledge, 2004.

Sinclair, John, Susan Jones and Robert Daley. *English Collocation Studies: The OSTI Report*. London: Continuum, 2004.

Stewart, Dominic. *Frequency in the Dictionary: A Corpus-Assisted Contrastive Analysis of English and Italian*. Bern: Peter Lang, 2021.

---. "Grammar Labels for Verbs in English Monolingual Learners' Dictionaries." *Iperstoria* 16 (2020): 192-212.

Stubbs, Michael. *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell, 1996.

---. "The Search for Units of Meaning: Sinclair on Empirical Semantics." *Applied Linguistics* 30.1 (2009): 115-137.

Tognini-Bonelli, Elena. *Corpus Linguistics at Work*. Amsterdam: John Benjamins, 2001.

Zhang, Rui-Hua. "Form, Meaning and Learners' Dictionaries." *Studies in English Language and Literature* 32 (2013): 29-50.

English Dictionaries

Collins Reverso Dictionary Online. https://dictionary.reverso.net/english-cobuild/. All websites last visited on 08/03/2022.

Longman Dictionary of Contemporary English Online. https://www.ldoceonline.com.

Macmillan Dictionary Online. https://www.macmillandictionary.com.

Oxford Advanced Learner's Dictionary Online.

https://www.oxfordlearnersdictionaries.com/?cc=it.

Oxford Lexico UK Dictionary Online. https://www.lexico.com.