A diachronic corpus-based study into the effects of age and gender on the usage patterns of verb-forming suffixation in spoken British English

Article

Accepted Version

Laws, J. ORCID: https://orcid.org/0000-0001-7275-116X, Ryder, C. and Jaworska, S. ORCID: https://orcid.org/0000-0001-7465-2245 (2017) A diachronic corpus-based study into the effects of age and gender on the usage patterns of verb-forming suffixation in spoken British English. International Journal of Corpus Linguistics, 22 (3). pp. 375-402. ISSN 1569-9811 doi: https://doi.org/10.1075/ijcl.22.3.04law Available at https://centaur.reading.ac.uk/67314/

To link to this article DOI: http://dx.doi.org/10.1075/ijcl.22.3.04law

Publisher: John Benjamins Publishing Co.

A diachronic corpus-based study into the effects of age and gender on the usage patterns of verb-forming suffixation in spoken British English

Jacqueline Laws, Chris Ryder & Sylvia Jaworska

University of Reading

Abstract

The aim of this paper is to ascertain the degree to which lexical diversity, density and creativity in everyday spoken British English have changed over a 20-year period, as a function of age and gender. Usage patterns of the four verb-forming suffixes, *-ate*, *-en*, *-ify* and *-ize*, were compared in contemporary speech from the BNC2014 with its 20-year old counterpart, the Demographically-Sampled (DS) component of the British National Corpus. Frequency comparisons revealed that verb suffixation is denser in the BNC2014 than in its earlier equivalent (DS), with the exception of the *-en* suffix, the use of which has decreased, particularly among females and younger speakers in general. Males and speakers in the 35-59 age range showed the greatest type diversity; there is evidence that this peak is occurring earlier in the more recent corpus. Contrary to expectations, females rather than males produced the largest number of neologisms and rare forms.

Keywords: diachronic analysis, everyday spoken language, verb-forming suffixation, age, gender

1. Introduction

The linguistic process of interest in this paper is verb formation using derivational morphology. The derivation of complex words, such as *categorize*, allows speakers to package a variety of concepts into a single lexical item: in this case, the noun stem *category* has been converted, with the application of the suffix *-ize*, to a verb meaning *to put into a category*. In addition, derivation provides opportunities for generating neologisms, some of which may become established lexemes, such as *selfie*. There are several hundred derivational morphemes in English (Stein 2007), but the current study focused on the four principal verb-forming suffixes in English: *-ate*, *-en*, *-ify*, and *-ize* (Bauer et al. 2013, Marchand 1969), which constitute a finite set of suffixes that produce complex verbs denoting a change of state, e.g., *activate*, *broaden*, *intensify* and *categorize*.

The distribution of different suffixes varies according to a variety of factors. For example, register variation reflected in speech and writing significantly affect the frequency of derived forms, in particular the four verb-forming suffixes (Biber et al. 1999, Schmid 2011). Thus, the three properties of derivation that form the linguistic focus of this paper are lexical diversity (through the process of derivation), lexical density (through repetition) and creativity (through the creation of neologisms).

Sociolinguistic variationist research has identified that the degree of lexical diversity, density and creativity employed by speakers can be affected by both age and gender. Together with social class and ethnicity, age and gender are the prime social variables considered in research on language variation and change. Gender, in particular, has been extensively investigated producing a vast body of knowledge. Despite the great interest in the impact of age and gender on linguistic choices also over time, most research in the area has focused on phonological variation and lexical choices at the word level. Thus, the aim of this paper is to extend previous variationist research to the linguistic domain of derivational morphology by evaluating the evidence for the effects of age and gender on lexical diversity, density and creativity through the examination of complex verb choices and whether these linguistic patterns have changed over the last 20 years. To our knowledge, there is no published research that has evaluated the impact of these sociolinguistic variables on morphology, derivational morphology in particular, over time. This study, therefore, utilised the Demographically-Sampled (DS) component of the British National Corpus (BNC) corpus and its recent counterpart, the Spoken BNC2014, to achieve this goal.

The release of the new Spoken BNC2014 provides an ideal opportunity to investigate evidence for sociolinguistic variation and change in linguistic features of everyday speech in British English over the 20-year period since its counterpart, the DS, was compiled.

Before discussing the methodology and results, Section 2 discusses the feature in focus, that is, the four verb-forming suffixes and their role in lexical productivity.

Section 3 offers a brief overview of recent research concerned with the effects of the two sociolinguistic variables on speakers' linguistic choices and how the study of derived forms can be utilised to evaluate previous findings. Section 4 summarises the key factors under consideration and predicted outcomes. Section 5 outlines the composition of the datasets being compared together with the methodological procedures undertaken. In Section 6, the results of the study are reported and interpreted in light of the predicted outcomes. The concluding remarks are set out in Section 7.

2. Feature in focus: Complex verb forms

In English, the four principal verb-forming suffixes denote a range of change-of-state events (Bauer et al. 2013, Marchand 1969), for example, ornative (*chlorinate*), causative (*darken*), inchoative (*acidify*) and locative (*containerize*). The reader is referred to Plag (1999, 2004) for a detailed analysis of these four semantic interpretations that relate to change-of-state verbs, together with three further readings (resultative, performative and similitive). A description of these semantic verb categories is available in Laws & Ryder (under review), along with an overview of the etymological, morphological and functional characteristics of these suffixes.

The Latinate suffix *-ate* mainly produces denominal derivatives (*nitrogenate*) and exhibits low productivity (Bauer et al. 2013). This suffix category constitutes the smallest vocabulary set in spoken English and is rarely used in the formation of neologisms (Plag 1999).

The suffix *-en* is of Old English origin and tends to attach to monosyllabic bases (Dixon 2014). This suffix has produced a number of complex verb forms by attaching to nouns (*threaten*) and more predominantly adjectives (*sweeten*), but is now considered to be only marginally productive (Bauer 1983). The vocabulary set of *-en* derivatives in spoken English has been shown to be greater than that formed with *-ate*, but again it is unlikely that neologisms formed with *-en* will be observed in contemporary speech (Biber et al. 1999).

The final two suffixes, *-ify* and *-ize*, are both of Latinate origin and are able to generate verb forms from all seven semantic categories mentioned above. This greater versatility is reflected in their higher productivity: they attach to a larger number of bases than *-ate* and *-en*, producing a richer set of complex verb types. However, as Biber et al. (1999) note, *-ify* derivatives are not commonly used in spoken language; therefore it is expected that its type count will be low, but it is nevertheless a candidate for generating neologisms, such as *Shelleyfy* (Plag 2004). The *-ize* suffix category is roughly seven times greater than the *-ify* category in both speech and writing (Biber et al. 1999:401) and it is deemed to be the most productive of all four verb-forming suffixes (Bauer et al. 2013, Plag 1999). It is therefore expected that this suffix will produce the largest category, and is very likely to generate neologisms in both speech corpora, particularly in the more recent BNC2014.

In summary, the study of complex verb usage patterns can provide evidence of lexical diversity, density and creativity. The following section outlines the research on age and gender which forms the basis for predicting contrasts in these variables between speakers from different sociolinguistic groups.

3. Age and gender in language variation and change

As it is beyond the scope of this paper to review all studies and methodological approaches to the influence of gender and age on language use, variation and change, the following overview discusses predominantly those studies that have explored the effects of gender and age on linguistic choices, and summarises the predictions they provide relating to usage patterns of complex verb derivatives.

3.1 Age

Despite the fact that speaker age is considered to be the prime social correlate of real time language change, compared to other social variables, age has received considerably less attention in sociolinguistics (Eckert 1997, Chambers 2002, Barbieri 2008). The age group which has been most investigated is adolescence. This is motivated by the fact that younger speakers are generally seen as linguistic innovators advancing intergenerational language change as opposed to older speakers who are regarded as more conservative (e.g. Eckert 1997, Kirkham & Moore 2013). Strong affiliation with peers, symbolic resistance to norms in the process of coming-of-age as well as media, music and technology are some of the reasons behind changes leading to the emergence of new linguistic youth styles. These new styles are characterised by linguistic creativity, playfulness, polyphony and bricolage (Androutsopoulos & Georgakopoulou 2003).

Less attention has been paid to other age groups and linguistic features beyond phonology. Work by Rayson et al. (1997) and Barbieri (2008), present an exception. Rayson et al. (1997) investigated change in vocabulary in two age groups under and over 35 in conversations recorded for the DS subset of the BNC. Features that are used significantly more often by younger British speakers included politeness markers *please* and *sorry*. At the same time and perhaps not surprising, they show a marked tendency for the use of taboo words, particularly swearwords. Older speakers used taboo words too, but ones that could be perceived as less offensive.

A wider range of lexical and grammatical features in language use was investigated by Barbieri (2008) who utilised the LSWE American English conversation corpus. Similar to Rayson et al. (1997), she divided the speakers into two groups: younger (15-25) and older speakers (35-60). As in previous studies, her results show that younger speakers use slang and swearwords significantly more often than older groups. Youth language shows a strong tendency to exhibit features of stance, especially emotional and personal involvement indicated by greater use of attitudinal and affective adjectives, intensifiers, discourse markers and first and second personal pronouns. Older speakers, in contrast, appear to use more conventional markers of stance, of which modal verbs seem particularly salient. The author also noted that younger speakers tend to rely on a relatively small set of multipurpose stance markers that they repeat more frequently, whereas older speakers have a wider range at their disposal.

All in all, there is strong evidence for age-dependent language variation based on speaker age, which is not surprising given the different stages of life and challenges that they involve. The fact that adolescent speakers tend to use more swearwords, slang, and are more creative could be a sign of the resistance to established norms and patterns. In contrast, older speakers have a larger vocabulary, which is due to longer life experience, and might be more likely to preserve the norms that they are accustomed to.

3.2 Gender

As with other complex social variables, gender is difficult to define precisely. Whereas in earlier sociolinguistic research the term ‘sex’ was adopted, from the 1970s researchers began increasingly to use ‘gender’, which is more nuanced and accounts for social and cultural conditions that impose social roles and ‘produce’ gendered personae which might not necessarily correspond to biological sex (Cameron 2007). Influenced by Butler’s (1990) performativity theory, gender is nowadays considered something that people *do* as opposed to what they *have*. Several scholars argue therefore that research on language and gender needs to depart from the simple categorisation of speakers into women and men and investigate instead the dynamic ways in which gender is performed contextually looking also at how gender intersects with other social categories (Queen 2014). We are in agreement with the criticism that binary categorisation is somewhat static and insufficient to provide richer insights into the relationship between gender and language use. At the same time, in line with Cheshire (2002), there are still good reasons to include this categorisation in research. As Cheshire points out, most social life is organised alongside binary distinctions, of which the classification into ‘female’ and ‘male’ is the most fundamental and a significant source of social stereotyping. Studying differences and similarities between male and female speakers could help deconstruct some of the myths that such general social categorisations create in the public imagination. Cheshire (2002) also highlights that categorising speakers into male and female, even if this appears to be rather crude, ensures a better replicability and can help draw stronger comparisons in studies interested in the gender factor. For the purpose of the present comparative and diachronic study, we treat ‘gender’ as an independent variable and in line with Milroy & Milroy (1997) see it as an exploratory and purposefully broad and aggregated categorisation, which points to general patterns and does not explain individual linguistic behaviours (cf. Baker 2010, Brezina & Meyerhoff 2014).

Interest in gender and language use goes back to the pioneering work by Fischer (1958) who was the first to look systematically at links between gender and allomorphic phonological variation in verbs. This gave impetus to the development of variationist research in sociolinguistics interested in the differences in language use by men and women. By far, most of this research focused on phonetic variations (e.g. Labov 1972, Mansfield & Trudgill 1994, Nordberg & Sundgren 1998) providing ample evidence that women are more likely to use standard variants than men and favour innovative phonological variants in the situation of a language change from below.

Rayson et al. (1997) studied the impact of gender, amongst other variables, on everyday spoken language by analysing the DS component of the BNC. The authors found that in contrast to men, women tend to use more pronouns, as well as more proper nouns (first names) and verbs. In contrast, men show a greater preference for nouns and noun phrases and tend to use more swearwords. The analysis confirms marked differences between male and female speech in that women tend to use more involved language (Biber 1995) or *rapport speech* (Tannen 1991) orientated towards building and maintaining relationships, whereas men seem to be more concerned with facts and information or use, in Tannen's words, *report speech*. Interestingly, the impact of gender was shown to be much greater than that of any other variables investigated including age and social status.

While most research on gender have been preoccupied with speech, Argamon et al. (2003) studied a large subset of fictional and non-fictional texts from the written BNC and identified marked gender differences in writing. Using the stylistic variation framework devised by Biber (1995), the authors found that texts by women exhibit significantly more features of involved style including personal pronouns, present tense verbs and contractions, whereas men tend to be more factual as evidenced in greater use of common nouns and noun modifications with 'of'.

Taken together, more than sixty years of variationist research concerned with gender and language use points to a consistent and stable pattern of differences across contexts in both speech and writing. Generally speaking, women tend to use more standard variants, appear more conservative and formal in expression, adhere to the rules of politeness and seem less assertive. They orient more strongly towards building and maintaining relationships, cooperation and personal involvement reflected in the increased use of personal pronouns, verbs, politeness devices, apologies and hedges. Contrastingly, men appear to prefer a factual and informational style and make greater use of nouns and noun phrases: they tend to be more assertive and use more taboo words. Because male speakers seem to be less concerned with the norms of politeness and like to say things for impact, they also tend to use more humour and novel expressions (Holmes 1997).

3.3 Summary relating to sociolinguistic variables

Most of the research concerned with the effects of age and gender on language variation and change has been interested in the influence of these variables on selected linguistic units at the level of phonology, syntax, grammar and lexis, with phonological variation being the most studied area. To our knowledge, there is no published study beyond the level of words that examines the effects of the two variables on morphology, specifically derivational morphology. This area warrants special attention, as it could shed light on age / gender effects on the mechanisms of lexical diversity, density and creativity also over time. We endeavour to contribute to this neglected area by investigating the use of four verb-forming suffixes by women and men across a range of ages and whether this has changed over a 20-year period.

4. The current study

This section summarises the variables employed in this study of sociolinguistic variation and change, and how the properties of complex verb derivatives were used to evaluate those factors. The section concludes with the research questions and predictions upon which the study was based.

Research on sociolinguistic variation has provided evidence that although younger speakers use a smaller vocabulary set than older speakers, these forms are repeated more frequently (Barbieri 2008). There is also evidence that females use verbs and pronouns more frequently than males (Rayson et al. 1997). Finally, younger speakers generally, and males, in particular, have been shown to be more inclined to be creative with lexical forms (Holmes 1997).

Usage patterns of complex words are of particular interest to researchers because they provide a language sample set for exploring lexical diversity (the use of derived forms that represent a complex concept), density (token frequency) and lexical creativity (the creation of a neologism) as a function of speaker characteristics, such as age and gender. The four principal verb-forming suffixes in English, -*ate*, -*en*, -*ify* and -*ize* provide a manageable, well-defined set of verb derivatives for the comparison of lexical choices relating to age, gender and diachronic factors. On the basis of these two sets of observations, the following research questions and predictions were posed, where type frequency and token frequency are measures of suffix category diversity and density, respectively:

1) What is the relationship between speaker age and verb-forming suffix usage over time?

   Younger speakers will show less diversity in complex verb choice and will use the same types more frequently than older speakers.

2) What is the relationship between speaker gender and verb-forming suffix usage over time?

   Female speakers will show a greater density of complex verb forms in general but less diversity in complex verb choice than males.

3) What are the characteristics of new coinages in the BNC2014 dataset for each of the four verb-forming suffixes, and in what way are these linked to age and/or gender?

   (i) Younger speakers and males will exhibit a greater number of creative uses of complex verb forms;

   (ii) Verb types unique to the BNC2014 dataset will reflect diachronic changes in terminology, in particular, in relation to technological developments;

   (iii) The verb suffixes that are still deemed to be productive, -*ify* and -*ize*, will generate the majority of neologisms.

5. Data and Methodology

This section presents details of the DS and BNC2014 datasets, and the procedure for extracting the complex verb forms from each of these corpora.

5.1 The BNC datasets

For the purpose of this study, we utilised the larger corpora of spoken everyday British English, the first being the demographically-sampled (DS) sub-corpus, which is part of the original BNC and its recent counterpart the Spoken BNC2014. The DS corpus makes up about 40% (around 4 million tokens) of all spoken material in the BNC and represents everyday spoken conversation between native speakers of British English.

The BNC2014 spoken corpus (Love et al. 2017 forthcoming) was compiled from transcribed recordings of everyday speech over the period 2012 to 2015. The dataset is similar to the DS component of the BNC both in size (around 5 million words), content (informal conversations), speaker first language (British English) and environmental context (mainly the home). Table 1 provides an overview of the distributions of tokens for the metadata categories relevant to the current study. The token totals were extracted using the BNC*web* Spoken Restrictions search facility for the DS, and metadata provided by the BNC2014 project team for the BNC2014 dataset.

As can be seen, for the DS, the ages of 28% of the 1,405 speakers are unknown (unk'n) and gender information relating to 24% of those speakers is also unknown. This contrasts with missing age and gender information on only 1% of the 380 speakers in the BNC2014.

With respect to token counts in the DS, 13.62% of the age sub-category tokens and 12.18% of the gender tokens are unknown, thus reducing the usable corpus size from 4,233,938 to 3,657,427 for age and 3,718,438 for gender comparisons. By contrast, given the far smaller proportion of unknown utterances in the BNC2014, the usable corpus size is minimally reduced from 4,784,691 to 4,755,970 and 4,784,594 tokens for age and gender, respectively. Therefore, since the aim of this study was to compare speaker subgroups, the necessity to remove utterances provided by speakers of unknown age and gender has resulted in the smaller DS corpus being reduced even further in size, compared with the BNC2014.

Table 1. Speaker and token distribution of Age and Gender subcategories in the DS and BNC2014

Cate-	**No.**	**Sp**	**DS**	**DS%**	**Cate-**	**No.**	**Sp**	**BNC**	**BNC**
gory	**Spks**	**%**	**tokens**	**tokens**	**gory**	**Spks**	**%**	**2014 toks**	**2014%**
0-14	201	14	355,673	8.40	**0-10**	3	1	1,281	0.03
15-24	211	15	500,619	11.82	**11-18**	17	4	191,987	4.01
25-34	163	12	690,720	16.31	**19-29**	146	38	1,961,779	41.00
35-44	147	10	705,882	16.67	**30-39**	54	14	834,379	17.44
45-59	153	11	733,141	17.32	**40-49**	44	12	463,022	9.68
60+	142	10	671,392	15.86	**50-59**	41	11	375,368	7.85
unk'n	388	28	576,511	13.62	**60-69**	48	13	625,013	13.06
					70-79	13	3	254,263	5.31
					80-89	7	2	45,066	0.94
					90-99	2	1	3,812	0.08
					unk'n	5	1	28,721	0.60
	1,405	**100**	**4,233,938**	**100.00**		**380**	**100**	**4,784,691**	**100.00**
Male	509	36	1,454,344	34.35	**Male**	171	45	1,911,836	39.96
Female	559	40	2,264,094	53.47	**Female**	207	54	2,872,758	60.04
unk'n	337	24	515,500	12.18	**unk'n**	2	1	97	<0.01
	1,405	**100**	**4,233,938**	**100.00**		**380**	**100**	**4,784,691**	**100.00**

Table 1 also illustrates that the age sub-category boundaries employed for the DS do not coincide with those used for the BNC2014. In order to make direct comparisons between age ranges across the two corpora, it was therefore necessary to regroup the BNC2014 token counts so that they corresponded to the age ranges employed for the DS dataset; this was achieved by halving tokens within some BNC2014 age bands and allocating half the tokens to the lower band and the other half to the next highest age band. For example, to create the age band 0-14, the token count for 0-10 (1,281) was added to half the 11-18 tokens (191,987/2 = 95,993.5), making a total of 97,274.5 tokens.

To create a sub-corpus corresponding to the age group 15-24, the other half of the 11-18 tokens (95,993.5) were added to half of the 19-29 group tokens (1,961,779/2 = 980,889.5) making a total of 1,076,833 tokens. This process, which was also applied to create the age bands 25-34 and 35-44, has the effect of providing a conservative estimate of the number of tokens in the resultant age bracket with respect to the age groups on either side, due to the averaging of token counts at the boundaries of the created age group. The 45-59 band comprised the sum of half the 40-49 tokens and all the 50-59 band tokens. Since, in the DS corpus, the tokens of all participants over the age of 60 are grouped together in a 60+ category, an equivalent 60+ band was also created for the BNC2014 dataset that consisted of the tokens from the 60-69, 70-79, 80-89 and 90-99 age bands. The resultant sub-group totals and the percentage of the overall corpus from which they are derived are presented in Table 2.

Table 2. Speaker and token distribution of Age and Gender subcategories without unknown categories

Cate-gory	**No. Spks**	**Sp %**	**DS tokens**	**DS%**	**Cate-gory**	**No. Spks**	**Sp %**	**BNC 2014 toks**	**BNC 2014%**
0-14	201	20	355,673	9.72	**0-14**	11.5	3	97,275	2.04
15-24	211	21	500,619	13.69	**15-24**	81.5	22	1,076,883	22.64
25-34	163	16	690,720	18.89	**25-34**	100	26	1,398,079	29.40
35-44	147	14	705,882	19.30	**35-44**	49	13	648,701	13.64
45-59	153	15	733,141	20.05	**45-59**	63	17	606,879	12.76
60+	142	14	671,392	18.36	**60+**	70	19	928,154	19.52
	1,017	**100**	**3,657,427**	**100.00**		**375**	**100**	**4,755,970**	**100.00**
Male	509	48	1,454,344	39.11	**Male**	171	45	1,911,836	39.96
Female	559	52	2,264,094	60.89	**Female**	207	55	2,872,758	60.04
	1,068	**100**	**3,718,438**	**100.00**		**378**	**100**	**4,784,594**	**100.00**

Since the token counts for the 0-14 age bands in both datasets were relatively low (9.72% for DS and 2.04% for the BNC2014), in the analyses reported below, these were combined with the 15-24 age band tokens in each case, thus creating a 0-24 age category containing 856,292 tokens (20.22%) for DS and 1,174,158 tokens (24.54%) for the BNC2014 dataset, which are more similar to the proportion of tokens provided by the other age groups in both corpora (see Table 2).

5.2 Retrieval of verbal suffixes

Complex words bearing the suffixes *-ate*, *-en*, *-ify* and *-ize* were extracted from the BNC2014 dataset using *CQPweb*. The extraction process involved using search strings for each suffix (**suffix*) and all possible inflections appropriate to verbs. For verbs ending in *-ize*, the spelling variant *-ise* was also included in the search set; all retrieved verbs with *-ise* were internally standardised to the *-ize* form. Raw data files were compiled with the headwords for each suffix set together with the corresponding Part of Speech (PoS) and token frequency for each entry.

PoS assignments were derived from the grammar tagger analysis: this was CLAWS-5 in the case of the DS and CLAWS-6 in the case of the BNC2014. The tag set for CLAWS-6 is more extensive than that used in CLAWS-5, however, the subsets of grammar tags relevant to verb analysis are equivalent, even though some of the labels differ slightly. The main difference in these two PoS outputs is that CLAWS-5 reports ambiguous tags when the word class of the headword is not clear cut, for example the headword *nationalizing* could be tagged VVG-NN1 since it may represent the present participle of the verb, or it may represent a singular noun formed from the gerund. According to the *BNCweb* website, 3.00% of the PoS allocations result in ambiguous tagging. To resolve all ambiguities of this type, each ambiguously-tagged headword was checked in context. By contrast, CLAWS-6 does not provide ambiguously tagged PoS assignments; the natural consequence of this is that the dataset extracted from the BNC2014 may contain more PoS errors than the DS.

Simplex words ending in the target suffix strings, such as *plate*, *hen*, or *prize*, were then eliminated from the raw word lists and potential complex word candidates were checked against the *Oxford English Dictionary* (*OED*): a headword was included if its etymological formulation in the *OED* was clearly *base* + *suffix*. Proper nouns bearing the four verb-forming suffixes were excluded from the dataset. Multi-morphemic complex words were included in the dataset, for example, *industrialize*. The usage patterns of complex verb forms relating to the DS and BNC2014 datasets were then compared.

The term neologism was initially defined as an item not occurring in the *OED*. As the analysis progressed, the authors found that certain items, such as *zombify*, were listed in the *OED*, but were 'extremely rare' and could even be considered to English speakers as 'invented', in fact this example was rated by the *OED* as having a frequency of less than 0.01/million words, i.e., band 2. Therefore, lexemes classified as frequency band 2 or below were identified for inclusion in the qualitative analyses conducted here.

5.3 Procedures of Data Analysis: consideration of speaker effects and Type 1 errors

As mentioned in Section 4, category diversity and category density were represented by normalised type and token counts, respectively. Normalised token counts are used routinely in corpus linguistics to compare the relative frequency of target items between corpora (McEnery & Hardie 2012: 49-50); normalised values are obtained by dividing target token counts by the total number of tokens in the respective corpora and multiplying the result by a suitable base, such as 1 million. However, with respect to the comparison of normalised type counts, the same procedure cannot be applied because the denominator used in the normalisation calculations would be equivalent to the total number of types in each corpus; such a computation would not be feasible. Therefore normalised type counts were calculated by dividing raw type counts by the total number of tokens in the respective corpora (thus following the normalised token count procedure). However, as is well-documented by Baayen (2008: 222-4), there is a non-linear relationship between type and token growth rate: as token frequency increases, type frequency asymptotes. The disadvantage of this phenomenon with large corpora, such as the DS and BNC2014, is that the normalisation procedure results in deflated normalised type count values. It is therefore worthy of note that the normalised type counts reported in this study are conservative as a result.

The log-likelihood (LL) test, which relies on the normalisation procedure based on total corpus size, is designed to deal with large quantities of non-parametric data and is ideal for testing the significance of differences between token counts of specific linguistic features across corpora (Dunning 1993). The LL test was used in the current study not only for comparing token frequencies but also type frequencies; however, as stated above, the normalisation procedure based on total token count has the effect of deflating normalised type frequencies, thus reducing the likelihood of type count comparisons reaching statistical significance in the current study.

In corpus-based sociolinguistic studies, the analysis of aggregate data often involves the use of the LL test. However, as illustrated by Brezina & Meyerhoff (2014), when analysing aggregate spoken data it is important to ensure that significant LL values derived from the comparison of token counts of a specific lexical item have not occurred as the result of a disproportionate contribution by a single speaker, or sub-group of speakers. Aggregating token counts across speakers may lead to a false positive, or Type I error, if the effect is attributable to a small subset of the speaker sample. Their solution to this potential problem is to compare speaker-based token frequencies using the Mann-Whitney test, rather than the LL statistic. Although the current authors recognise that this approach is very effective in minimising the chances of Type I errors, it was deemed that a speaker-based approach would not be feasible, because it is not possible to extract the complete dataset for individual speakers from the BNC*web* and *CQPweb* interfaces. To adopt this approach, it would be necessary to extract the speaker characteristics for each of the 3,354 complex verb tokens analysed in this study, a task that was deemed impractical.

Instead, we adopted an alternative approach which allows the study of aggregated data without losing track of the effect of individual speakers: whenever a between-group comparison yielded a significant LL value for a particular complex verb type, the distribution of speaker tokens was extracted and examined. If the token count for any speaker exceeded the group mean by two standard deviations, those token counts were replaced by the remaining group mean without the outliers (Jegerski & VanPatten 2014). In the results reported here, all raw token counts were adjusted for outliers before calculating the normalised value.

Another potential cause of Type I errors occurs when the same dataset is used for multiple comparisons. Since the objective of the research reported here was to compare token differences between two corpora and between the two genders or age groups, the corpus x gender and corpus x age band sub-totals were necessarily used twice. To overcome the possible inflation of significant outcomes from multiple comparisons, the Bonferroni correction was applied: for the 2 x 2 design used, the standard α of 0.05 is divided by 2, yielding a corrected α’ of 0.025. To be even more conservative, the *p* value of 0.01 was used as the minimum significance threshold throughout. The following notation has been used to indicate the level of statistical significance: NS: not significant; * *p*<0.05; ** *p*<0.01; *** *p*<0.001; **** *p*<0.0001.

6. Results and Discussion

This section is divided into three parts. Firstly, the overall type and token frequency patterns of the four verb-forming suffixes across the two time periods represented by the DS and BNC2014 corpora are reported. The second section explores the effect of age on the usage patterns of each suffix category and differences in those patterns between the DS and BNC2014; the subsequent section reports on the equivalent analysis with respect to gender, together with an analysis of the age and gender-related neologistic forms encountered.

6.1 Diachronic analysis of complex verb usage patterns

The normalised (per million) type and token frequencies for the DS and BNC2014 datasets are presented in Table 3 in decreasing order of magnitude; shaded cells indicate the larger value, where a significant pair-wise difference was observed.

Table 3. Normalised type and token frequencies of the verb-forming suffixes from the DS and BNC2014

		Types / million					**Tokens / million**			
Suffix	**Example**	**DS**	**2014**	**LL**	***p***		**DS**	**2014**	**LL**	***p***
-ize	*criticize*	16	25	8.59	**		220	371	72.99	****
-en	*frighten*	12	8	3.65	NS		72	54	10.50	**
-ify	*classify*	7	8	0.01	NS		30	56	28.39	****
-ate	*activate*	2	3	1.31	NS		5	21	44.81	****
	Totals	**37**	**43**	**1.93**	**NS**		**327**	**447**	**76.84**	********

The first observation that warrants comment is that normalised type counts of complex verb forms did not increase significantly across the two time periods. However, the largest verb-forming suffix category, *-ize*, has increased significantly in diversity between the DS and BNC2014. These findings confirm the expectation that for the most productive category, *-ize*, diversity is significantly richer in contemporary everyday spoken language, whereas less productive categories have undergone minimal change (Bauer 1983, Plag 1999).

Unsurprisingly, the increase over time was considerably more robust for tokens where the difference between the DS and BNC2014 token counts for three of the four verb-forming suffixes was highly significant. By contrast, Table 3 illustrates that the verb-forming suffix *-en* showed a significant decrease in frequency between the two time periods. A decrease over time is also detected for the type counts that occurred for this suffix (DS: 12 and BNC2014: 8), although this difference failed to reach significance. These observations suggest that, in general, contemporary speakers of British English have increased the frequency with which they use complex verb forms, but that the preference for *-ize*, *-ify* and *-ate* contrasts with a significant decrease in the selection of *-en* verb derivatives.

The next section addresses the first research question which explores the relationship between speaker age and the choice of verb-forming suffixation categories and the extent to which these have changed over the 20-year timeframe.

6.2 Diachronic analysis of age-related usage patterns

In Tables 4 to 8, the peak type and token frequency value across age groups for each corpus have been underlined. Since complex verb types reoccur across age bands, the mean number of types, rather than the total, has been added at the bottom of each table to provide an indication of the number of types that were observed in each time-based corpus. For comparative purposes, means of token counts have also been added below the token age breakdown for each corpus.

When all suffix categories are considered together, the normalised type frequencies in Table 4 indicate that in the age groups 35-44 and 45-59 diversity in verbal suffixation has increased significantly between the DS and the BNC2014.

Furthermore, younger speakers exhibit a more restricted vocabulary set than older speakers.

Table 4. Normalised type and token frequencies of all suffixed verbs from the DS and BNC2014 by age group

	All Types / million				**All Tokens / million**				**2014**
Age	**DS**	**2014**	**LL**	***p***	**DS**	**2014**	**LL**	***p***	**TTR**
0-24	83	74	0.49	NS	274	374	15.06	***	0.20
25-34	81	86	0.12	NS	321	413	10.74	**	0.21
35-44	92	156	11.45	***	381	519	14.48	***	0.30
45-59	90	154	11.39	***	332	516	26.91	****	0.30
60+	88	97	0.35	NS	334	475	19.09	****	0.21
Means	**87**	**113**	**1.22**	**NS**	**328**	**459**	**14.68**	*****	

A comparison between age bands within each corpus revealed no significant differences in type counts between age groups in the DS, whereas for the BNC2014 significantly more complex verbs types occurred in the 35-44 age band than the 25-34 group (LL=19.22, *df*=1, *p*<0.0001) and in the 45-59 than the 60+ group (LL=9.77, *df*=1, *p*<0.01), but no other significant differences were observed between adjacent age groups. Therefore, in the DS corpus, overall diversity failed to distinguish between age groups. By contrast in the more contemporary BNC2014 corpus, in accordance with Barbieri's (2008) findings, older speakers between the ages of 35 and 59 tend to use a considerably more diverse set of complex verbal forms than other age groups.

Table 4 also illustrates that token frequency in the BNC2014 is significantly denser in each age group than the DS. Comparisons revealed a significant increase between the 0-24 and 35-44 age group in both the DS (LL=13.51, *df*=1, *p*<0.001) and BNC2014 dataset (LL=20.19, *df*=1, *p*<0.001). Therefore, repetitive use of certain complex verbs appears to increase in participants over the age of 24 and peak between the ages of 35 and 59. Type/Token ratios (TTR) in Table 4 reveal that complex verbs are represented less densely in the speech of younger than older participants in the BNC2014 (Barbieri 2008), with the exception of the oldest group; this observation holds true for each individual suffix discussed below.

In the following sub-sections, relative type and token frequency patterns across the age bands are discussed with respect to each suffix category in decreasing order of magnitude (see Table 3).

6.2.1 *Age-related usage patterns of -ize*

Table 5 illustrates that the patterns of type and token counts of complex verbs suffixed with *-ize* reflect very closely the overall patterns observed for all complex verb frequencies presented in Table 4. Again, type counts in the 35-44 and 45-59 age groups in the BNC2014 increased significantly compared with the DS type counts, and all age groups demonstrate a significant increase in token frequency in the more contemporary dataset.

Table 5. Normalised type and token frequencies of *-ize* suffixes from the DS and BNC2014 by age group

	***-ize* Types / million**				***-ize* Tokens / million**				**2014**
Age	**DS**	**2014**	**LL**	***p***	**DS**	**2014**	**LL**	***p***	**TTR**
0-24	42	39	0.13	NS	207	264	6.71	**	0.15
25-34	32	47	2.72	NS	215	293	10.94	***	0.16
35-44	41	86	11.16	***	249	377	18.01	****	0.23
45-59	38	83	11.49	***	210	385	35.10	****	0.22
60+	43	46	3.13	NS	220	324	15.63	****	0.14
Means	**39**	**60**	**2.10**	**NS**	**220**	**329**	**14.10**	***	

Inter-age band comparisons between the BNC2014 age bands revealed a significant increase in the diversity of *-ize* verbs used by the 35-44 group compared with the 25-34 group (LL=10.70, *df*=1, *p*<0.001) and between the 45-59 group and the 60+ group (LL=7.98, *df*=1, *p*<0.01). As was observed for the overall normalised token frequencies (Table 4), the greatest density of *-ize* forms occurred in the 35-59 age group compared with all other age bands. Given that *-ize* verbal derivatives contribute 45% of all complex types and 54% of all tokens in the DS, and 68% and 73% respectively in the BNC2014, it is not surprising that *-ize* usage patterns are reflected so strongly in the overall frequency distribution (see Table 3 for the basis of these percentages).

Speakers between 35 and 59 tend to use a wider repertoire of *-ize* suffixed verbs and all age groups have increased their usage of this verbal form in the BNC2014. Interestingly, of the 32 *-ize* derivatives that were unique to the BNC2014, 42% of them were used by the 25-34 age group, whereas only 39% were used by the combined 35-59 groups who used the forms less repetitively. This concurs with Barbieri (2008): the younger group employed the suffix more extensively, but used these forms more frequently than the older participants, as reflected in the TTR values. As noted in Section 2, the *-ize* suffix is highly productive (Bauer et al. 2013, Plag 1999, Biber et al. 1999).

6.2.2 *Age-related usage patterns of -en*

In contrast to the results for *-ize*, Table 6 shows that no significant differences in diversity were obtained for verbs suffixed with *-en* between the corpora for any age group; instead, there is a general decrease in the use of these complex forms between the two time periods tested. This trend is also reflected in frequency of use: all speakers tended to use *-en* derivative complex forms less in the BNC2014, but the only age group in which significantly fewer were identified was the 35-44 age band (the difference observed in the 25-34 range is below criterion); here, complex verbs bearing this suffix occurred significantly more frequently in the DS.

No significant differences were obtained in type counts between DS age groups. The peak in type count (35) observed in the 45-59 group in the BNC2014, was greater than the type count for the 0-24 (LL=5.07, *df*=1, *p*<0.05) and 25-34 bands (LL=5.31, *df*=1, *p*<0.05), although these comparisons do not meet the minimum 0.01 significance criterion set here.

Table 6: Normalised type and token frequencies of *-en* suffixes from the DS and BNC2014 by age group

	***-en* Types / million**				***-en* Tokens / million**				**2014**
Age	**DS**	**2014**	**LL**	***p***	**DS**	**2014**	**LL**	***p***	**TTR**
0-24	29	17	3.25	NS	50	49	0.02	NS	0.35
25-34	28	17	2.29	NS	74	50	4.51	*	0.34
35-44	30	30	0.00	NS	89	48	8.58	**	0.63
45-59	26	35	0.83	NS	60	53	0.34	NS	0.66
60+	30	24	0.54	NS	89	75	0.93	NS	0.32
Means	**28**	**24**	**0.62**	**NS**	**73**	**55**	**1.88**	**NS**	

These findings imply that in contemporary adult speech, a wider range of *-en* verb forms tend to be used by the 35-59 groups than by younger participants. Regarding token density, a significant difference was observed between the DS 0-24 and 35-44 groups (LL=8.52, *df*=1, *p*<0.01). There is also a tendency for speakers to use the *-en* category verb set more frequently as age increases.

Therefore, the vocabulary set size of *-en* complex verbs has not increased between the times the corpora were compiled; in fact, a decreasing trend in usage was observed across age groups. This is consistent with Bauer's (1983) observation that this suffix is only marginally productive, and therefore unlikely to generate new coinages; the decrease in *-en* lexemes may indicate that other verb forms are replacing it in speech. It appears that speakers use *-en* forms less frequently now than 20 years ago, particularly in the 35-44 age range. Furthermore, in contemporary speech, peak usage occurs in older speakers (60+) than was the case 20 years ago, where these forms were employed more frequently by a larger range of speakers (35-60+).

6.2.3 *Age-related usage patterns of -ify*

Table 7 indicates a slight increase in the diversity of *-ify* derivatives between the two time periods, although type frequencies failed to reach significance in any age group. Token frequencies, however, increased significantly in all age groups except for the 35-44 (difference below criterion) and 45-59 age bands. This result can be explained by the shift in peak token counts between the two corpora: in the DS, the peak token value (50) falls in the 45-59 band, whereas in the BNC2014 dataset, the peak (66) occurs in the 35-44 group. A similar trend is observable in the type counts: the DS peak (22) occurs in the 45-59 band and the BNC2014 peak (31) occurs on the younger 35-44 group.

Table 7. Normalised type and token frequencies of *-ify* suffixes from the DS and BNC2014 by age group

	***-ify* Types / million**				***-ify* Tokens / million**				**2014**
Age	**DS**	**2014**	**LL**	***p***	**DS**	**2014**	**LL**	***p***	**TTR**
0-24	11	14	0.50	NS	15	47	16.42	****	0.30
25-34	16	15	0.01	NS	25	53	9.44	**	0.28
35-44	18	31	2.15	NS	38	66	5.17	*	0.47
45-59	22	25	0.12	NS	50	54	0.10	NS	0.46
60+	13	16	0.02	NS	24	56	10.18	**	0.29
Means	**16**	**20**	**0.17**	**NS**	**30**	**55**	**5.67**	*****	

No type count differences were observed in inter-age comparisons between DS age bands, but a below-criterion increase in frequency was found between 25-34 and 45-59 (LL=6.44, *df*=1, *p*<0.05) and a significant decrease between 45-59 and 60+ (LL=6.81, *df*=1, *p*<0.01). By contrast, no token frequency differences were found across the

BNC2014 age bands. These results suggest that in the DS, the speakers who used *-ify* verb forms most frequently were between 35 and 59, whereas in contemporary speech, usage patterns are distributed evenly across all age groups.

To summarise, it appears that although the diversity of *-ify* forms has not increased significantly between the two time periods tested, the groups that use these derivatives more frequently have shifted from the 45-59 to the 35-44 age band in contemporary speech. There is no strong evidence that younger speakers employed this suffix category more creatively than older speakers, but TTR values indicate that the former groups used a smaller set of *-ify* forms more frequently, as has been observed consistently across all four suffixes.

6.2.4 *Age-related usage patterns of -ate*

Table 8 shows that token frequency of *-ate* suffixed verbs has increased significantly in most age groups between the periods in which the corpora were compiled, but that only a below-criterion increase in diversity is detectable in the 60+ age group. The distribution of type counts across the BNC2014 age groups indicates that these verb forms tend to increase in frequency the older the speaker, but inter-age group differences failed to reach statistical significance. A trend, similar to that noted in section 6.2.3 in relation to *-ify*, is also observable here: peak usage of *-ate* forms (11) occurs in the 45-59 age range for DS, but this peak is shifted to the younger age band of 35-44 for the BNC2014 (28). Again, this finding coincides with no overall significant increase in diversity.

Table 8. Normalised type and token frequencies of *-ate* suffixes from the DS and BNC2014 by age group

	***-ate* Types / million**				***-ate* Tokens / million**				**2014**
Age	**DS**	**2014**	**LL**	***p***	**DS**	**2014**	**LL**	***p***	**TTR**
0-24	1	4	1.80	NS	1	14	11.83	***	0.29
25-34	6	6	0.01	NS	7	18	4.08	*	0.33
35-44	3	10	2.90	NS	4	28	13.05	***	0.35
45-59	4	12	2.49	NS	11	23	3.07	NS	0.52
60+	1	11	5.92	*	1	19	13.49	***	0.57
Means	**3**	**9**	**1.77**	**NS**	**5**	**20**	**7.24**	**	

The number of types in this suffix category is very small and so fluctuations in type counts must be interpreted with caution; furthermore, similar to *-en*, the *-ate* suffix is no longer deemed to be productive (Bauer et al. 2013). On the other hand, the significant differences in token counts between the two time periods indicate a robust increase in the frequency with which these verbal derivatives are used in contemporary speech, in particular with respect to the usage of the items *activate*, *evaluate* and *motivate*. A significant increase in token frequency was observed in the DS between the 0-24 and 45-59 age bands (LL=7.34, *df*=1, *p*<0.01), but in the BNC2014 age band comparisons failed to reach significance.

6.3 Diachronic analysis of gender-related usage patterns

Table 9a presents the time-based comparisons of type and token frequencies by gender across verbal suffix categories.

Table 9a. DS vs. BNC2014 comparison of normalised type and token frequencies of all suffixed verbs

	All Types / million				**All Tokens / million**			
Gender	**DS**	**2014**	**LL**	***p***	**DS**	**2014**	**LL**	***p***
Male	65	83	3.63	NS	321	442	31.54	****
Female	49	57	1.42	NS	331	450	45.42	****

Normalised type frequencies of all suffix categories grouped together failed to show any significant differences between the DS and the BNC2014 corpora for either gender. By contrast usage frequencies increased significantly across the two time periods for both genders.

Table 9b. Gender comparison of normalised type and token frequencies of all suffixed verbs

	All Types / million				**All Tokens / million**			
Corpus	**M**	**F**	**LL**	***p***	**M**	**F**	**LL**	***p***
DS	65	49	3.85	NS	321	331	0.28	NS
2014	83	57	11.26	**	442	450	0.15	NS

In accordance with expectations (Labov 1972, Mansfield & Trudgill 1994, Norberg & Sundgren 1999), inter-gender comparisons revealed that males produced a more diverse complex verbal set in both time periods (Table 9b), although this difference only reached significance in the BNC2014 dataset. The suffix types that contribute to this effect are explored in Sections 6.3.1 to 6.3.4. Gender differences were not, by contrast, reflected overall in the token counts for either time period, providing counter-evidence to the findings of Rayson et al. (1997) that females use verbs more frequently. These results indicate that male speakers tend, in general, to use a wider repertoire of complex verbs forms, as reflected by type count differences, but that overall frequency patterns, as measured by token counts, are not associated with gender.

6.3.1 *Gender-related usage patterns of -ize*

The normalised type values in Table 10a indicate that both males and females use a larger vocabulary set of *-ize* derivatives in contemporary speech and that these forms are used more frequently compared with the earlier DS dataset, although the significance level is below criterion.

Table 10a. DS vs. BNC2014 comparison of normalised type and token frequencies of *-ize* suffixes

	***-ize* Types / million**				***-ize* Tokens / million**			
Gender	**DS**	**2014**	**LL**	***p***	**DS**	**2014**	**LL**	***p***
Male	33	49	4.94	*	211	307	29.09	****
Female	19	30	6.20	*	226	323	44.06	****

Significant gender differences were obtained in both corpora with respect to type frequencies (Table 10b): Males produced significantly more *-ize* types than females in the DS and the BNC2014. However, no significant gender differences were obtained for either corpus with respect to usage frequency.

Table 10b. Gender comparison of normalised type and token frequencies of *-ize* suffixes

	***-ize* Types / million**				***-ize* Tokens / million**			
Corpus	**M**	**F**	**LL**	***p***	**M**	**F**	**LL**	***p***
DS	33	19	6.91	**	211	226	0.83	NS
2014	49	30	10.49	***	307	323	0.99	NS

These findings lend support to the outcomes of previous studies that males tend to be more verbally creative than females (Holmes 1993); here, certainly with respect to the most productive verb-forming suffix -*ize*, evidence for this phenomenon can be observed. The types of complex verbs that males used more than once in the DS corpus and which were not found in the female vocabulary set, included: *Americanize, equalize* and *vaporize*; in the BNC2014 the following were identified: *galvanize*, *idealize*, *jeopardize*, *legitimize*, *mobilize*, *normalize*, *revolutionize*, *sensationalize*, *synchronize*, *theorize* and *vandalize*.

Although no significant overall gender effect was obtained between token frequencies, complex verb types used more frequently by females than males across the corpora were: *apologize*, *criticize*, *economize*, *organize*, *realize*, *socialize* and *traumatize*. Males used the following forms more frequently than females: *authorize* and *privatize* and *recognize*. Unexpectedly, females produced as many neologisms or rare words with the -*ize* suffix as males: *alkalize*, *favouritize*, *innoculize* (neologism) and *technologize*; males produced: *academicize*, *civilianize*, *euthanize* and *quantize*, none of which are neologisms.

6.3.2 *Gender-related usage patterns of -en*

No significant differences were found between the DS and BNC2014 for either gender, or between genders in either time period, with respect to type counts of -*en* suffixed verb forms (Table 11a).

Table 11a. DS vs. BNC2014 comparison of normalised type and token frequencies of -*en* suffixes

	-*en* Types / million				**-*en* Tokens / million**			
Gender	**DS**	**2014**	**LL**	***p***	**DS**	**2014**	**LL**	***p***
Male	17	16	0.05	NS	69	62	0.54	NS
Female	18	11	3.37	NS	74	49	13.16	***

Although the type frequencies between the DS (18) and BNC2014 (11) for females failed to reach significance, the significant decrease in the frequency of items that appeared in the former (74) but not in the latter corpus (49) was attributable in part to the lack of the following derivatives in the BNC2014: *deepen, glisten, sharpen, stiffen and waken*, and in part to the items *fasten, frighten, quieten* and *shorten* that occurred more frequently in the earlier corpus. Unsurprisingly, no neologisms were found in this suffix category.

Table 11b. Gender comparison of normalised type and token frequencies of -*en* suffixes

	***-en* Types / million**				***-en* Tokens / million**			
Corpus	**M**	**F**	**LL**	***p***	**M**	**F**	**LL**	***p***
DS	17	18	0.01	NS	69	74	0.37	NS
2014	16	11	1.88	NS	62	49	3.61	NS

Therefore, in contemporary speech, females have decreased their usage of -*en* verbal derivatives more markedly than males over the last 20 years; a follow-on study could investigate what lexical choices females are making in place of this verbal form in contemporary speech.

6.3.3 *Gender-related usage patterns of -ify*

The verbal suffix -*ify*, is still considered to be productive in English (Bauer et al. 2014), but Table 12a indicates that no significant diachronic increases were observed between the corpora with respect to the relative number of verbs bearing this suffix for either gender. Females, however, used certain verbal -*ify* derivatives in the BNC2014 more frequently than males: *clarify*, *classify*, *identify*, *justify*, *qualify*, *satisfy*, *simplify*, *specify* and *terrify*.

Despite the significant increase in usage of -*ify* complex verbs by females presented above, Table 12b shows that although the token frequency of females (58) exceeds that of males (49) in the BNC2014, this difference is not significant. Therefore, over the intervening 20-year period, the inter-gender gap in usage patterns has been eliminated.

Table 12a. DS vs. BNC2014 comparison of normalised type and token frequencies of -*ify* suffixes

	***-ify* Types / million**				***-ify* Tokens / million**			
Gender	**DS**	**2014**	**LL**	***p***	**DS**	**2014**	**LL**	***p***
Male	12	12	0.01	NS	36	49	2.88	NS
Female	10	11	0.21	NS	27	58	30.23	***

Table 12b. Gender comparison of normalised type and token frequencies of -*ify* suffixes

	-ify **Types / million**				***-ify*** **Tokens / million**			
Corpus	**M**	**F**	**LL**	***p***	**M**	**F**	**LL**	***p***
DS	12	10	0.39	NS	36	27	2.83	NS
2014	12	11	0.03	NS	49	58	1.93	NS

Therefore, compared to males, females increased their usage of -*ify* verbal derivatives more markedly than males over time. Furthermore, females produced the only neologisms for this suffix, *popify* and *wintrify* and also the very low frequency lexeme *zombify*.

6.3.4 *Gender-related usage patterns of -ate*

Table 13a illustrates that the range of -*ate* derivatives used by both genders between the two time periods increases, but these differences are statistically negligible. Nevertheless, 67% of the BNC2014 -*ate* category set includes items which did not occur in the DS: *calibrate, encapsulate, facilitate, fixate, formulate, hydrate, orientate, pixelate, pollinate* and *rejuvenate.*

Table 13a. DS vs. BNC2014 comparison of normalised type and token frequencies of -*ate* suffixes

	-ate **Types / million**				***-ate*** **Tokens / million**			
Gender	**DS**	**2014**	**LL**	***p***	**DS**	**2014**	**LL**	***p***
Male	2	6	2.93	NS	5	24	22.74	***
Female	2	4	1.10	NS	5	19	22.68	***

By contrast, Table 13a also shows that the frequency of usage of -*ate* forms increases significantly for both genders between the times the corpora were compiled; this is partly attributable to the additional types in the latter, identified above, in that the new items constituted 33% of the tokens with this suffix in the BNC2014. In addition, certain -*ate* types common to both corpora occurred significantly more frequently in the BNC2014: *evaluate* was used more frequently by males and the item *activate* was used more frequently by females in contexts referring to technology, e.g., phone apps.

Table 13b. Gender comparison of normalised type and token frequencies of -*ate* suffixes

	-*ate* Types / million				-*ate* Tokens / million			
Corpus	**M**	**F**	**LL**	***p***	**M**	**F**	**LL**	***p***
DS	2	2	0.01	NS	5	5	0.00	NS
2014	6	4	0.91	NS	24	19	1.35	NS

Table 13b illustrates that between-gender comparisons revealed no significant differences in type or token counts in the use of -*ate* derivatives for either time period. As expected, no neologisms were observed with this suffix category.

7. Conclusions

This study has demonstrated that complex verb forms provide a useful linguistic gauge for tracking diachronic variations in vocabulary diversity, frequency of use and creativity as a function of age and gender. The results demonstrated that both the diversity and frequency with which speakers choose to employ complex verb types bearing the three Latinate suffixes, -*ize*, -*ify* and -*ate*, have increased over the 20 year period, and that relative increases are associated with the productivity potential of each suffix (Biber et al 1999, Bauer et al 2013). There appears to be evidence that the repertoire of verbs formed with the native suffix -*en* is declining, particularly amongst females and younger speaker groups, and this is accompanied by a commensurate decrease in the frequency of -*en* verb forms; a follow-on study could explore what lexical choices are replacing these native verb forms.

The expectation that younger age groups would exhibit a more restricted range of verbs was borne out consistently across the four suffix categories in both time periods and, in general, greatest diversity was observed in the 35-44 and 45-59 age bands, particularly with respect to -*ize*, although there was evidence that in the case of -*ify* and -*ate*, the diversity peak has shifted from the latter to the former age group over the 20-year period, indicating that vocabulary sets may expand more in the 30s and 40s and remain more fixed in the 50 age group. TTR values demonstrated that the greatest repetition of items occurs in the youngest and oldest groups.

With respect to gender, the observation from previous studies that females use verbs more frequently than males was not confirmed; however there was evidence that males tend to use a wider repertoire of complex verb forms than females in both time periods, although this effect was only robustly significant with the most productive suffix *-ize*. Both genders tend to use complex verbs more frequently in contemporary speech, but statistically significant effects were limited to *-ize* and *-ate*, the largest and smallest category sets, indicating that this outcome is not linked to the past productivity of these suffixes. Females, however, appear to use *-ify* more frequently and *-en* less frequently between the two time periods tested, suggesting that patterns in female speech may be undergoing change, whereas male speech patterns are remaining relatively more constant; this may be related to the difference in the linguistic environment that females have been exposed to over the last 20 years as a result of a greater participation in the workplace than the home. Interestingly, and contrary to expectations, females demonstrated greater creativity than males in the use of new or rare verb forms.

In conclusion, the analysis of complex verb forms in British English reveals evidence of age and gender differences in lexical diversity, creativity and frequency patterns between the two time periods tested, which may be related to the changes in the status of women in British society over the last 20 years, especially their increased participation in higher education and the labour market (Office for National Statistics 2013), as well as increased access to ‘packaged’ information via the media and the internet, which is experienced by all age groups. The current research has opened a new chapter in sociolinguistic variationist research by extending this endeavour to morphological aspects of language.

Acknowledgements

We are grateful to the BNC2014 project team at the University of Lancaster and Cambridge University Press for granting the authors permission to use the *Early-Access Subset of the Spoken BNC2014*.

References

Androutsopoulos, J. & Georgakopoulou, A. (2003). *Discourse Constructions of Youth Identities*. Amsterdam: John Benjamins.

Argamon, S., Koppel, M., Fine, J. & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text*, 23(3), 321-346.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R* Cambridge: Cambridge University Press.

Baker, P. (2010). *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Barbieri, F. (2008). Patterns of age-based linguistic variation in American English. *Journal of Sociolinguistics*, 12(1), 58-88.

Bauer, L. (1983). *English Word-Formation*. Cambridge: Cambridge University Press.

Bauer, L., Lieber, R., & Plag, I. (2013). *The Oxford reference guide to English morphology*. Oxford: Oxford University Press.

Biber, D. (1995). Dimensions of register variation: A cross-linguistic comparison. Cambridge: Cambridge University Press.

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

BNC*web* (CQP-Edition) Version 4.3, November 2013. Retrieved from https://bncweb.lancs.ac.uk/ (last accessed October 2016).

British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved from http://www.natcorp.ox.ac.uk/ (last accessed October 2016).

Brezina, V. & Meyerhoff, M. (2014). Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, 19(1), 1-28.

Butler, J. (1990). *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge.

Cameron, D. (2007). *The Myth of Mars and Venus*. Oxford: Oxford University Press.

Chambers, J. K. (2002). Patterns of variation including change. In In J.K. Chambers & N. Schilling (Eds.), *The Handbook of Language Variation and Change* (pp. 297-322). Oxford: Blackwell.

Cheshire, J. (2002). Sex and gender in variationist research. In J. K. Chambers, P. Trudgill & N. Schilling-Estes (Eds.), *The Handbook of Language Variation and Change* (pp. 423-443). Oxford: Blackwell.

Dixon, R.M.W. (2014). *Making new words: morphological derivation in English*. Oxford: Oxford University Press.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*, 61-74.

Eckert, P. (1997). Age as a sociolinguistic variable. In F. Coulmas (Ed.), *The Handbook of Sociolinguistics* (pp. 151-167). Oxford: Blackwell.

Fischer, J.L. (1958). Social influences on the choice of a linguistic variant. *Word*, 14: 47-56.

Holmes, J. (1997). Women, language and identity. *Journal of Sociolinguistics*, 1(2), 195-223.

Jegerski, J. & VanPatten, B. (2014). *Research methods in second language psycholinguistics*. New York: Routledge.

Kirkham, S. & Moore, E. (2013). Adolescence. In J.K. Chambers & N. Schilling (Eds.), *The Handbook of Language Variation and Change* (pp. 277-296). Oxford: Blackwell.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Laws, J.V. & Ryder, C. (under review) Register Variation in Spoken Language: The Case of Verb-forming Suffixation.

Love, R., Dembry, C., Hardie, A., Brezina V. & McEnery T. (2017 fc). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. Special Issue of *International Journal of Corpus Linguistics*, xx(x), xx-xx.

Mansfield, P. & Trudgill, P. (1994). A sex-specific linguistic feature in a European dialect. *Multilingua*, 13, 381-6.

Marchand, H. (1969). *The categories and types of present-day English word-formation: A synchronic-diachronic approach* (2nd ed.). Munich: C. H. Beck'sche Verlagsbuchhandlung.

McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice* Cambridge: Cambridge University Press.

Milroy, J. & Milroy, L. (1997). Varieties and variation. In F. Coulmas (Ed.), *The Handbook of Sociolinguistics* (pp. 47-64). Oxford: Blackwell.

Nordberg, B. & Sundgren, E. (1998). *On Observing Language Change. A Swedish Case Study*. FUMS Rapport. Uppsala: Uppsala Universitet.

Office for National Statistics. (2013). Women in the labour market. Retrieved from http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/dcp171776_328352.pdf (last access January 2017).

Oxford English Dictionary (OED online) Oxford: Oxford University Press. Retrieved from http://www.oed.com/ (last accessed October 2016).

Plag, I. (1999). *Morphological productivity: structural constraints in English*. Mouton de Gruyter, Berlin.

Plag, I. (2004). Syntactic category information and the semantics of derivational morphological rules. *Folia Linguistica*, *38*(3-4), 193-225.

Queen, R. (2014). Gender, sex, sexuality, and sexual identities. In J.K. Chambers & N. Schilling (Eds.), *The Handbook of Language Variation and Change* (pp. 368-387). Oxford: Wiley-Blackwell.

Rayson, P., Leech, G., & Hodges, M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, 2(1), 133-152.

Schmid, H-J. (2011). *English morphology and word-formation. An introduction*. Berlin: Erich Schmidt Verlag.

Stein, G. (2007). *A Dictionary of English Affixes: Their Function and Meaning*. Munich: Lincom Europa.

Tannen, D. (1991). *You Just Don't Understand: Women and Men in Conversation*. New York: Ballantine.

Trudgill, P. (1972). Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society*, 1, 179-95.

Authors' addresses

Jacqueline Laws
Department of English Language and Applied Linguistics
University of Reading
Whiteknights
Reading RG6 6AW
UK
j.v.laws@reading.ac.uk

Chris Ryder
Department of English Language and Applied Linguistics
University of Reading
Whiteknights
Reading RG6 6AW
UK
c.s.ryder@reading.ac.uk

Sylvia Jaworska

Department of English Language and Applied Linguistics

University of Reading

Whiteknights

Reading RG6 6AW

UK

s.jaworska@reading.ac.uk