This is a draft of a chapter that has been accepted for publication by Oxford University Press in the following book. Estimated publication date October 2017.

Hayes, Mary & Allison Burkette (eds.). 2017. *Approaches to teaching the history of the English language: Pedagogy in practice*. Oxford University Press.

https://global.oup.com/academic/product/approaches-to-teaching-the-history-of-the-english-language-9780190611057?cc=gb&lang=en&

Word classes in the history of English

David Denison
University of Manchester

Introduction

Word classes or parts of speech (POS) may seem a rather banal aspect of language, but they lead us into some intriguing aspects of language history. It's probably an advantage of this topic that most people will have at least some idea of which words are nouns, which adjectives, and so on, though there may be some haziness about how such identifications are arrived at. If verbs are 'doing words', how come *demolition* is a noun? If adjectives are 'describing words', how come *very* in *the very idea* is an adjective? – to which the answer, of course, is that notional definitions like those are no longer linguistically respectable. Most modern linguistic approaches privilege form over semantics. Word classes are primarily determined by distributional properties – that is, the kinds of syntactic environment a word can fit into – and, where appropriate, the potential for inflectional endings.

Any respectable introduction to language will discuss suitable tests for word class, though we should be aware that semantic properties have a way of sneaking back in when we look more closely at how language works. Most research in the various formal grammatical frameworks, just like most school-level analysis, assumes that every single word in every grammatical sentence belongs to one and only one part of speech. Even if (like me) you don't accept the assumption, it's still a useful abstraction that allows us to make generalisations about whole classes of words.

In this chapter we will look at research that students can carry out with the aid of, or sometimes actually in search of, part-of-speech information. And students **should** be carrying out research. In the history of the English language, just as with scholarship generally, what teachers or textbooks say should not be taken for granted. Experts can be wrong, or their facts out of date. Even with valid assertions, it is unrewarding for a student simply to parrot back to their teacher the examples they have been given. Where possible, students should be finding their own data to help make sense of – and test critically – the surveys or generalisations or hypotheses on offer. If they can do this from time to time, it's more fun all round, and everyone (instructors too!) learns more.

Tagged corpora

What data sources can students use? Certainly conversations around them, newspapers, film and TV dialogue, novels, emails, texts … any of these can provide useful examples, as appropriate, of change in progress or change in the past. It's surely a good idea to be alert to language around us, and

there's always the chance of discovering something useful or even a usage previously unnoticed. But the language we encounter casually is usually too haphazard to help with a specific research question. Incidentally, quite apart from problems of inefficiency, there are other risks in relying on casually observed data, as Arnold Zwicky has observed:

> […] the Recency Illusion, the belief that things YOU have noticed only recently are in fact recent. This is a selective attention effect. Your impressions are simply not to be trusted; you have to check the facts. Again and again -- retro *not*, double *is*, speaker-oriented *hopefully*, split infinitives, etc. -- the phenomena turn out to have been around, with some frequency, for very much longer than you think. It's not just Kids These Days.
>
> Professional linguists can be as subject to the Recency Illusion as anyone else. […] Another selective attention effect, which tends to accompany the Recency Illusion, is the Frequency Illusion: once you've noticed a phenomenon, you think it happens a whole lot, even "all the time". Your estimates of frequency are likely to be skewed by your noticing nearly every occurrence that comes past you. People who are reflective about language -- professional linguists, people who set themselves up as authorities on language, and ordinary people who are simply interested in language -- are especially prone to the Frequency Illusion. (Zwicky 2005)

We needn't stop listening and looking, but it is well to bear Zwicky's warnings in mind.

A standard way of garnering data is to search a **corpus** (plural *corpora*), a systematic collection of linguistic material, usually, these days, in electronic form. In principle a corpus search allows us to make factual statements about the relative frequency of a particular usage at a particular time and provides genuine rather than invented examples of relevant usage. There are a number of corpora of English available over the web, some of which are freely available, possibly after registration, others requiring a paid individual or institutional subscription.

Corpora differ in the linguistic **mark-up** they offer. It is common for words to be **POS-tagged**, where each word is associated with a tag that identifies its word class (and perhaps more). The set of classes used, the **tagset**, may be akin to the conventional parts of speech but usually reflects a rather more fine-grained classification. For example, the CLAWS-5 tagset used in the British National Corpus (BNC) has 57 simple word class tags, and some others have many more still. Thus four kinds of noun word are differentiated in the CLAWS-5 tagset, as shown in Table 1.

NN0	Common noun, neutral for number (e.g. *aircraft, data, committee*)
NN1	Singular common noun (e.g. *pencil, goose, time, revelation*)
NN2	Plural common noun (e.g. *pencils, geese, times, revelations*)
NP0	Proper noun (e.g. *London, Michael, Mars, IBM*)

Table 1: Kinds of noun the CLAWS-5 tagset

Even this short extract from the tagset illustrates some items of interest. The word *data* may be singular or plural ('The data is …' or 'The data are …'), a fair representation of usage in a 100-million-word corpus but not to the liking of purists conscious of its etymology in a Latin plural. A collective noun like *committee* may take either singular or plural agreement in British English ('A committee which is …' vs. 'A committee who are …', the latter less likely in American English). Concrete and abstract nouns are not distinguished, as the tagging is distributional rather than meaning-based. On a practical note, tags for subclasses of a particular part of speech are usually grouped together

alphabetically so that a wildcard such as the asterisk (= 'any 0 or more characters') can be used to capture related tags in a single search, e.g. in the list above, **N*** for all nouns or **NN*** for all common nouns.

Tagged corpora make searching much more efficient. If you are interested in the modal verbs *can* and *may*, you can exclude the noun and verb *can* associated with food preservation and the noun *May* 'fifth month', or vice versa. If you are interested in *impact* as a verb rather than a noun, you can search more efficiently with tags, and for *impact* as a transitive verb you can search for it followed directly by a pronoun, determiner or noun. Note that such a rough-and-ready search strategy will give you some false positives on the one hand (the so-called **precision** of the search is sub-optimal) and will miss some valid data on the other (sub-optimal **recall**). Ideally both recall and precision should be as close to 100% as possible, but compromises are normal.

A corpus that is not just tagged but also **parsed** – that is, with each sentence given a structural analysis – would allow us to look for structures rather than just strings of words and/or tags. Then we could look directly for verbs that take direct objects. Depending on the parsing adopted, we could search for relative clauses, say, or pronoun subjects. Note that nominative forms of pronouns – which can be searched for without parsing – are not the same as subjects, as the following examples show. In (1), accusative *me* is subject of a verb; in (2), nominative *I* is not subject.

(1) a. And then maybe you and me could go to Hawaii or something (2006, COCA)
    b. So what do you want me to do? (2015, COCA)
(2) and [they] presently guessed that it was I (1832, COHA)

For reasons of practicality, however, this chapter will not discuss parsed corpora further. Much useful work can be accomplished simply by searching for text strings, and more still with the possibility of using POS tags as well.

Some corpora have a rudimentary semantic tagging, but even if not, if you were interested, say, in synonyms for telling a joke, you could try searching for a verb before a phrase like *a joke* to get such verbs as *attempt, crack, make, play, repeat, share, tell*. (And/or you could search the definitions of the *OED*.) Furthermore, most tagged corpora are **lemmatised**, so that all forms of a single lexical item can be picked up in a single lemma search, *cat/cats* or *good/better/best* or *buy/buys/buying/bought*. Be aware of lemmatisation in the way a corpus is set out. Usually, for example, the contracted negative *n't* is separated from its verb, contrary to conventional spelling. This allows one to search for both forms of negative particle, *not* and *n't*, with the same tag. Thus in COHA, a search for **would n't** or more generally ***.[VM] *.[XX]** (modal verb followed by *not* or *n't*) will produce thousands of hits, but **wouldn't** finds none.

There is much more to be said about corpus searching if you want to take it further, for example with more complex searches involving alternatives or wildcards or regular expressions. Details vary with corpus and search engine – the interface through which you put your queries. There are many textbooks; Hoffmann et al. (2008) is excellent, of general utility even though designed for work with a particular corpus.

Note that tagging of large corpora is done by software programs ('taggers'), and a proportion of tags, typically 4-6% or so, will inevitably be incorrect. Where and how tagging goes wrong is also of interest, as we shall see.

Some corpora and other data sources

If at all possible, students should have access to the online *Oxford English Dictionary* (*OED*). I will also illustrate this chapter from corpora made conveniently available by Mark Davies of Brigham Young University, especially his Corpus of Historical American English (COHA) and Corpus of Contemporary American English (COCA). The same search engine is used for a number of corpora stored at BYU, present-day and historical, including some really huge ones. The British National Corpus (BNC) is one of those put out by BYU, and also, with arguably a more manageable search engine, at Lancaster University.

Another source is the web, good for up-to-the-minute usage but presenting various challenges: unknown size of 'corpus', which makes statistics difficult; ever-changing material, so that results are irreproducible; often unknown provenance, leaving doubt as to the variety of English in question. For a discussion of the problems of internet source material and how to mitigate them, see for instance (Keller, Lapata & Ourioupina 2002, Kilgarriff & Grefenstette 2003, Véronis 2005, Hundt, Nesselhauf & Biewer 2007). A convenient search engine for internet material is WebCorp, better than using an ordinary content-oriented search engine like Google or Bing.

Other BYU corpora include Google Books, a Time Magazine archive, and – soon – a new version of Early English Books Online (EEBO, all books published in England 1453-1700) and Hansard (records of UK parliamentary proceedings). Here are some other suggestions, by no means a complete list, some of them referenced at the end of the chapter. From the Survey of English Usage at University College London you can purchase ICE-GB (1m words of current British English, written and spoken) and the Diachronic Corpus of Parsed Spoken English (DCPSE, two matched corpora of spoken English from the 1960s and the 1990s, nearly 900k words). These are very carefully produced corpora with a lot of spoken material, fully tagged and parsed. The search engine is comprehensive but idiosyncratic, and the corpora are not free.

The Corpus of Late Modern English Texts (CLMET3.0) is 34m words 1710-1920, divided into three subperiods of 70 years each, tagged. Four carefully produced corpora from the Universities of Pennsylvania and York totalling 5.3m words cover the periods from Old English through to 1914 (YCOE, PPCME2, PPCEME, PPCMBE). They are fully tagged and consistently parsed, but the search engine, CorpusSearch, takes serious training to master and is probably not suitable for any but the most advanced classes. A Representative Corpus of Historical English Registers 3.2 (ARCHER) is a multi-genre corpus of 3.3m words 1600-1999. The project is coordinated from Manchester and available in a tagged version at Lancaster and a tagged and parsed version at Zurich. The Old Bailey Corpus (OBC) is some 134m words of transcribed London court proceedings 1720-1913, a remarkable resource.

Searching a corpus

To make good use of a corpus you should have some idea of its coverage of material – periods, genres, media, etc. – and of the tagset used. A good corpus will offer documentation that explains the procedures followed in collecting, transcribing and marking it up. Online search engines differ in features and screen interface. Some corpora are available for use on your own computer rather than online, whether with their own search engine or by means of a general-purpose concordance program.

The online search engine for the BYU corpora has its idiosyncrasies. (Fortunately a new interface is promised for May 2016; what follows may therefore need revision.) For COHA you enter a search word or string, optionally specifying word classes or periods or genres, and click **SEARCH**. The **LIST** search tabulates in decreasing order of frequency up to 100 forms that satisfy your search, with detailed numbers by decade between 1810-19 and 2000-2009. All or a selection of the hits can be displayed, a screenful at a time. (Note that the collections for the first couple of decades are rather patchy and unreliable.) The **CHART** search produces a decade-by-decade histogram. The **KWIC** (= key word in context) search shows the number of hits and displays them in context with a handful of words on either side of the search item colour-coded to indicate the POS: nouns in turquoise highlighting, verbs magenta, adjectives green, prepositions yellow, adverbs brown, function words grey. The **COMPARE** search seems to show for two different search strings or periods or genres the frequency of collocates (words in the contexts of the hits).

Derivation

One way that new words get made is by **derivation**, the addition of an affix. The usual effect of a **suffix** (an affix attached at the end) is to change the word class, as in *sketch*$_{N}$ > *sketchy*$_{Adj}$ or *tender*$_{Adj}$ > *tenderise*$_{V}$, but not always, as in *rocket*$_{N}$ > *rocketry*$_{N}$ (different **kinds** of noun, however, in this case). Via derivation we get families of words like *abstain*$_{N}$ –*abstinent*$_{Adj}$ – *abstention*$_{N}$, *permit*$_{V}$ – *permission*$_{N}$ – *permissive*$_{Adj}$, *strong*$_{Adj}$ –*strongly*$_{Adv}$ –*strength*$_{N}$ – *strengthen*$_{V}$, *invite*$_{V}$ – *inviting*$_{Adj}$ – *invitation*$_{N}$, *attend*$_{V}$ – *attendance*$_{N}$ – *attention*$_{N}$ – *attentive*$_{Adj}$, and so on.

Semantic change may affect whole families of words, as when the *strong* family is extended from describing human muscle-power to, say, the force of an argument – or coffee. Sometimes, though, different members of the family go different ways. Thus the verb *abstain* is most typically used in two senses: to forego something (usually pleasurable), and later, to refrain specifically from voting. The noun *abstention* eventually follows the verb into both of these contexts, but the adjective *abstinent* has apparently never been used of someone who doesn't vote. There is endless scope for research in the *OED* into the semantic histories of word families. For two examples of relevance here, see Denison (submitted) on the negative connotations of recent senses of *sketchy* in American English vs. the entirely neutral *sketch* (noun or verb), and on the frequently negative connotations of *discrimination* and *discriminate* vs. the positive connotations of the adjective *discriminating*.

To return to word-formation, an interesting exercise is to look for trends in the use of different suffixes. For example, choose some suffixes that form abstract nouns, whether general ones like –*ment* and *-ity* or more specialised ones like *-age, -ance/-ence, -ation, -hood, -ism, -ship*, etc. For each such suffix, use the *OED Online* to find new formations, with dates of first use from which you can calculate numbers in a given decade or half-century. To do this, try the **Advanced search** facility, specifying

- Search in **Entries**
- search string **???*age** = wildcards followed by the suffix in question
  a minimum of 6 letters should avoid monosyllables like *phage*
- in **Headword**
- Part of speech **Noun**
- Date of entry **1800-** or **1950-1999** for example (don't shorten ranges as 1950-99)

If you choose a longish timespan or leave the date filter blank, you can then click **Timeline** to see a convenient graphic representation (histogram) in numbers by 50-year periods, though more careful work would discount irrelevant hits which come up in the sample search described above, such as the entries *road rage* or *monophage*, as well as borrowings like *ménage à quatre*.

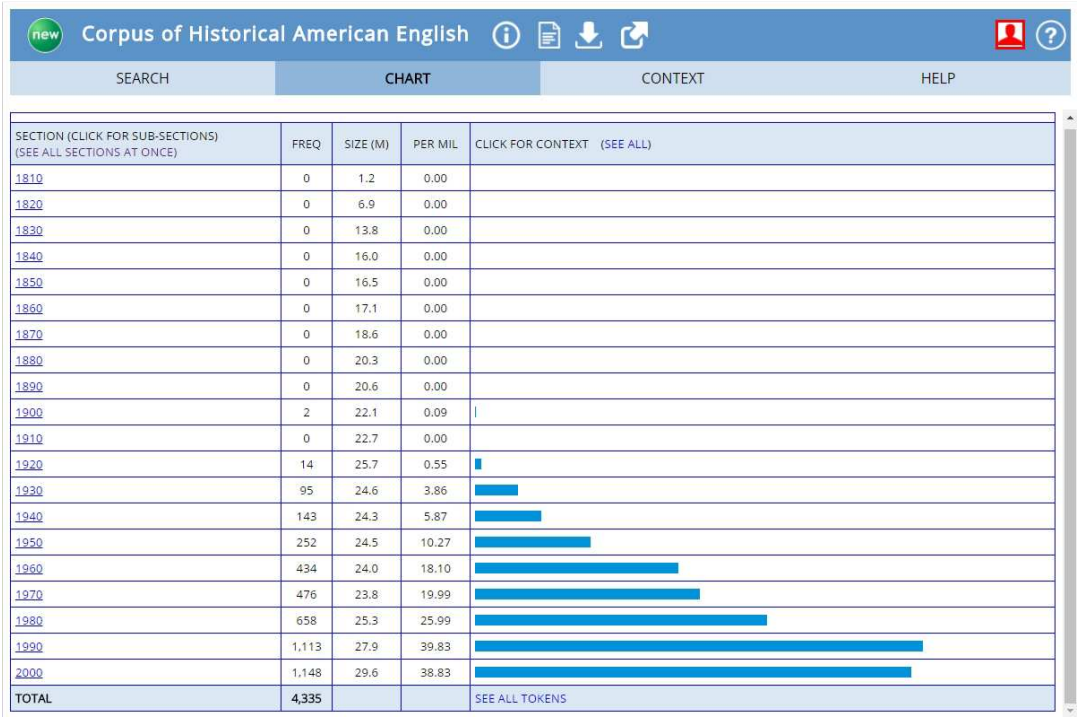
In addition to the dictionary you can use a large corpus to get a better idea of frequency of use at a given time: COHA or CLMET3.0 or EEBO or (with more caution) *Google Books*, or even the quotation database of the *OED* treated as a corpus. For example, there have been student fads in recent decades in the US and Britain where non-standard words in *-age* have been coined humorously despite the prior existence of words which should have blocked them (Denison 2008: 208-10). Something similar is illustrated in (3):

(3) He makes up words out of his head. At lunchtime, he says it's time for eatage. When we get head call, it's time for pissage. Lights-out, and it's time for sleepage. (2003, COCA)

Here, collected from the *OED* (s.v. *–age* suffix), are some recent *–age* words (i) first found on or after 1900; (ii) consisting of a single word; (iii) not obviously French borrowings; (iv) which happen to occur in COCA:

coverage|creepage|ecotage|footage|gallonage|megatonnage|narratage|ohmage|plottage|plussage|signage|wordage

An upright bar **|** in a search string separates alternatives. It did not seem necessary to add **.[nn*]** to each item to limit the part of speech to nouns, though in any case that would have exceeded some machine limit and caused an error, as also would the desirable improvement of enclosing each word in square brackets to search for lexemes, thus getting possessives and plurals as well as the singular nouns. The search string indicated above gives the chart shown in Figure 1.


Corpus of Historical American English

SEARCH | CHART | CONTEXT | HELP

SECTION (CLICK FOR SUB-SECTIONS) (SEE ALL SECTIONS AT ONCE)	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT (SEE ALL)
1810	0	1.2	0.00	
1820	0	6.9	0.00	
1830	0	13.8	0.00	
1840	0	16.0	0.00	
1850	0	16.5	0.00	
1860	0	17.1	0.00	
1870	0	18.6	0.00	
1880	0	20.3	0.00	
1890	0	20.6	0.00	
1900	2	22.1	0.09	
1910	0	22.7	0.00	
1920	14	25.7	0.55	
1930	95	24.6	3.86	
1940	143	24.3	5.87	
1950	252	24.5	10.27	
1960	434	24.0	18.10	
1970	476	23.8	19.99	
1980	658	25.3	25.99	
1990	1,113	27.9	39.83	
2000	1,148	29.6	38.83	
TOTAL	4,335			SEE ALL TOKENS

Figure 1: Twelve nouns in with *–age* suffix in COHA

Sadly the recent word *flamage* 'vitriolic argument or ranting [online]' does not occur in COHA or COCA, though there are 7 examples in GloWbE.

If students are aware of derivational innovations in their own circles, get them to investigate the distribution or history of similar patterns. Or try plotting the spread of the *–gate* suffix (from *Watergate*, 1972-3) for scandals and cover-ups more generally.

Conversion

One property that somewhat distinguishes English from 'Standard Average European' is the ease of **conversion** from one word class to another without any overt change of form, allowing the word to be used in either class. Viewing conversion (a technical term) as the addition of an affix which happens to be zero explains the alternative term, **zero-derivation**. The commonest case is transfer between noun and verb, in either direction (historically, that is). The words *bottle, fashion* and *hammer*, which intuition says must have started as nouns, may be used as verbs. The verbs *dive* and *fling* may be used as nouns. See Nevalainen (1999: 424-30). Alteration of stress (compare *import*$_V$ and *import*$_N$) can be included in 'zero'.

An obvious contributor to the ease of conversion in English is the loss of much inflectional marking in the language, especially in the centuries between late Old English and early Modern English, since one sure-fire way of distinguishing parts of speech in, say, the classical languages is the different inflectional paradigms associated with verbs, nouns and adjectives. Our intuitions on which word class has historical priority must always be checked, for example in the *OED*; sometimes the first appearances are more or less simultaneous. Other conversions include adverbial particle to noun (*ups and downs*), adjective to noun (*musical, uniform*), adjective to verb (*big up, gross out*), particle to verb (*to out someone*).

Recent denominal verbs – or at least, conversions **thought** to be recent – arouse particular ire from prescriptivists. Business and computer Englishes favour such usages as *action* as a verb, *progress* as a transitive verb, and so on.

(4) Air Force Space Command is transitioning its current mix of medium and heavy lift expendable boosters to the new Evolved Expendable Launch Vehicle for space launch. (2003, COHA)
(5) Put your request into an email and I will action it. (2012, GloWbE)
(6) " We still want people to take risks and progress the sport, " says Gaffney. (2012, COCA)

Rather than use the suffixed derivative *invitation*, many speakers use the zero-derived noun *invite*, a usage criticised for example at https://grammarusage.wordpress.com/2012/04/03/invite-vs-invitation/. The *OED* shows that *invite* n. goes back to the seventeenth century and was used in the eighteenth, for example, by Fanny Burney:

(7) Every body Bowed, & accepted the invite but me..for I have no Notion of snapping at invites from the Great. (1778, *OED* s.v. *invite* n. 1)

A tagged corpus like COHA should allow students to find examples for discussion, but care is needed. Why does **[invite].[nn*]** (= the lexeme *invite* tagged with POS noun) get so many hits that are verbs? Conversely, when we search for **[action].[vv*]** (= the lexeme *action* with POS tag lexical verb), why

are almost all the hits actually nouns? It is instructive to try to work out why the tagger may have gone wrong in particular cases. For a nice example of how the **human** parser can go astray with a word that is ambiguously verb or noun, see Liberman (2016).

Stepwise change

Conversion is instantaneous and complete: the occurrences of *invite* in (7) can only be parsed as nouns. A different and little-noted type of transition is what I call stepwise change of word class (Denison 2013, submitted). We need a brief introduction.

Consider nouns and adjectives. Both can occur as modifiers of a head noun; the nouns *drug* and *health* in (8) and the adjective *serious* in (9) are all modifiers of the head noun *problem*:

(8) These individuals. mask an increasing drug problem and claim that it is necessary to solve the problem of drugs as a health problem. (2007, COHA)
(9) It became a serious problem. (2000, COHA)

Both word classes can occur as head of a predicative complement, and either a mass noun like *nonsense* or an adjective like *worthless* can occupy that position by itself:

(10)The claim is nonsense (1982, COHA)
(11)the claim is worthless (2002, COHA)

The underlined words in (10)–(11) are distinguishable in other contexts that show they are respectively noun and adjective, for example:

(12)a.	They sent us some nonsense/*worthless.	(only N possible)
b.	The claim is very/so *nonsense/worthless.	(only Adj possible)

The asterisk here is a notational convention for something ungrammatical.

The fact that nouns and adjectives share certain parts of their distributions, as illustrated in (8)-(11), means that those contexts leave the word class of their filler underdetermined or vague if there isn't clear evidence of the correct class elsewhere (as there was with (12)). And with certain words the external evidence is conflicting. A number of nouns in the last few decades have been gaining the possibility of functioning as full-fledged adjectives in addition to their earlier status, e.g. *ace, amateur, apricot, bandaid, cardboard, champion, core, corker, cowboy, designer, dinosaur, draft, freak, fun, genius, key, killer, landmark, luxury, niche, pants, powerhouse, rubbish, surprise, toy, Velcro* (list from Denison submitted). They don't become adjectives instantaneously. There is a period of transition, during which some evidence points to them not yet being adjectives, other evidence suggests that they are. And they retain the option of being used as nouns.

Early in the process the word can be coordinated with items that are definitely adjectives, which suggests but certainly does not prove that it has become an adjective itself:

(13)There's the expectation that it's this fun and easy thing that takes a few days. (2015, COCA)

Or the word occurs to the left of another modifier which is definitely an adjective:

(14)Key new evidence comes from two directions. (2005, COCA)

Even stronger evidence that a word has achieved adjectival status is when it can itself be modified by intensifiers like *so, very, too* and the like:

(15)They both knew that Molly was thinking about Fred's scars, which looked too amateur to be from surgery (1995, COCA)

Likewise the ability to occur as a post-modifier:

(16)["] What'd he say to her? " " Probably said something genius, like' Pardon me. " (2014, COCA)

And the clincher is when the transitioning word occurs as a comparative or superlative, an inflectional property unavailable to nouns:

(17)For many of the country's nichest and most powerful, a recession might be preferable to what it takes to get the economy moving again (1994, COCA)

Even a syntactic comparative or superlative will do:

(18)Organizational services are moving from the realm of being discretionary to being more key, more critical to an improved quality of life (1995, COCA)

Adjectival properties are added at different speeds and different periods for different words, and speakers at any one time may be more or less conservative. Thus most speakers would accept *something fun*; younger speakers would probably accept *so fun* where their older relatives would say *such fun*; but so far relatively few would accept the attested comparative *funner*. The examples above have not been selected to reflect the chronology of particular changes. The process for a given word can sometimes be tracked in dictionaries and especially historical corpora.

I suggest the word *key* in senses like 'essential, crucial' for an exercise of this kind. The process by which it develops an adjectival use has been discussed from several points of view by Leech & Li (1995), De Smet (2012: 621-8) and Kiparsky (2014: 000), amongst others. The usage is frequent enough and the transition close enough to completion among younger speakers for data to be easily found. Students can search for suitable diagnostic contexts, perhaps cautiously using the POS tags assigned to *key* to help them find pertinent examples. An additional exercise that will teach them a lot is to get them to explain why they think the tagger might have gone wrong in cases where they judge a sentence to have been mistagged. Examples will quickly appear if you search COHA for **[key]** as, respectively, noun, adjective or verb. The BNC at Lancaster with its BNCweb interface allows you to see not just the simple tags but the so-called **ambiguity tags** used when the tagger is unable to decide on an analysis with reasonable certainty. For example, out of 2280 occurrences of the form *keys* in BNC, 134 are tagged **NN2-VVZ** (= either plural of a common noun or 3 sg. pres. of a verb, but more likely the former), while 37 are tagged **VVZ-NN2**.

Another possibly stepwise change is seen in the past participles of verbs of mental disposition. They can behave either as adjectives or as verbs. Compare

(19)Jim was frightened by a spider.
(20)Jim was frightened of spiders.
(21)Jim was frightened.

In (19), *frightened* is a verb. The sentence is the passive of *A spider frightened Jim*, and the likely interpretation is of an instantaneous event; we could say *What happened to Jim was that he was frightened by a spider*. That would not be an appropriate transformation of (20), which describes a state. In (20), *frightened* is an adjective; it can be modified readily by *so* or *very*. Sentence (21) is ambiguous. This area of the language has seen a gradual change over the last four centuries. In Shakespeare's time it was normal to write *much interested, much concerned*, etc., the modifier *much* being appropriate for a verb. Over time it became increasingly acceptable, and nowadays more or less obligatory, to use modifiers like *very* that are appropriate to adjectives; see Denison (1998: 229-30). This kind of change could be tracked in COHA or ARCHER or PPCMBE, and students could be encouraged to think of other evidence that would show whether a participle was verbal or adjectival.

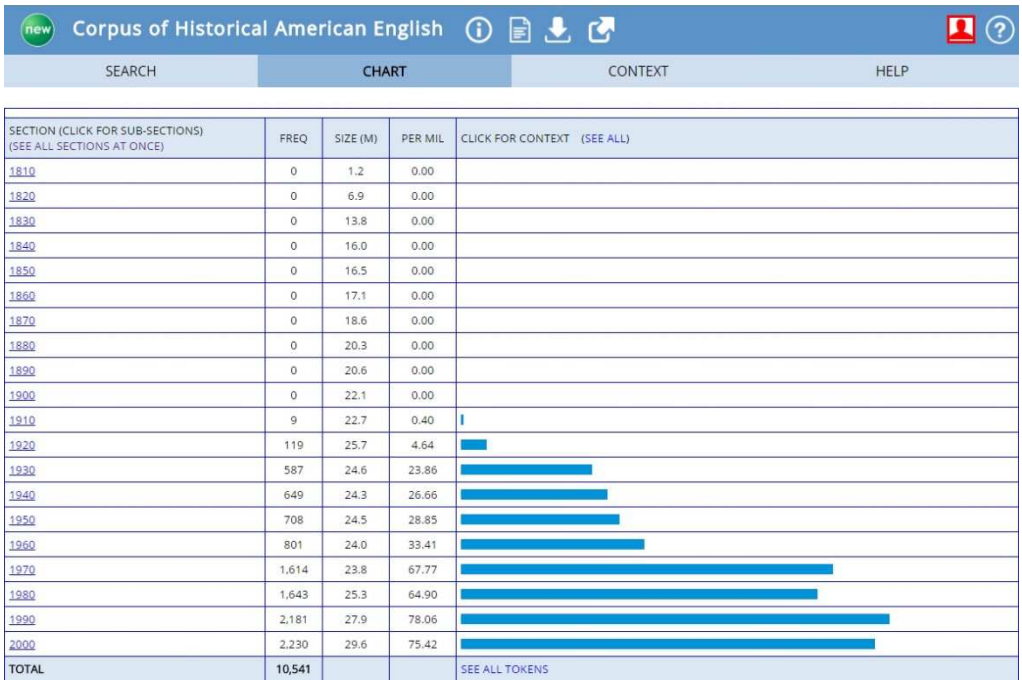
Cliticisation/grammaticalisation/unconventional spellings

Unstressed function words are often combined as **clitics** with the preceding word. Thus the strings *have to* and *has to* are typically pronounced [hæftə, hæstə] with devoicing of the final consonant of the verb, but only when they mean 'must':

(22)a. the crops they have to [hæv tə] sell 'the crops for sale that they possess'
b. the crops they have to [hæftə] sell 'the crops that they must sell

Now there is usually no sign in writing of the devoicing, though in fact haf(f) *to* (106×), *haf t* (4×), *haf ter* (2×) and *haf tuh* (2×) do occur in COHA, some no doubt representing dialect or foreign speech. However, informal pseudo-phonetic spellings like *gonna* for *going to*, *gotta* for *got to*, *kinda* for *kind of*, and so on are increasingly common as representations of colloquial speech and/or a non-standard speaker. All are interesting topics for research.

The informal spelling *gonna* is lemmatised as two 'words' (technically, two tokens) separated by a space in COHA, **gon na**. Its rise is shown in Figure 2:



SECTION (CLICK FOR SUB-SECTIONS) (SEE ALL SECTIONS AT ONCE)	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT (SEE ALL)
1810	0	1.2	0.00	
1820	0	6.9	0.00	
1830	0	13.8	0.00	
1840	0	16.0	0.00	
1850	0	16.5	0.00	
1860	0	17.1	0.00	
1870	0	18.6	0.00	
1880	0	20.3	0.00	
1890	0	20.6	0.00	
1900	0	22.1	0.00	
1910	9	22.7	0.40	
1920	119	25.7	4.64	
1930	587	24.6	23.86	
1940	649	24.3	26.66	
1950	708	24.5	28.85	
1960	801	24.0	33.41	
1970	1,614	23.8	67.77	
1980	1,643	25.3	64.90	
1990	2,181	27.9	78.06	
2000	2,230	29.6	75.42	
TOTAL	10,541			SEE ALL TOKENS

Figure 2: The form *gonna* (gon na) in COHA

(Incidentally, the infinitive marker with POS tag [TO] lemmatised as **na** occurs in both *gonna* and *wanna*.) Students can compare the frequencies of standard and informal spellings by date or genre; they can see whether particular subjects or lexical verbs promote the use of the non-standard spellings. To make a fair comparison with standard *going to* as an auxiliary of the future, it would make sense to ensure that the following word is a verb. In COHA one could search for **going to *.[vv*]** or perhaps **going to.[TO]** (= the infinitive marker *to* as opposed to the preposition), in order to avoid irrelevant verb phrases like *going to bed* or *going to London* which do not have equivalents with *gonna*. Then improve the search to catch split infinitives like (23) as well:

(23)a. She thought he was going to really hurt her (1978, COHA)
b. It's not too sharp, so it's probably gon na really hurt when it hits you (1999, COHA)

Another good topic is the strings *sort of, kind of* and *type of*, sometimes known as SKT constructions. What is of interest here it is not so much the phonetically similar contraction to *kind a / kinda / sorta*, etc. as the word classes involved, originally noun and preposition, respectively, **kind.[nn1] of.[io]**. Such a tagging is fully justified in a sentence like (24), less obviously appropriate in (25), but it would be completely inappropriate in (26) (where in fact the two words are 'ditto-tagged' as an adverb in COHA).

(24)in making a particular kind of wine (1834, COHA)
(25)" I don't generally do those kind of things, " answered Randal (1839, COHA)
(26)" Didn't you kind of hate to give it up? " (1918, COHA)

The pattern of (25), where the number of the determiner *these* or *those* agrees not with the SKT noun but the noun after *of*, is surprisingly old – and a prescriptivist bugbear. The pattern of (26), where the word after *of* is not a noun but a verb, adverb or adjective, has been seen with *sort* and *kind* for quite a long time and is just starting to appear with *type* as well; SKT-word + *of* seems to be lexicalised as a kind of adverb or hedge. British and American Englishes differ in their rates of use (see here Biber et al. 1999: 867-71) and perhaps their formality judgements of *sort of*. For a fuller discussion of the range of SKT constructions than I have space for here, see among many other works Denison (2002), Keizer (2007), Brems & Davidse (2010). Exercises here could include seeing just how frequent SKT constructions are and have been in the past, finding examples of the main kinds of use, discussing how taggers have (or have not) coped with them, comparing frequencies of *sort of* and *kind of* in British and American English or in different genres.

Finally in this section, consider the unreal conditional sentence in (27). Contracted versions, (28), would also be standard English, especially in fast speech.

(27)If I had known sooner, I would have gone to see you. (1834, COHA)
(28)a. If I'd known sooner, I would've gone to see you.
b. If I'd known sooner, I'd've gone to see you.

In colloquial usage, separate developments have been taking place in both clauses over a long period; for some discussion see Denison (1998: 140-2, 210-2), (2012: §4.4).

In the *if*-clause, many speakers insert an extra *have*, nearly always unstressed:

(29)If I'd've known sooner, I'd've gone to see you.

This can make for parallel verb strings in each clause. The apparently superfluous morpheme also appears in various other guises in writing, as in (30), with or without apostrophes. The spelling *of* arouses particular prescriptive ire. For some speakers, apparently, the first contracted verb *'d* can be expanded, as in (31), though others (a majority?) find both *would* and *had* ungrammatical here, even – remarkably – if they're happy with the contraction *'d*.

(30)a.	If I'd of known sooner, …	pronounced [əv] with schwa, the same as (29)
b	If I'd'a known sooner, …	pronounced [ə] with schwa and loss of [v]
(31)a.	If I had've known sooner, …	
b	If I would've known sooner, …	

The second morpheme can also be written or spoken in full as *have*, though less commonly than the reduced forms, and there is even corpus evidence that some speakers can pronounce it as stressed *of*, with the vowel of *hot* rather than a schwa. There are other permutations of these variants.

In the second, main clause of (28), what is historically the infinitive of the auxiliary verb *have* is perceived by many speakers not just phonetically but also lexically and syntactically as the preposition *of* – and can appear as such in writing. Whether or not it has been reanalysed as *of*, once again the [v] can disappear under lack of stress

(32)a. … I would of gone to see you.
b. … I woulda gone to see you.

The same possibilities arise after *could, should*, *might* and the infinitive marker *to*.

This maelstrom of possibilities is a good topic for student investigation and discussion. It raises questions of 'correctness' and the role of prescriptive and proscriptive grammar. It is something that students can question in their own experience and usage. Linguistically it is a classic example of grammaticalisation. It is also a nightmare for the lemmatisation and tagging of corpora!

This perhaps is the most elaborate test case given in this chapter, suitable for classes with a good knowledge of English grammar and a willingness to tangle with variation. Other groups will be able to find more modest projects to expand their knowledge of the history of English.

Data sources and abbreviations

ARCHER = A Representative Corpus of Historical English Registers version x. 1990–1993/2002/2007/2010/2013. Originally compiled under the supervision of Douglas Biber and Edward Finegan at Northern Arizona University and University of Southern California; modified and expanded by subsequent members of a consortium of universities. Current member universities are Bamberg, Freiburg, Heidelberg, Helsinki, Lancaster, Leicester, Manchester, Michigan, Northern Arizona, Santiago de Compostela, Southern California, Trier, Uppsala, Zurich. Information available from http://manchester.ac.uk/archer/.

BNC = The British National Corpus, version 3 (BNC XML Edition). 2007. Oxford University Computing Services on behalf of the BNC Consortium. Available online at http://bncweb.lancs.ac.uk/ and at http://corpus.byu.edu/bnc/.

CLMET3.0 = De Smet, Hendrik , Hans-Jürgen Diller & Jukka Tyrkkö. [date?]. The Corpus of Late Modern English Texts, version 3.0, Available at https://perswww.kuleuven.be/~u0044428/

COCA = Davies, Mark. (2008-) *The Corpus of Contemporary American English: 520 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/.

COHA = Davies, Mark. (2010-) *The Corpus of Historical American English: 400 million words, 1810-2009*. Available online at http://corpus.byu.edu/coha/.

ECCO = Eighteenth Century Collections Online. Available online at http://find.galegroup.com/ecco/.

EEBO = Early English Books Online. Available online at http://eebo.chadwyck.com/home and soon from BYU.

GloWbE = Davies, Mark. (2013) *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries*. Available online at http://corpus.byu.edu/glowbe/.

OBC = Huber, Magnus, Magnus Nissel, Patrick Maiwald & Bianca Widlitzki. 2012. The Old Bailey Corpus. Spoken English in the 18th and 19th centuries. Available at www.uni-giessen.de/oldbaileycorpus.

OED = Simpson, J. A. & E. S. C. Weiner (eds.). 2000-. *The Oxford English dictionary online*. Available at www.oed.com/.

PPCEME = Kroch, Anthony, Beatrice Santorini & Lauren Delfs. 2004. Penn-Helsinki Parsed Corpus of Early Modern English. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/index.html.

PPCMBE = Kroch, Anthony, Beatrice Santorini & Ariel Diertani. 2010. Penn Parsed Corpus of Modern British English. http://www.ling.upenn.edu/hist-corpora/PPCMBE-RELEASE-1/index.html .

PPCME2 = Kroch, Anthony & Ann Taylor. 2000. Penn-Helsinki Parsed Corpus of Middle English, 2nd edition. http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-3/index.html

WebCorp = WebCorp: Linguist's search engine. Birmingham City University. Available at http://www.webcorp.org.uk/.

YCOE = Taylor, Ann, Anthony Warner, Susan Pintzuk & Frank Beths. 2003. The York-Toronto-Helsinki Parsed Corpus of Old English Prose. http://www-users.york.ac.uk/~lang22/YCOE/YcoeHome.htm.

Secondary references

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson.

Brems, Lieselotte & Kristin Davidse. 2010. The grammaticalisation of nominal type noun constructions with *kind/sort of*: Chronology and paths of change. *English Studies* 91.2, 180-202.

De Smet, Hendrik. 2012. The course of actualization. *Language* 88.3, 601-33.

Denison, David. 1998. Syntax. In Suzanne Romaine (ed.), *The Cambridge history of the English language*, vol. 4, *1776-1997*, 92-329. Cambridge: Cambridge University Press.

Denison, David. 2002. History of the *sort of* construction family. Paper presented at ICCG2: Second International Conference on Construction Grammar, Helsinki.

Denison, David. 2008. Patterns and productivity. In Susan M. Fitzmaurice & Donka Minkova (eds.), *Studies in the history of the English language IV: Empirical and analytical advances in the study of English language change* (Topics in English Linguistics 61), 207-30. Berlin and New York: Mouton de Gruyter.

Denison, David. 2012. Non-inflecting verbs in Modern English. Paper presented at Autour du verbe / Around the verb: Colloque en l'honneur de Claude Delmas, Paris.

Denison, David. 2013. Parts of speech: Solid citizens or slippery customers? *Journal of the British Academy* 1, 151-85.

Denison, David. submitted. Ambiguity and vagueness in historical change. In Marianne Hundt, Sandra Mollin & Simone Pfenninger (eds.), *The changing English language: Psycholinguistic perspectives* (Studies in English Language). Cambridge: Cambridge University Press.

Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David Lee & Ylva Berglund Prytz. 2008. *Corpus linguistics with BNCweb - a practical guide* (English Corpus Linguistics 6). Frankfurt am Main: Peter Lang.

Hundt, Marianne, Nadja Nesselhauf & Carolin Biewer (eds.). 2007. *Corpus linguistics and the web* (Language and Computers - Studies in Practical Linguistics 59). Amsterdam: Rodopi.

Keizer, Evelien. 2007. *The English Noun Phrase: The nature of linguistic categorization* (Studies in English Language). Cambridge: Cambridge University Press.

Keller, Frank, Maria Lapata & Olga Ourioupina. 2002. Using the Web to overcome data sparseness. In J. Hajic & Y. Matsumoto (eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 230-7. Philadelphia PA:

Kilgarriff, Adam & Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics* 29.3, 333-47.

Kiparsky, Paul. 2014. New perspectives in historical linguistics. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics* (Routledge Handbooks in Linguistics). Routledge.

Leech, Geoffrey & Lu Li. 1995. Indeterminacy between Noun Phrases and Adjective Phrases as complements of the English verb. In Bas Aarts & Charles F. Meyer (eds.), *The verb in contemporary English: Theory and description*, 183-202. Cambridge, etc: Cambridge University Press.

Liberman, Mark. 2016. Nounification of the week. Language Log. http://languagelog.ldc.upenn.edu/nll/?p=24726.

Nevalainen, Terttu. 1999. Lexis and semantics. In Roger Lass (ed.), *The Cambridge history of the English language*, vol. 3, *1476-1776*, 332-458. Cambridge: Cambridge University Press.

Véronis, Jean. 2005. Web: Google's counts faked? *Technologies du Langage*. http://aixtal.blogspot.com/2005/01/web-googles-counts-faked.html <Accessed 26 Jan 2005>.

Zwicky, Arnold. 2005. Just between Dr. Language and I. Language Log. http://itre.cis.upenn.edu/~myl/languagelog/archives/002386.html <Accessed 4 May 2006>.