14 Commentators and Corpora: Evidence about Markers of Formality

David Minugh
Stockholm University

1. Introduction

Dictionary-makers and stylists have long singled out various terms for special notice, and at times had strong opinions about their use and abuse. These comments were in many cases essentially a matter of taste (often masquerading as logic), but until corpus linguistics and powerful computers arrived on the scene, no tools existed to demonstrate actual usage, beyond collections of (laudable and reprehensible) examples. Logical and sentential connectors have not escaped such scrutiny, and here we shall focus on three fairly formal such terms, all of which have interesting characteristics from a learner perspective: *albeit*, *notwithstanding* and *thus*. After briefly considering their origins, we will examine some of the comments about them, particularly by grammarians and style police, and then bring in data from recent corpora to examine their actual use, which will not always prove to be in formal settings.

2. Origins

The lexical items *albeit, notwithstanding* and *thus* are not particularly obscure in their development, although they do have a reasonably venerable pedigree. The *OED Online* considers the etymology of ***albeit*** as straightforwardly deriving from *all* as a conjunction and the present subjunctive of *be*, with the first instances surfacing with clauses in the late 14th century:

> [1] "But syn my name is lost thurgh you," quod she,
> "I may wel lese a word on yow or letter,

How to cite this book chapter:
Minugh, D. 2015. Commentators and Corpora: Evidence about markers of formality. In: Shaw, P., Erman, B., Melchers, G. and Sundkvist, P. (eds) *From Clerks to Corpora: essays on the English language yesterday and today*. Pp. 239–265. Stockholm: Stockholm University Press. DOI: http://dx.doi.org/10.16993/bab.n License: CC-BY.

Al be it that I shal be neuer the better"
[Chaucer, *Legend of Good Women*, 1361–63][1]

Further instances soon show it introducing other constructions, such as PPs:

[2] We dyd graunte (albeit not for this argumentacyon) that…
[Marshall, 1535]

or as an adverb, as in the OED's quite recent final citation:

[3] Young skunks begin to spray, albeit inaccurately, at about one month of age.
[1995, *Animals' Voice* Spring 13/1]

Notwithstanding is also a compound form, straightforwardly derived, as Johnson noted,[2] from *not* + *withstand*, on the pattern of Anglo-Norman and Old French *non obstant* and post-classical Latin *nōn obstante*, with the same sense, appearing shortly after *albeit*:

[4] Natwith-stondinge his grene mortal wounde, He ros ageyn.
[c1425, Lydgate *Troyyes Bk.*]

Its most striking grammatical feature, the ability to function as a postposition (or adverb, depending on your analysis),[3] is also documented from within less than a century later, as in Caxton:

[5] This notwystondyng, alwaye they be in awayte.
[*Eneydos*, 1490]

a variation which remains its hallmark until the present day.

Our final (and oldest) item, ***thus***, apparently has its roots in the demonstratives (the OED suggests derivation via either *that* or *this*). Some early examples from the OED:

[6] *Sicini* [*siccine*], ac ðus
[*c725 Corpus Gloss.* 26]

---

[1] Much of the detailed *OED* information has been removed from these citations; the Chaucer quote follows the text in Fisher 1977:643. The *OED* also mentions variants such as *al were it, albe* (both with further citations from Chaucer), but the clearest view of the range of this type of construction actually emerges from the examples cited in Jespersen 1940.

[2] "*[Notwithstanding]* is properly a participial adjective, as it is compounded of *not* and *withstanding*, and answers exactly to the Latin *non obstante*" (1783, Vol. II).

[3] Cf. Rissanen 2002 and Weber (2010:181–86) for Middle English developments, Minugh 2002 for modern English use.

[7] & tuss ȝho seȝȝde inn hire þohht..Þuss hafeþþ drihhtin don wiþþ me.
[?*c*1200 *Ormulum* (Burchfield transcript) l. 235–7]

[8] Here vn-to you þus am I sente.
[*c*1440 *York Myst.* vii. 6]

3. Learners' perspectives

Albeit. From the start, the term ***albeit*** stands out for phonetic reasons. For foreign learners and young native speakers alike, the word is normally first encountered in written form, so that the trisyllabic pronunciation /ɔːlˈbiːɪt/, with a clear *be*, often comes as a distinct surprise, particularly if they have previously paid attention to items with reduced stress, such as the RP pronunciation of *secretary*. To most speakers of English, the etymological links to *al-* (as in *although*) and subjunctive *be* are not at all obvious, particularly since the latter's primary current use, the mandative subjunctive (e.g. *I move that the meeting be adjourned*), is not frequent (Hundt 1998); in addition, the pronunciation of the final *-it* as a distinct final syllable is unexpected. Placing the stress on the first syllable (as in *alien, alias*) would lead to something like */ˈeɪlbɪt/, which has apparently never been current. John Wells (2008: 19) records the frequent but "non-RP" pronunciation /ælˈbiːɪt/ (he also notes it for AmE, a form that e.g. Elster [1999:13] takes violent exception to). Once learned, its pronunciation is easy enough (in parallel to *although*), and as regards usage, it presents no particular difficulties, functioning as a synonym of *even though* or *although*.

Notwithstanding. For learners, the pronunciation of ***notwithstanding*** ought to be straightforward (once they grasp that it is a single unit), and as for usage, its preposed placement predominates in BrE, and causes no problem. This position allows it to control fairly long (and relatively complicated) constructions, whereas its postposed use tends to be limited to controlling short NPs. The postposed use is above all found in more formal AmE (cf. Minugh 2002 for statistics).

Thus. The voiced initial consonant of ***thus*** follows the normal deictic patterns seen in *the, this, those, thy* and so on. It has no direct cognates in Romance or Germanic languages (the sole exception is Dutch *dus*).[4]

[4] *Dus* is considerably more frequent than its English counterpart: the *Dutch Web Corpus* (via the commercial program called *SketchEngine*) reports it as having an occurrence of 1,299 per M words; by comparison, the *Oxford English Corpus* (again via *SketchEngine*) reports *thus* as having an occurrence of 153 per M words.

Its simple monosyllabic form and its use parallel those of other logical connectors such as *so*. But like *therefore* and *as a result*, *thus* has a distribution heavily slanted towards formal written English; this (and the lack of cognates) appears to delay its acquisition, at least in Sweden, where informal English is given priority in the school system.

Of these three items, only *thus* (which has the widest functional range) was regarded as sufficiently important to be included in the classical General Service List (West 1953), the first reasonably modern word list for learners. When the *Academic Word List* was developed (Coxhead 2000), all GSL words were excluded, as already covered, so that *thus* was not included in the AWL; the latter does, however, include both *albeit* and *notwithstanding*.[5]

4. Stylistic comments by dictionaries, grammars and style manuals

In this section, we will briefly survey what various reference works have had to say about our three terms, and what claims, if any, they make about the validity of their comments about the use of *albeit, notwithstanding* and *thus*. It should be noted at the outset that there is no significant disagreement about the semantics of these terms; what is at issue are matters of register and style.

4.1 Major dictionaries

Johnson's epoch-making *Dictionary* (1755) records all three items without further ado, notably without any comments on their stylistic level. The reader is reminded that he was by no means above pronouncing judgments about usage: while *thus* is merely recorded, compare his comment on the very next word, *thwack*: "A ludicrous word" (1799, Vol. II).

The first edition of the *OED* (1933) passes over the stylistic value of *albeit* and *notwithstanding* in silence, but begins the article for *thus* with the note "now chiefly literary and formal" (1933:XI, 397); more interestingly, no changes in this judgment are to be noted even in the contemporary *OED Online*. In addition, the one-volume *New*

[5] The increasing impact of the *AWL* is seen not least at the English Department of Stockholm University, where an "AWL Vocabulary" test is administered to entering students early in their first semester. Not until 2013 did its first serious competitor appear (http://www.academicvocabulary.info/); cf. Gardner & Davies 2013. Note also section 4.4, below.

Oxford Dictionary of English (1998) similarly has a note only for *thus*: "poetic/literary or formal". *Webster's New International Dictionary of the English Language* (1941) records all three without further comment, as do *Webster's Collegiate Dictionary* (1988, 1993) the *American Heritage Dictionary of the English Language* (1970) and Australia's *Macquarie Dictionary*.[6]

4.2 Grammars

Turning now to earlier 20th-century grammars, we find that Poutsma comments "[i]n Present-day English *[albeit]* is used only in the higher literary style, mostly without *that*" (1929:I, ii, 712), with a similar comment on *notwithstanding* (711). Curme notes that conjunctions used in concessive clauses include *notwithstanding* and "[i]n older or archaic English: *albeit* (i.e., *all be it = be it entirely*) *that* or simple *albeit, albe"* (1931:II, 333). Jespersen (1940:51) remarks on the alternate pre-/postposition of *notwithstanding* and provides numerous examples of *albeit* and related subjunctive constructions (1940:364). Interestingly enough, his volume on pronunciation (1949) does not mention *albeit*.

For our triad of terms, Swedish-based university grammars of English have a long tradition of silence as regards form and use, although register is occasionally touched on. Elfstrand & Gabrielsson (1960) only mention *notwithstanding that* as a concessive conjunction. Svartvik & Sager (1977, 1996) mention *albeit* functioning to link adjectives (§353D) in "formal language" and *thus* as a linking adverbial (§439E) "in formal style". More recently, Estling Vannestål merely mentions *thus* as one of the linking adverbials (2008:269), omitting *albeit* and *notwithstanding*.

In Quirk et al., the first major grammar with a dawning awareness of corpus data, *albeit* is dismissed in a footnote: "the following archaic subordinators still have a limited currency: *albeit, whence, whereat, wherefore, whither" (1985:998,* note [b]). *Notwithstanding* is mentioned several times, usually with the label "formal"; note particularly: "*Notwithstanding* ['in spite of'] is formal and rather legalistic in style, particularly when postposed" (1985:706). Together with other prepositional phrases (*despite, in spite of, irrespective of, regardless of*),

[6] The *New Oxford Dictionary* also includes an entry on *thusly*, which is labeled "informal" (1998:1935b), while the *American Heritage* goes further, labelling it "nonstandard," noting that it "is termed unacceptable by 97 per cent of the Usage Panel" (1970:1342); cf. Menken's comments, in section 4.3, below.

notwithstanding is "considered stylistically clumsy" (1985:1098). *Thus* is consistently labeled "formal", e.g. "The form *thus* is largely formal" (e.g. 1985:557, note [b]).

In their brief discussion of register, Celce-Murcia & Freeman remark: "In any kind of informal situation, a native speaker of English would be surprised to hear somebody say *notwithstanding the fact that* to express the notion of concession. A connector such as *even though* would be much more likely" (1983:323). They nevertheless list *albeit* and *notwithstanding* under "Concession," without any comments on register (326).

Turning to modern general learner grammars, we find that *A Communicative Grammar of English* (2002) does not include *albeit*, but does mention *notwithstanding* ("very formal" [2002:113]) and *thus* ("formal" [2002:110]). The *Longman Student Grammar of Spoken and Written English* (Biber, Conrad & Leech 2002) appears to contain no information on our three terms.[7] The *Cambridge Grammar of English* (Carter & McCarthy 2006) is silent on *albeit* and *notwithstanding*. They list *thus* as an option among many, but the only concrete information given is that initial *thus* can allow inversion:

> [9] Thus does Mr Major find himself ever more closely closeted with Mr Campbell. (2006:782)

4.3 Prescriptive stylists and manuals of style

Like most of the grammars cited above, nearly all of the works cited below were written in the pre-corpus era. With a single exception to be discussed below, however, they rarely cite extensive examples to bolster their claims. Fowler & Fowler (1930:29), for example, using guilt by association, dismiss *albeit* as an archaism, listing it with the likes of *bashaw, certes, damsel* and *quoth(a)*, terms few would wish to champion as shining examples of modern English. They are silent on *notwithstanding*, but object strongly to *thus* in one case:

> In this use *thus* is placed before a present participle (*thus enabling* &c.), & its function, when it is not purely otiose, seems to be that of apologizing for the writer's not being quite sure what noun the participle belongs to, or whether there is any noun to which it can properly be attached (cf. UNATTACHED PARTICIPLES); (1929:652)

[7] A caveat: since lexical items are not included in the index, the search by subject area may have missed a minor comment on these words.

This actually sounds rather like the discourse markers sometimes referred to as *shell nouns*, i.e. a way of summing up a form of logical relationship previously presented in detail in the text (Schmid 2000), a use which they find to be too vague. In all other respects, *thus* is passed over in silence.

The Americanist H.L. Mencken found nothing to comment about on *albeit* and *nevertheless*, but was interested enough in the American use of *ly*-less adverbs to comment that: "the use of *illy* and *thusly* is confined to the half educated" (1936:467).[8] Copperud (1964) only warns against the use of *for* or *thus* at the beginning of sentences: "…an affectation by some writers, particularly columnists. This is warranted only when the sentence draws a conclusion based on what has gone before" (1964: 165). The *Longman Guide to English Usage* (Greenbaum & Whitcut 1988), silent on *notwithstanding*, does warn against "the FACETIOUS variant *thusly*", and waxes truly eloquent on *albeit*:

> This is often regarded as pretentious when used, unless for humorous effect, as an alternative to *(even) though*. It is perhaps justified as a convenient way of linking pairs of adjectives (*a small albeit crucial mistake*), although *but, yet,* and *though* will also do in this case (1988: 27).[9]

Oxford's *Authors' and Printers' Dictionary* (1956), the Chicago University Press *Manual of Style* (1969), Michael Swan's *Practical English Usage* (2005) and Collins COBUILD *English Usage* (1992) are among the numerous works silent on these three words. As for student writing manuals, an examination of the popular *Writing Academic English* revealed only that *thus* appears in several lists of "connecting words and transition signals" (Oshima & Hogue, 2006, Appendix C), while *albeit* and *notwithstanding* are passed over in silence.

However, one work stands out in its detailed comments on *albeit*, as well as its extensive use of 20th century citations (almost unique among style manuals): the *Merriam-Webster's Dictionary of English Usage* (1994). Their opening shot deserves quotation *in extenso:*

> Copperud 1970, 1980 observes that "a generation ago" *albeit* was considered archaic but is "now being revived." The source of the

[8] In Supplement Two, he records a congressman using *thusly*, but adds, "However, it is often difficult to tell whether a congressman is serious or spoofing" (1948: 390, n. 3).

[9] Also noted by Svartvik & Sager 1977 (see section 4.2, above).

> notice of revival is Gowers (in Fowler 1965). This is a most curious business, since *albeit* seems never to have gone out of use, though it may have faded somewhat in the later 19th century. If it did, the revival began decades before the commentators noticed. (1994:65)[10]

They go on to trace a lineage of *albeit* quotes from 1907 to Krapp's grammar in the late 1920s, with a last example from the 1980s. As noted in section 4.2, above, as late as 1985 Quirk et al. labelled *albeit* as archaic, despite such evidence.

What we seem to find, then, is a series of fairly random objections to specific uses or forms (such as *thusly*), while "allowing" others. This is hardly surprising, given that these writers were unable to systematically trawl through large amounts of text from many different domains for matters of interest. To do so, we must turn to recently-compiled corpora for documentation. In doing so, we will concentrate on these three terms and their frequencies over the last two centuries. The first indications of what this can result in may be seen in modern learner dictionaries, to which we now turn.

4.4 Learner dictionaries

Starting with the first edition of the Collins COBUILD dictionary (1987), but increasingly in the period after 2005, learner dictionaries have based their labelling on data drawn from (usually in-house) large corpora, i.e. corpora now normally in excess of 100 M words. It is nevertheless worth noting the comments from the editors of the Oxford *Advanced Learner's Dictionary (ALD)*: when it comes to deciding on the recently-introduced "Oxford 3000" keywords ("the words which should receive priority in vocabulary study because of their importance and usefulness"), they based their decision on corpus frequency and range of text types—but also as being "very familiar to most users of English", as judged by "language experts and experienced teachers" (2005:R99). In other words, for Oxford, the corpus is definitely not considered the sole arbiter in adjudicating on such matters.

What, then, do learner dictionaries say about our three terms? A pre-corpus edition of the Oxford *ALD* (2nd ed., 1963) labels *albeit*

[10] Gowers states that "*[albeit]* has since been picked up and dusted and, though not to everyone's taste, is now freely used, e.g. *It is undeniable that Hitler was a genius, a. the most evil one the modern world has known*" (1965:16). Note also that Copperud 1964 was silent on *albeit*; in later editions he is clearly aware of the changing perception of *albeit*'s status.

as "not colloq[uial]", but is otherwise silent on this issue. By the 7th (corpus-aware) edition of 2005, it labels all three as "formal", while including *thus* as one of its "Oxford 3000" keywords. In the 8th edition (2010), *albeit* and *notwithstanding* are additionally labeled as AW (i.e., part of the Academic Word List, which has now made its entry into the *OALD*).

COBUILD editions (1987, 1995) are relatively consistent, labelling *albeit* and *notwithstanding* as "a formal word", but *thus* as "a fairly formal word".

By its 4th edition, the Longman *Dictionary of Contemporary English* labels all three as "formal". *Thus* is noted as W1, i.e. among the 1000 most common words of written English, but with a warning triangle indicating that when it is used as a sentence adverb, "in spoken English it is more usual to use *so*". In the 5th edition (2009), *albeit* and *notwithstanding* receive the additional label AC (i.e., part of the Academic Word List, which has now made its entry into *LDOCE*, as well).

The *Cambridge Advanced Learner's Dictionary* (3rd ed., 2008) labels all three as "formal". *Thus* is additionally noted as 1, for "improver", the middle category in its high-frequency words.[11]

Table 1. Learner dictionary labels for albeit, notwithstanding and thus.

	albeit	***notwithstanding***	***thus***
ALD (1963)	not colloq.	(no label)	(no label)
ALD (2005)	formal	formal	formal **Oxford 3000**
ALD (2010)	formal **aw**	formal **aw**	formal **Oxford 3000**
COBUILD (1987)	formal	formal	fairly formal
COBUILD (1995)	formal	formal	fairly formal
LDOCE (2005)	formal	formal	formal
LDOCE (2009)	formal **ac**	formal **ac**	formal **w1**
CALD (2008)	formal	formal **I**	formal
MEDAL (2007)	formal ★	formal	formal ★★★
CDAE (2000)	(omitted)	(no label)	formal
OADCE (1999)	(no label)	(no label)	formal

[11] More specifically, this applies to *thus* in the senses 'in this way' and 'with this result', with a frequency typically of 200–400 per 10 million words (2008:VIII).

The *Macmillan English Dictionary for Advanced Learners* (2nd ed., 2007) also labels all three as "formal", but includes *albeit* among the 7000 most common words of English (one star) and *thus* among the 2500 most common words (three stars, its highest frequency rating).[12]

As for these dictionary-makers' American offshoots (which are invariably smaller, presumably in order to sell better in America), the *Oxford American Dictionary of Current English* (1999) notes that *thus* is formal, but has no labels for *albeit* or *notwithstanding*. The *Cambridge Dictionary of American English* (2000) omits *albeit* altogether, but includes *notwithstanding* (with only one example—a postposed one!) and *thus* (considered "formal").

Summarizing, we obtain the table below, from which it appears clear that the dictionary-makers are in agreement on both register and frequency. This should not lead to conspiracy theories about borrowing from one another, but rather is a consequence of their now having access to large, proprietary corpora yielding similar results. However, it has recently become possible for scholars independently to check on these results, thanks to large-scale publicly-available corpora, to which we now turn.

Table 2. Frequencies from the BNC (Lancaster interface), including comparative data for *therefore*.

Term	Written	%	Per M wds	Spoken	%	Per M wds
albeit	1330	96.6%	15.13	47	3.4%	4.51
notwithstanding	701	97.4%	7.97	19	2.6%	1.83
thus	20,127	99.6%	228.97	84	0.4%	8.07
therefore	21,406	93.2%	243.52	1567	6.8%	150.53

[12] This is the BrE version; the AmE edition is *Macmillan English Dictionary for Advanced Learners of American English*, which had no second edition, their subscription website having instead taken over all updating.

5. Corpus data for *albeit*, *notwithstanding* and *thus*

5.1 Contemporary corpora[13]

The earliest "large" (= 1M word) corpora of contemporary written English are the *BROWN* series: *BROWN* and *LOB* have matched AmE/BrEtexts from 1961, and *FROWN* and *FLOB* similarly matching texts from 1991: They produce quite small numbers (only *thus* yields results larger than 20 examples per corpus), but they will turn out to be quite close to the results from the much larger corpora now available.[14]

The major 90s corpus, the *British National Corpus (BNC)*, at 100M words (British English only) yields our first solid data on how these words are distributed along the written/spoken dimension:[15]

This clear preponderance of written instances suggests that we are dealing with what is tantamount to words found in written English only, particularly if one considers that some of the "spoken" data consists of prepared transcripts for radio and TV. In absolute numbers, *thus* is once again more common than the other two by more than an order of magnitude. Using a further analysis from the Brigham Young (BYU) interface, we can break down the results into different domains (Figure 1):[16]

[13] The corpora in this section all seek to portray modern English from the 1990s onward. The most purely synchronic of these corpora are of course the BROWN group, each of which samples only one year. The diachronic corpora in section 5.2, on the other hand, cover a much larger temporal range, precisely in order to track changes over time.

[14] For descriptions of these earlier corpora, see any standard undergraduate textbook on corpus linguistics, e.g. McEnery, Xiao & Tono 2006.

[15] The now much larger Collins COBUILD *Bank of English* was the first modern corpus of English, but the open-access policy of the *BNC* continues to be crucial to scholarship; meanwhile, the publicly available component of the *BOE* has evolved into the 57M Collins *Wordbanks Online*, currently openly available at http://www.collinslanguage.com/content-solutions/wordbanks.

[16] A technical note: by comparing instances per 1M words, and selecting for each word the domain with the largest number of instances as 100%, we can graphically compare all domains for each word individually.

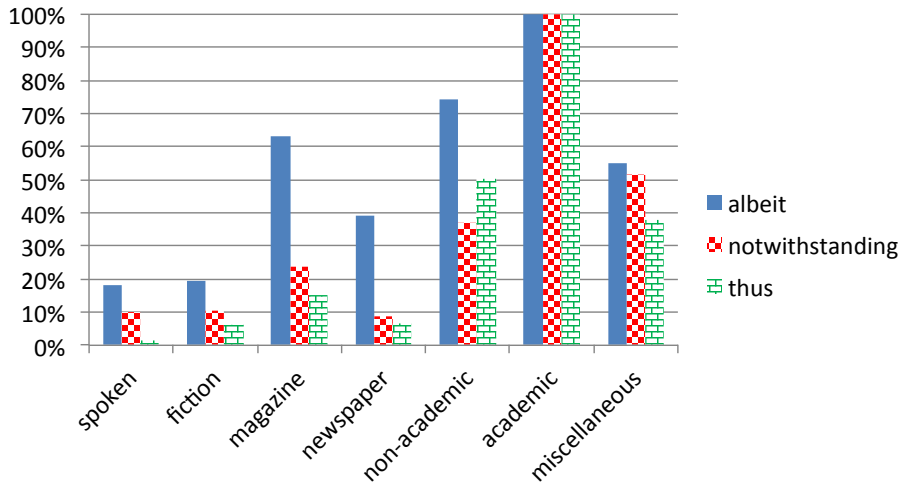

Figure 1. *BNC* distribution frequencies (BYU interface).

As expected, all three connectives occur most frequently in the academic domain. However, the most striking aspect of this comparison is that *albeit* is considerably more evenly distributed than the other two; the discrepancy is so large that it is difficult to ascribe it merely to being an artefact of the domain definitions.[17] The chi-square test returns a significance of well below p< .001.

Since the *BNC* is specifically limited to BrE, let us next turn to the *Corpus of Contemporary American English (COCA)*, which now covers a little over two decades, from 1990 on.[18] Containing 450 M words, including a "spoken" section (largely derived from radio/TV transcripts), it is the largest broadly-based contemporary corpus with free access, although also limited geographically. Not surprisingly, in raw numbers, *thus* again dominates by more than an order of magnitude, with 62,764, compared to *albeit* and *notwithstanding*, with 4,061 and 2,683, respectively. We therefore again choose to display the data as percentage comparisons to the largest category for each item, again based on frequency per 1M words (Figure 2):

[17] The values for the category "miscellaneous", on the other hand, clearly indicate that something is escaping this categorization.

[18] It has been expanding as time passes, now including up to 2012, so that this is an evolving synchronic record of "contemporary" American English. Like several other contemporary corpora, it thus will not yield replicable results over time, since the corpus itself is growing; cf. the "monitor corpus" solution adopted by John Sinclair and the COBUILD team (for an early description, see e.g. Clear 1998).

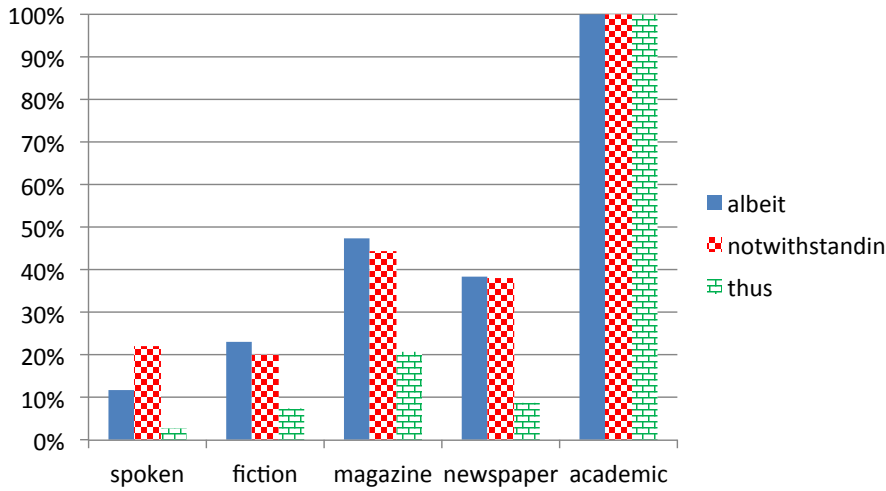

Figure 2. *COCA* distribution frequencies.

From Figure 2 it is clear that for all three items, the academic domain dominates, having more than twice the frequency found in the other domains. For *thus*, the dominance of the academic domain is overwhelming, while both *albeit* and *notwithstanding* have a certain currency in magazines and newspapers, perhaps due to their feature

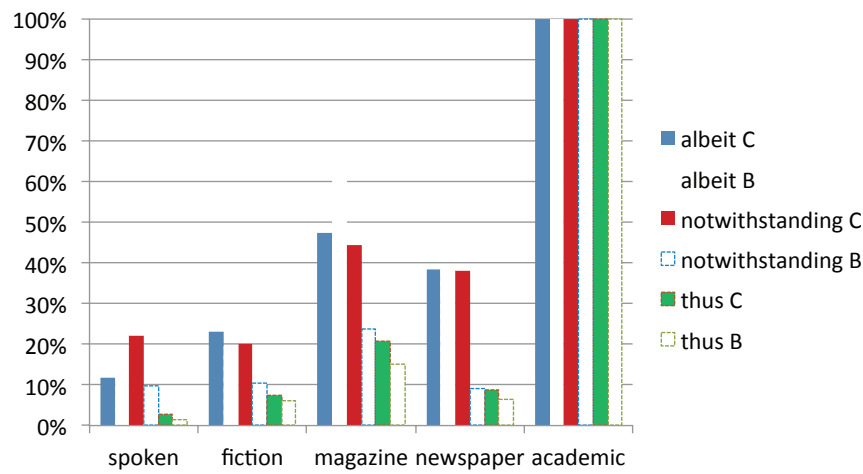

Figure 3. *COCA* (C) and *BNC* (B) distribution frequencies (comparable domains).

articles. With the possible exception of spoken *notwithstanding*, the three items show similar distribution patterns, with *thus* having the widest gap between the other domains and academic English. Via the chi-square test, both *albeit* and *thus* distributions are significant at $p < .001$, while *notwithstanding* has $p < .0157$.

Naturally, it is interesting to compare the British and American data. Since both corpora are available with the same (BYU) interface, this would appear to be simple, but the *BNC* data has the two extra categories of **non-academic** and **miscellaneous**, which in unknown fashion are redistributed in *COCA*'s fewer domains. Omitting those two *BNC* categories, our comparison looks like this (Figure 3).

The fit between these two geographical domains is quite good for both *albeit* and *thus*, which is not surprising, given that four of the five are written, the domains where BrE and AmE are traditionally considered to have the smallest differences. The odd man out is

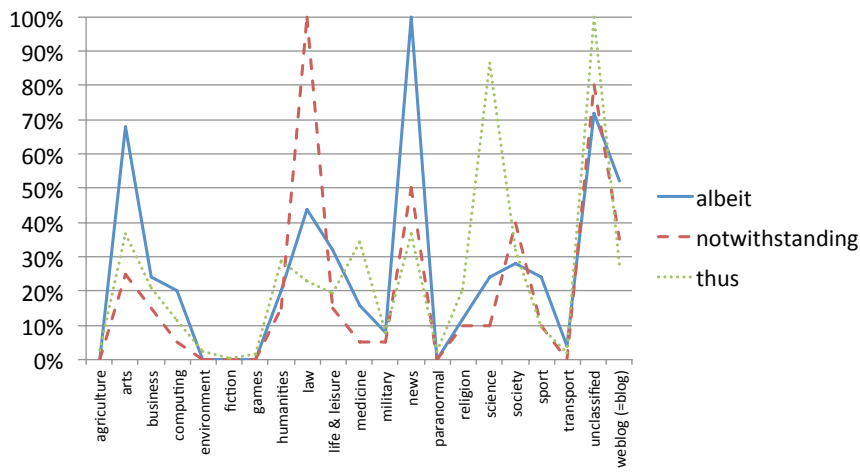

Figure 4. *OCE* distribution frequencies (SketchEngine interface).

notwithstanding, which seems to be more favored in AmE, again with the reservation for the *BNC* data loss.

Stepping up to an even larger corpus, the Oxford Corpus of English, a corpus from the early 2000s based on material from the Web, now supplemented to reach 1736 M, and including significant input from English in other parts of the world, we again find that *thus* is more frequent by an order of magnitude: 152.4 words per million, versus *albeit*

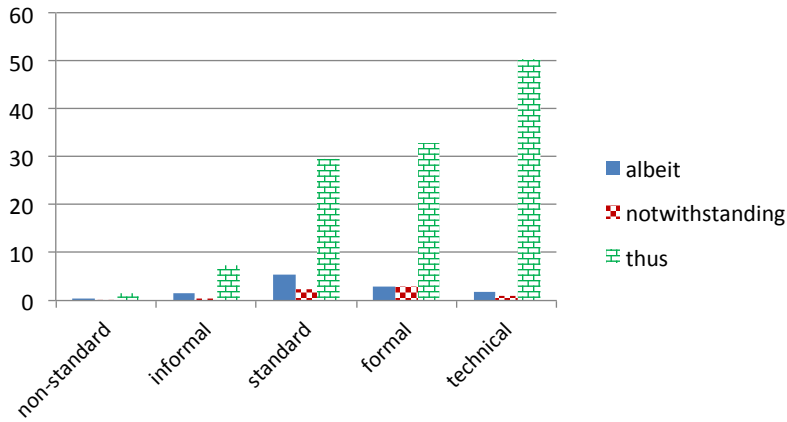

Figure 5. *OCE* frequencies vis-à-vis formality, per 1M words (SketchEngine interface).

and *notwithstanding* at 13.7 and 8.4, respectively. Comparing their frequency relative to the largest domain (SketchEngine provides 21 different domains), we see a largely parallel pattern for the three, including great differences between domain frequencies (Figure 4).[19] *Albeit* weighs in heavily in the arts, and above all, in news—perhaps an attempt to give news greater weight? Not surprisingly, *notwithstanding*'s single largest component is legal texts. *Thus*, however, turns out to be relatively evenly distributed, with the single spectacular exception of science texts, where it occurs three times more frequently than in any other domain (except the dubious "unclassified"). All three terms are relatively well represented in "weblogs", perhaps because these are relatively early blogs, when they had not yet reached the demotic level of today's twittering. Also of interest is that for all three terms, only a few domains reach levels more than 33% of the most frequent domain, again indicating that the distribution of these words is quite domain-sensitive.[20]

The SketchEngine software for the *OCE* also allows us to look at this data via degrees of formality, ranging from *non-standard* to *formal* and

[19] The values for the category "unclassified" are uniformly high, again suggesting that something fairly formal about them is escaping the categorization.

A technical note: to keep the diagram legible, the type of graph has been changed, but the domains are of course independent of one another.

[20] In terms of raw numbers, these are quite robust samples, with 16,293 instances of *albeit*, 9,024 of *notwithstanding*, and 189,969 of *thus*, so that even one of the smallest, the fiction examples of *thus*, weighing in at 0.1 per million, still totals 109 separate tokens.

technical.[21] Here, the most striking distribution is that of *thus*, whose use peaks in the *technical* texts (suggesting that this group, rather than *formal*, includes most scientific texts), but which is still a clear presence from the *standard* level on upward. The other, somewhat surprising factor is that *albeit* seems above all to be a marker of *standard* texts, as seen in the peaks in the arts and news domains; domains such as science and law are less entranced with its quasi-literary flavor, it seems.

5.2 Diachronic corpora

Here, we shall consider three very recent diachronic corpora: the *TIME* corpus (100 M words, 1923–1996), the *Corpus of Historical American English* (*COCA*; 406 M words, 1810–2009), and the *GOOGLE US/UK* corpus (in two parallel parts: AmE 155 B words, BrE 34 B words, 1810–2000), all of them created at Brigham Young University, and with the same interface.

The *TIME* corpus is one of the few corpora that chart a single source over a long period.[22] *Time Magazine* began publication in 1923, and this corpus includes all the texts in *Time* (excluding ads, picture captions, etc.) from its inception until 1996. Two factors are of particular importance when using this corpus: first, nearly all of its articles

[21] Their category *technical* is clearly not automatically "more formal" than, say, *formal*, but presumably much narrower in domain. Since SketchEngine makes this division of the entire corpus, *technical* is included in the present discussion.

[22] There is a small, but clear overlap between this corpus and both *COCA* and *COHA*, as the latter two corpora could hardly ignore *Time* when dealing with contemporary and historical American magazine writing.

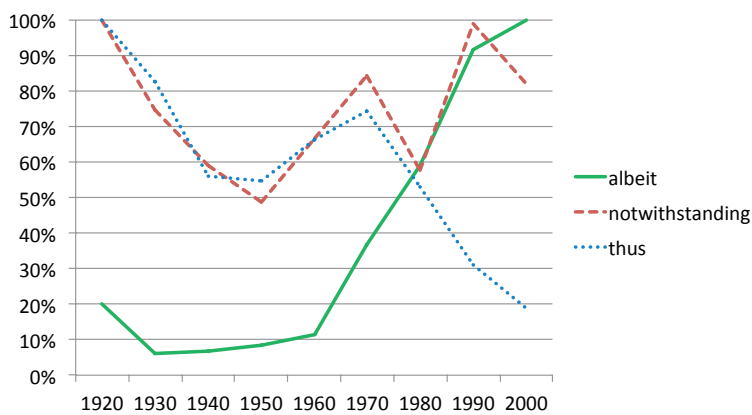

Figure 6. *TIME* distribution frequencies.

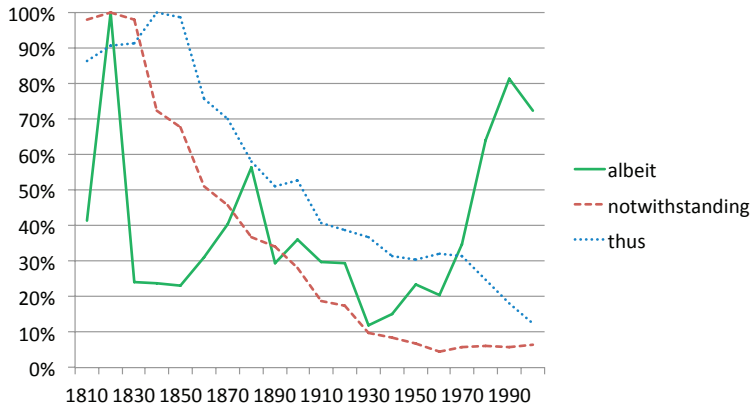

Figure 7. *COHA* distribution frequencies.

are collectively written, with a small staff of writers and editors interacting on many articles; second, it has been read in a large number of middle-class American homes for generations. As such, it is hardly representative of all American writing, but is disproportionately influential (although with a penchant for word play and *bons mots*). Again, the most revealing way to look at its statistics is to compare changes in relative frequencies per decade (Figure 6).

There are two striking changes for our word trio: first, *thus* has undergone a steep decline, broken only by a resurgence during the 60s and 70s, and ending up at less than 20% of its frequency in the 20s, from 347 per M words to 65 per M words after 2000. This is clearly in line with the specialization (tantamount to domain loss) we see in the contemporary corpora, where the vast majority of the modern instances are in science articles, a domain that does not feature prominently in *Time*. The second is the rise of *albeit*, which, at 0.6 per M words in the 1930s (i.e. less than the frequency of recondite words such as *germane*, which in turn is almost 40 times less frequent than *relevant*), rises uninterruptedly to 10.4 per M words in the 2000s. This is almost double the peak frequency of *notwithstanding*, which fluctuates from 3.2 to 5.5 per M words throughout this time period. The fluctuations of *notwithstanding* suggest that we may not be seeing change that is a trend, but rather a fairly stable term with a variation of ± 1 per M.

Turning to the *COHA* corpus, we shift to a more traditional type of linguist's corpus, i.e. a sample selected for linguistic purposes. It covers two centuries, and again our comparison is of relative frequencies (Figure 7). From the perspective of this longer time scale, we see that

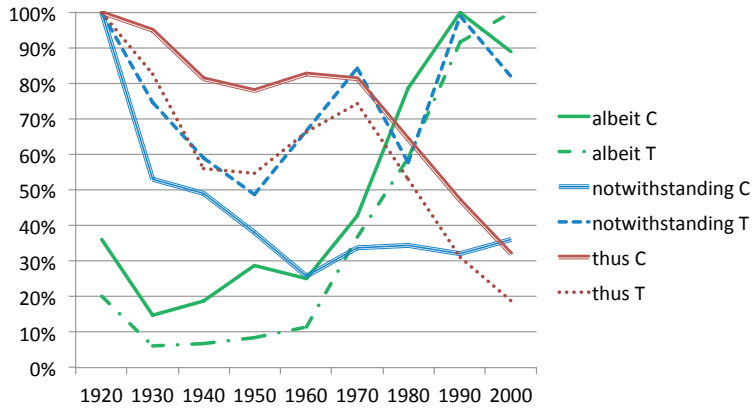

Figure 8. COHA and Time distribution frequencies, 1920s–2000s.

both *thus* and *notwithstanding* have declined drastically from the early 19th century to the present, and furthermore, rather consistently. The odd man out is *albeit*, which has an extraordinary peak in the 1820s (probably a result of its sampling,[23] and a more genuine higher level in the late 19th century, but which rises steadily from its low point of 1.22 per 1 M words in the 1930s to 7.41 in the 2000s. The shifts in *notwithstanding* and *thus* are both statistically significant (both with $p < .001$), but not the variation in *albeit*.

If we compare these two corpora during the time period 1920–2010 (ignoring for the moment the obvious distortion effects of comparing an entire range of written language with the language of a small group of editors and writers working at one publishing house), we find the following (Figure 8). Both *Time* and *COHA* begin with a high level of *thus* and *notwithstanding*, but quite a low level of *albeit*. They match quite well for both *albeit* and *thus*, the former dipping, then rising sharply, and the latter dipping, then dropping off sharply (probably a reflection of the domain loss suffered by *thus*). *Time*'s retention of

[23] These early decades have far fewer works to draw upon than the rest of the corpus, and are thus more vulnerable to sampling peculiarities. In particular, of the 71 instances of *albeit* in the 1820s, 55 are from a single work, *The Buccaneers: A Romance of Our Own Count[r]y in Its Ancient Day* . . . [by] Yclept Terentius Phlogobombos [pseud, actually Samuel Judah].

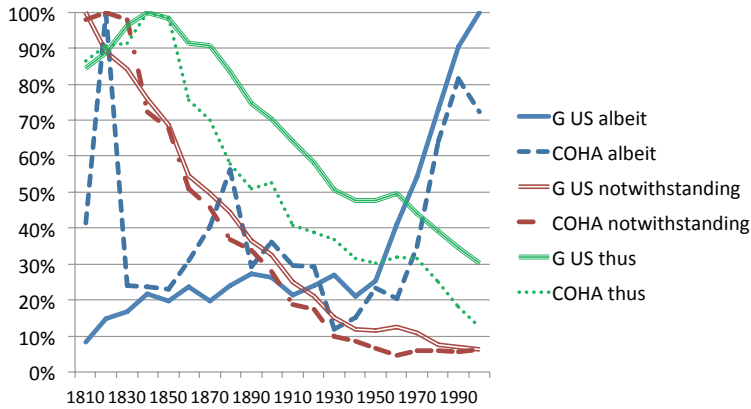

Figure 9. *Google US* and *COHA* distribution frequencies.

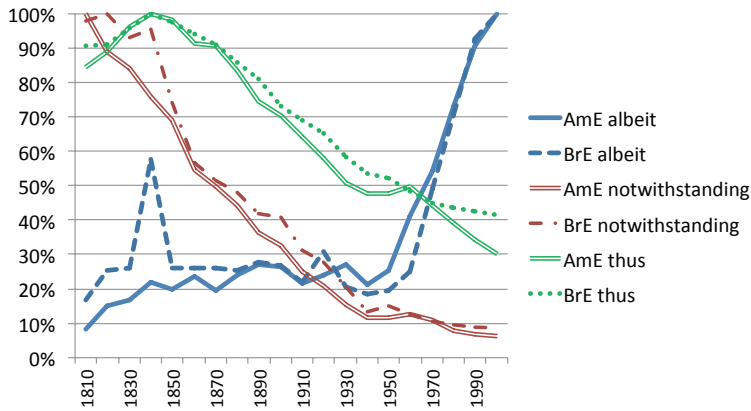

Figure 10. *Google* frequency distributions, 1810–2000.

notwithstanding, however, appears to be out of step with its reduced role in written language in general.[24]

Further comparison may now be made with material from the newly-released Google corpus from 1810 to 2000 (actually two parallel corpora of AmE and BrE, respectively). The vast amount of material available, with text masses in the billions of words, would seem to imply

[24] A word of caution about the 1920s issues of *Time*: a number of other searches indicate that the 20s was a period when the magazine was finding its level of readership, and is an atypical decade; e.g its use of *shall* dropped by 50% from the 20s to the 30s; no other decade-to-decade comparison indicates such a major shift for *shall*.

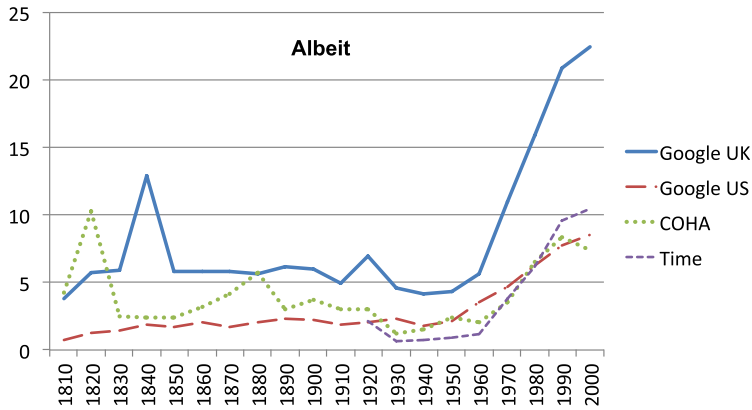

Figure 11. Comparison 1810–2000 for the three major diachronic corpora, plus *Time* (from 1923 to 1996).

a major improvement in quality of data, but in his comparison website, its creator Mark Davies remarks, "All three resources—Google Books (both versions) and COHA—give nearly the same results for [word and frequency] searches. The 400 million words in COHA is probably sufficient for nearly all searches of individual words and phrases."[25] It is consequently interesting to compare the *COHA* and *Google US* material for this period, as seen in Figure 9.

For *thus*, and even more so for *notwithstanding*, the fit appears to be quite good, but much less so for *albeit*, at least during several periods of the 19th century, whereas the 20th century appears to be a good fit. The explanation lies in the data mentioned in note 23, above: a single book in the 1820s accounts for 77% of the instances in the 1820s *COHA* data. This is a useful reminder when there are startling shifts in the data between adjacent periods.

Within Google, one can compare the relatively massive AmE corpus with the five times smaller BrE corpus for our three items (Figure 10). As the graph shows, the fit over two centuries is astonishingly good, with only a minor blip in the figure for *albeit* in BrE in the 1840s to disturb the picture.[26] The fit is also relatively good with the *COHA* data, and bears out Davies' prediction for both *thus* and *notwithstanding*.

[25] http://googlebooks.byu.edu/compare-googleBooks.asp, accessed March 10, 2014.

[26] The possibility that *albeit* was an OCR error for *(Prince Consort) Albert* was explored, but the readings are accurate; the Google image was quite clear in all instances checked—the BrE 1840s texts had a penchant for *albeit*.

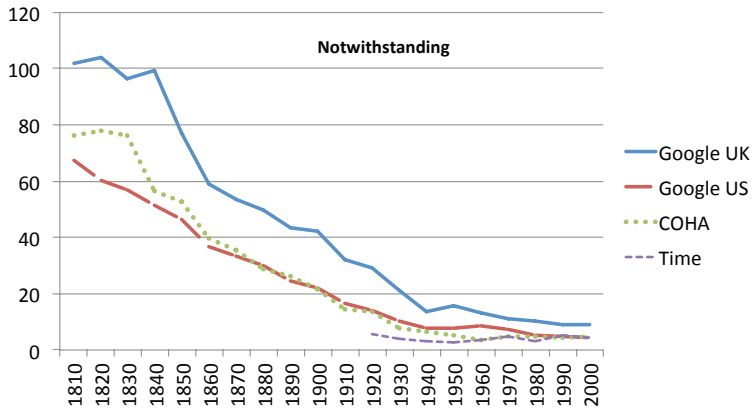

Figure 12. Comparison 1810–2000 for the three major diachronic corpora, plus *Time* (from 1923 to 1996).

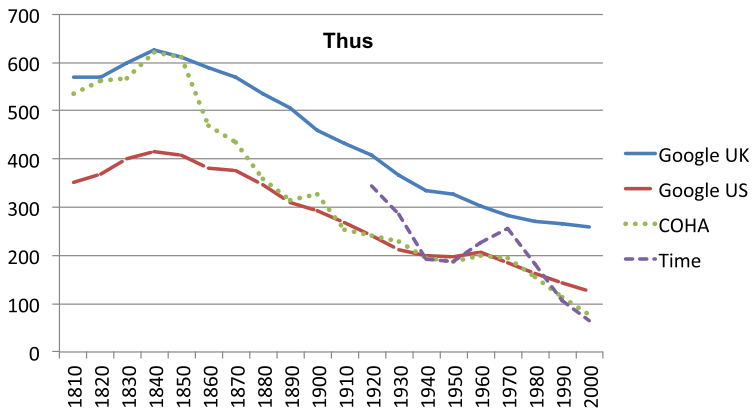

Figure 13. Comparison 1810–2000 for the three major diachronic corpora, plus *Time* (from 1923 to 1996).

The larger-sized Google data on *albeit* shows far less variation per decade, but from the 1950s onward shows the same sharp, consistent rise in its use.

Finally, combining these large corpora and looking at the four-corpora data for each word, i.e. the words in *Google US/GB, COHA* and *Time*, converted to frequencies per 1 M words, we see the same patterns in even stronger relief. The numbers are not fully consistent, particularly in the earlier decades, but what clearly emerges is that the

British frequencies are consistently a bit higher. This can of course indicate either that these three items are more markedly formal in AmE (and thus less used), or merely that the distribution of domains (and hence of formality) differs between BrE and AmE data in the corpora. Given that all three AmE corpora are in striking agreement for the 20th century, it seems probable that there is a difference between BrE and AmE invoved. Furthermore, whether BrE or AmE, it is clear that *notwithstanding* and *thus* are dropping in use (except for *thus* in scientific texts). As for *albeit*, the frequency per million is much lower than for most logical and attitudinal connectors, but rumors of its demise are clearly exaggerated (it is currently actually more frequent in AmE than *notwithstanding*). The key to its revival is to be found in the distributional data from the *OCE* corpus: the arts and news are the major domains for its use.

It seems reasonable to argue that earlier style mavens such as Fowler and Gowers (1965) were particularly aware of domains such as the arts, and less interested in stylistic uses in e.g. the sciences, so that it would not be surprising that Gowers should become aware of the revival of *albeit*—but the chronology is slightly wrong: the nadir in its use appears to have been the period 1930–1960, with the real rise taking place after 1965, the publication date of his revision of Fowler; moreover, it had always been more in use in BrE (his dialect) than in AmE, something that is even more clear today.[27]

6. Some final words

Since this paper has been written with a specific Stockholm Metaphor Festival scholar in mind, it may be worth mentioning that a review of the 2006 to 2010 articles from the Festival produced about 170,000 words in English (other languages discounted), and per million statistics of 22/M for *albeit*, 6/M for *notwithstanding* and 750/M for *thus*, figures which are quite close to the 2000 data for *albeit* and *notwithstanding*, but just about double the BrE figure for *thus*, which is not surprising, given that the majority of these papers were linguistically oriented, and follow the scientific pattern seen in e.g. Figure 4.

[27] Yet another recently-released web corpus, the *Corpus of Global Web-Based English* (also from BYU), reports post-2000 frequencies of 20.24 per M for BrE and 12.61 per M for AmE, which is in agreement with the other corpus data. Released in 2013, *GloWbE* contains 1.9 B words from the entire English-speaking world. See http://corpus2.byu.edu/glowbe.

It turns out that by and large, our corpus data from multiple corpora tends to be in agreement, although since these corpora are constructed with different metainformation, they will provide us with different types of information about matters such as style and domain. Even so, they are clearly of great help in enriching our picture of English, not to mention their forming the basis for the information found in our modern learner dictionaries. If there is one specific matter which the present corpus data suggests, it is that the Academic Word List needs to be re-examined, based on more extensive corpus data (as Gardner & Davies 2013 does).

Corpus data can of course only produce (massive) descriptive evidence of what people are doing with English at any given time, so that there will always be room for stylists and language police who wish to impose prescription upon us—even if they would do well to be far more heedful of the complexity of language, in particular the different domains within which language operates. All our data seems to indicate that while *notwithstanding* has indeed become relatively infrequent, *thus* has found a niche in scientific writing, where it seems to be flourishing. But the most astonishing of our triad is clearly *albeit*, which has returned from the moribund, to the joy of those who rejoice at seeing quirks of syntax live on, albeit in frozen form.

References

American Heritage Dictionary of the English Language. (1970). New York: American Heritage Publishing & Houghton Mifflin.

Biber, D., Conrad, S. & Leech, G. (2002). *Longman Student Grammar of Spoken and Written English*. Harlow, UK: Pearson Education.

Cambridge Advanced Learner's Dictionary. (2008). (3rd ed). Cambridge: Cambridge University Press.

Cambridge Dictionary of American English. (2000). Cambridge: Cambridge University Press.

Carter, R. & McCarthy, M. (eds) (2006). *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Celce-Murcia, M. & Larsen-Freeman, D. (1983). *The Grammar Book: An ESL/EFL Teacher's Course*. Rowley, Mass: Newbury House.

Clark, S. & Pointon, G. (2009). *Words: A User's Guide*. Harlow, UK: Pearson Education.

Clear, J. (1988). Trawling the Language: Monitor Corpora. M. Snell-Hornby (ed.), *ZuriLEX '86 Proceedings*. Tübingen: Francke.

Collins COBUILD English Language Dictionary. (1987). (1st ed.) London: Collins.

Collins COBUILD English Usage. (1992). Glasgow: HarperCollins.

Collins, F.H. (1956). *Authors' and Printers' Dictionary*. (10th ed.) London: Oxford University Press.

Copperud, R. (1964). *A Dictionary of Usage and Style*. New York: Hawthorn Books.

Corpus Of Contemporary American English (COCA), http://www.americancorpus.org/.

Corpus Of Historical American English (COHA), http://corpus.byu.edu/coha.

Coxhead, A. (2000). A New Academic Wordlist. *TESOL Quarterly*, 34(2), 213–238.

Curme, G.O. (1983 [1931]). *A Grammar of the English Language*, Vol. II: *Syntax*. Essex, Conn: Verbatim.

Davies, M. http://googlebooks.byu.edu/compare-googleBooks.asp, accessed Mar. 10, 2014.

Elfstrand, D. & Gabrielsson, A. (1960). *Engelsk Grammatik för universitet och högskolor*. (4th ed.) Stockholm: Läromedelsförlagen.

Elster, C.H. (1999). *The Big Book of Beastly Mispronunciations*. New York: Houghton Mifflin.

Estling Vannestål, M. (2008). *A University Grammar of English with a Swedish Perspective*. Lund: Studentlitteratur.

Fisher, J.H. (ed.). (1977). *The Complete Poetry and Prose of Geoffrey Chaucer*. New York: Holt, Rinehart and Winston.

Fowler, H. (1926). *A Dictionary of Modern English Usage*. Oxford: Oxford University Press. See also Gowers (1965).

Fowler, H. & Fowler, F. (1930). *The King's English*. (3rd ed.) Oxford: Oxford University Press.

Gardner, D. & Davies, M. (2013). A New Academic Vocabulary List. *Applied Linguistics*, *doi: 10.1093/applin/amt015*, published Aug. 2, 2013.

Gowers, E. (1965). *Fowler's Modern English Usage*. (2nd ed.) Oxford: Oxford University Press. See also Fowler (1926).

Greenbaum, S. & Whitcut, J. (1988). *Longman Guide to English Usage*. Harlow: Longman.

Hundt, M. (1998). It is Important that This Study (Should) Be Based on the

Analysis of Parallel Corpora: On the use of mandative subjunctive in four major varieties of English. H. Lindquist et al. (eds) *The Major Varieties of English*, Papers from MAVEN 97, Växjö: Växjö University,159–175.

Jespersen, O. (1940). *A Modern English Grammar on Historical Principles, V: Syntax* (Fourth Volume). Copenhagen: Ejnar Munksgaard.

———. (1949). *A Modern English Grammar on Historical Principles, I: Sounds and spellings*. Copenhagen: Ejnar Munksgaard.

Johnson, S. (1783). *A Dictionary of the English Language…Abstracted from the Folio Edition*. London.

———. (1799). *A Dictionary of the English Language*. (8th ed.) London.

Leech, G. & Svartvik, J. (2002). *A Communicative Grammar of English*, 3rd ed. Harlow, UK: Pearson Education.

Longman Dictionary of Contemporary English. (2005). (4th ed.) Harlow, UK: Pearson Education.

Longman Dictionary of Contemporary English. (2009). (5th ed.) Harlow, UK: Pearson Education.

McEnery, T., Xiao, R & Tono, Y. (2006). *Corpus-based Language Studies: An advanced resource book*. London: Routledge.

Macmillan English Dictionary for Advanced Learners. (2007). (2nd ed.) Macmillan Education: Oxford.

Macmillan English Dictionary for Advanced Learners of American English. (2004). Macmillan Education: Oxford.

Macquarie Dictionary. (1997). (3rd ed.). Sydney: The Macquarie Library.

Manual of Style, A. (1969). (12th ed.) Chicago: University of Chicago Press.

Mencken, H.L. (1936). *The American Language*. (4th ed.) New York: Alfred A. Knopf.

———. (1948). *The American Language, Supplement Two*. New York: Alfred A. Knopf.

Merriam-Webster's Collegiate Dictionary. (2003). (11th ed.) Springfield, Mass: Merriam-Webster.

Merriam-Webster's Dictionary of English Usage. (1994). Springfield, Mass: Merriam-Webster.

Minugh, D. (2002). "Her COLTISH Energy Notwithstanding": An examination of the adposition *notwithstanding*. L. E. Breivik & A. Hasselgren (eds) *Language and Computers: From the COLT's Mouth… and others'*. Amsterdam: Rodopi, 213–29.

New Oxford Dictionary of English. (1998). Oxford: Oxford University Press.

Oshima, A. & Hogue, A. (2006). *Writing Academic English*. (4th ed.) Harlow, UK: Pearson Education.

Oxford Advanced Learner's Dictionary (ALD). (1963). (2nd ed.) Oxford: Oxford University Press.

Oxford Advanced Learner's Dictionary (ALD). (2005). (7th ed.) Oxford: Oxford University Press.

Oxford Advanced Learner's Dictionary (ALD). (2010). (8th ed.) Oxford: Oxford University Press.

Oxford American Dictionary of Current English. (1999). New York: Oxford University Press.

Oxford English Dictionary. (1933 [1928]). (1st ed.) Oxford: Oxford University Press.

OED Online. Accessed Feb. 25, 2014, at www.oed.com.

Poutsma, H. (1929). *A Grammar of Late Modern English*. *Part I: The sentence*, 2nd ed. Groningen: P. Noordhoff.

Quirk, R., Greenbaum, S, Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Rissanen, M. (2002). On the Development of Concessive Prepositions in English. A. G. Fischer, G. Tottie, & H. M. Lehman (eds)*Text Types and Corpora : Studies in honour of Udo Fries*. Tübingen: Gunter Narr, 191–203.

Schmid, H.-J. (2000). *English Abstract Nouns as Conceptual Shells: From corpus to cognition*. Berlin: Mouton de Gruyter.

Svartvik, J. & Sager, O. (1996). *Engelsk universitetsgrammatik*. (2nd ed.) Stockholm: Almqvist & Wiksell.

———. (1977). *Engelsk universitetsgrammatik*. Stockholm: Esselte Studium.

Swan, M. (2005). *Practical English Usage*. (3rd ed.) Oxford: OUP.

Weber, B. (2010). *Sprachlicher Ausbau: konzeptionelle Studien zur spätmittelenglischen Schriftsprache*. Frankfurt am Main: Peter Lang.

Webster's New Collegiate Dictionary. (1993). (10th ed.) Springfield, Mass: Merriam-Webster.

Webster's New International Dictionary of the English Language. (1941). (2nd ed.) Springfield, Mass: Merriam.

Webster's Ninth New Collegiate Dictionary. (1988). Springfield, Mass: Merriam-Webster.

Wells, J. (ed.). (2008). *Longman Dictionary of English Pronunciation*. (3rd ed.) Harlow, UK: Pearson Education.

West, M. (1953). *A General Service List of English Words*. London: Longman, Green.

Corpora

British National Corpus, http://corpus2.byu.edu/bnc/.

British National Corpus, http://bncweb.lancs.ac.uk/bncwebSignup/user/login.php.

Brown Corpus (Brown University Standard Corpus of Present-Day American English), available via ICAME: http://icame.uib.no/newcd.htm.

Corpus of Global Web-Based English. http://corpus2.byu.edu/glowbe.

Dutch Web Corpus. 111 M words, accessed via SketchEngine, https://www.sketchengine.co.uk/.

Frown Corpus (Freiberg-Brown Corpus of American English), available via ICAME: http://icame.uib.no/newcd.htm.

FLOB Corpus (Freiburg-LOB Corpus of British English), available via ICAME: http://icame.uib.no/newcd.htm.

LOB Corpus (Lancaster-Oslo-Bergen Corpus of British English), available via ICAME: http://icame.uib.no/newcd.htm.

Oxford English Corpus (with BiWeC), 1736M words, accessed via SketchEngine, https://www.sketchengine.co.uk/.

Time Corpus. http://corpus.byu.edu/time/.

Wordbanks Online, http://www.collinslanguage.com/content-solutions/wordbanks.