ICAME Journal • 47(1) • 2023 • 63–79
DOI: 10.2478/icame-2023-0004

sciendo

ICAME Journal

TV series as disseminators of emerging vocabulary: Non-codified expressions in the TV Corpus

Research Article

Daniela Landert[1]*, Tanja Säily[2], Mika Hämäläinen[2]

[1]Heidelberg University
[2]University of Helsinki

Received 29 November 2022; accepted 16 January 2023

Abstract: This study presents a method for identifying words that appear in corpus data earlier than their first date of attestation in dictionaries. We demonstrate the application of this method based on a large diachronic corpus, the TV Corpus, and the *Oxford English Dictionary* (OED). Combining automatic extraction of candidate terms from the TV Corpus with comprehensive manual analysis and verification, the method identifies 32 words that were used in TV series before their first attestation in the OED. We present a detailed discussion of these words, analysing their distribution across decades and genres of the TV Corpus, their origins, semantic domains and word-formation processes. We also present extracts with their first uses in the TV Corpus and analyse how the words were presented to the large and anonymous mass audience. Our study shows that the method we present is suitable for identifying early attestations of words in large corpora, even though in the case of the TV Corpus, a great deal of manual analysis and verification is needed. In addition, we argue that TV series and other types of fictional texts are an important resource for studying the coinage and spread of terms, due to their function and the fact that they address a mass audience.

Keywords: *fiction • word formation • corpus linguistics • lexicography • non-codified lexis*

1 Introduction

Fictional texts have always played an important role in the study of emerging vocabulary (see, for instance, Busse and Busse 2012: 813–814 on the role of Shakespeare's works). While the strong dominance of fictional texts in first attestations of dictionaries may be partly due to lexicographers' overreliance on literary works (see Brewer 2012), factors relating to the functions of fiction – including, for instance, aesthetic functions, entertainment, emotional engagement – as well as the popularity of some fictional texts are likely to help promote the coinage and spread of new terms.

Neologisms are notoriously difficult to identify. We may come across words that we suspect to be neologisms and, by checking them against dictionaries and other resources, we can verify if they are, indeed, new words. But what if we do not have potential candidates as a starting point? What if we want to study a broader range of emerging vocabulary in a given domain or period? Previous studies have made use of a number of different approaches, including, for instance, using spell-checkers to identify words that are missing from standard dictionaries (Bednarek 2018: Ch. 9), or looking for significant increases in the frequencies of originally rare words in large diachronic datasets (Grieve et al. 2017; Kehoe et al. 2022). In this study, we use a two-step approach that combines a corpus-driven method with detailed philological analysis of candidate items. We base our study on the TV Corpus, a large collection of about 325 million words of user-generated subtitles of television series, which spans almost 70 years, from 1950 to 2018. We retrieve candidate items by comparing the words in the TV Corpus against the *Oxford English Dictionary* (OED) and its *Historical Thesaurus*. During the manual analysis of the resulting words, we further consult the *Merriam-Webster Dictionary*.

* *Corresponding author: Daniela Landert, E-mail: daniela.landert@as.uni-heidelberg.de*

The items we retrieve and discuss include neologisms as well as other instances of emerging vocabulary. It is possible that some words may have been established in some contexts already without having had wide enough circulation to be covered in the OED. In all cases, the terms we look at had not been codified to the extent to warrant their inclusion in the OED at the time. Nevertheless, these words were used in the context of TV series with a wide international appeal, which may have had an effect on their circulation. Given that all of the terms we study were included in the OED at a later point in time, it is at least conceivable that their use in TV series may have helped their spread and acceptance.

The aim of our study is twofold. On the one hand, we want to assess the suitability of the TV Corpus for identifying emerging vocabulary in English. Given the large size of the corpus and the fact that fiction is known to be a domain through which new words enter the language, we expect the TV Corpus (which mostly consists of fictional TV series) to be a promising resource for studying emerging vocabulary. However, we are aware that the corpus includes unmonitored data, a fact that poses problems for the analysis, especially when the dating of the subtitles is incorrect. We will explore this and other problems related to the corpus data and determine whether the approach is still worthwhile.

Our second aim is to explore the role of fictional language in introducing emerging words. For this purpose, we identify and discuss five overlapping groups of lexical items in our data. These groups were identified in a bottom-up approach and include word-formation processes as well as semantic domains of emerging lexis. Each group provides a different kind of insight into the role of fictional language for lexical innovation, and each group presents its own challenges for determining the status of a candidate item. In addition, we will take a closer look at the first instances of some of the retrieved words in the TV Corpus. In this part, our analysis will focus on how new words are introduced, as well as how their meaning is negotiated between the fictional characters of the TV series and how it is communicated to the audience.

Our paper starts with a brief overview of the relevant theoretical concepts and the role of linguistic innovation in TV series in Section 2. Following this, Section 3 presents information on the TV Corpus and the methods we used to automatically retrieve candidate items and to verify them manually. In Section 4, we present an overview of all retrieved words that antedate their earliest attestation in the OED, including observations on their distribution across metadata categories in the corpus and the OED, such as decades, genres, semantic domains and parts of speech. Section 5 is devoted to the analysis of five groups of words that present recurring patterns in our results and which can provide insight into the relationship between fictional language and lexical innovation. In Section 6, we present an analysis of how new words are first used in TV series, before coming to a brief conclusion in Section 7.

2 Neologisms and the language of fiction

2.1 Neologisms and linguistic creativity

The *Oxford companion to the English language* (McArthur et al. 2018) defines neologism as follows: “A new lexeme or sense of a word and the coining or use of new words and senses.” Neologisms can be created through word-formation (e.g. derivation, compounding, back-formation, blending), borrowing from other languages, root-creation, which refers to words coined “ex nihilo”, or shifting meaning. Neologisms are sometimes distinguished from nonce-words or nonce-formations (e.g. Fischer 1998: 3), which are ephemeral coinages for a specific text or by a specific writer (OED s.v. nonce-word). As noted by Kerremans (2015: 27–29), this is often the stance taken by lexicographers: dictionaries cannot include everything, so they concentrate on “lasting contributions to the language” (2015: 28). In practical terms, defining neologisms in such a way that they can be clearly distinguished from other vocabulary is fraught with difficulties and it has been noted that definitions in research literature are often absent or problematic (for discussions, see Fischer 1998: 3; Kerremans 2015: 27–32; Smyk-Bhattacharjee 2009: 37).

For the purposes of the present paper, we are interested in words that the OED does not consider to have been in use at the time of their occurrence in the corpus, but which later become sufficiently established in the speech community to warrant their inclusion in the dictionary. Therefore, we follow

Fischer (1998: 3) in excluding nonce-words from our definition. Unlike Fischer, however, we are interested in early attestations of new words, at which stage they may have been used before but may still be without wider currency. In many cases it is impossible to determine the exact time at which a new word is created. Whether or not an instance of a word is considered a neologism then depends on whether we have a known earlier record of the word in question. Moreover, words may be well-established in some contexts – regions, thematic areas, groups of speakers – but new to more wide-spread language use. Indeed, several of the items that we retrieve in our study had been in use in American English for some time already but had not been sufficiently circulating for them to be included as quotations in the *OED*. This is why we refer to our terms as "emerging vocabulary". Schmid (2008: 3) identifies three stages in the life of a neologism: creation, consolidation, and establishing (see also Smyk-Bhattacharjee 2009: 40). Our focus is on creation, at which stage the word may need contextual and cotextual anchoring to be understood (Baayen and Neijt 1997: 569–570; Schmid 2008: 11; Kerremans 2015: §2.2.1), and on consolidation, during which the form and meaning of the word become stabilised and the word spreads within the speech community. In our case, the speech community can be said to comprise the Anglosphere and other Anglophone countries broadcasting the series.

Previous research on the history of English neologisms has often relied on lexicographical data alone (e.g. Nevalainen 1999 and references therein; Durkin 2014). Corpus-based work, on the other hand, has typically focused on recent developments in newspapers (Renouf 2007; Kehoe et al. 2022) or Internet texts (Smyk-Bhattacharjee 2009; Kerremans 2015; Grieve et al. 2017, 2018). Some scholars have also gone further back in time by examining materials like early English letters (Palmer 2015; Säily et al. 2018, 2021) or scientific texts (Menzel 2018). The role of TV series has been less studied in this respect (but see Section 2.2 below).

Neologisms are closely connected to the notion of linguistic creativity. Körtvélyessy et al. (2022: 29) consider creativity to be a feature shared by all human beings, and they define word formation creativity as "the formation of *new* words […] that are *appropriate* signs of a class of objects to be named […] as a result of the *deliberate creativity* (cognitive activity) of language users; these signs are *useful* and *effective* because they serve the communication purposes of a speech community". Linguistic creativity is a graded phenomenon in that the means of creating neologisms can be more or less conventional and predictable (Körtvélyessy et al. 2022: 25), but also in that the use and functions of linguistic creativity may vary across social and situational contexts. The contexts related to TV series are discussed in the next section.

2.2 Linguistic innovation in TV series

It has often been argued that TV series are a promising type of data for observing linguistic innovation (for an overview, see Bednarek 2018: 28–31). For instance, Tagliamonte and Roberts (2005: 296), who study intensifiers in the TV series *Friends,* argue that this type of data "provides a kind of preview of mainstream language" and that "language is more innovative in the media than in the general population" (2005: 296). Other studies, also based on data from the TV series *Friends,* have found evidence for innovation in the form of the emerging plural address *you guys* (Heyd 2010) and the use of *in* instead of *for* in negative statements with present perfect aspect, such as in "I haven't heard from her in seven months" (Quaglio 2009: 118). While most previous studies on linguistic innovation in TV series have focused on grammar, lexical innovation is likely to behave similarly to grammatical innovation. One of the few studies focusing on lexical innovation in TV series is Mandala (2007), who shows that the innovative use of the adjective suffix *-y* in *Buffy the Vampire Slayer* is linked to characterisation and the establishment of in-group relationships. Bednarek (2018: Chapter 9) investigates non-codified language in the Sydney Corpus of Television Dialogue (SydTV) and finds that most of the non-codified forms are more frequent in SydTV than in the two reference corpora COCA (Spoken) and GloWbE (US) (2018: 194–197), which again points towards the innovative potential of language in TV series.

The leading role of TV language for linguistic innovation is sometimes explained with reference to the informality of language. For instance, Davies (2021) characterises the TV Corpus as "highly informal language" (2021: 27), which, due also to its large size, is particularly suitable for studying linguistic

innovation (2021: 32). If informality were the only reason for focusing on TV series, then the prominent role of this kind of data for studying linguistic innovation might simply be due to the lack of access to even more informal kinds of data. For instance, data from informal spoken conversation is very time-consuming to collect and compile into corpora and, thus, the size of spoken language corpora cannot compete with resources like the TV Corpus.

Since we cannot carry out a direct comparison, it is hard to assess how TV series compare to spontaneous spoken conversation when it comes to linguistic innovation. However, if we look at the overall functions of fiction in general and TV series in particular, there are some additional aspects that support the view that innovation may be of special relevance in such data. For instance, humour is a prominent function of many TV series, and it can be realised through word play and the creation of ad-hoc expressions (Bednarek 2018: 188–189). Other studies have shown that the use of neologisms can contribute to characterisation, i.e. the creation of fictional characters (Reichelt 2021) and that it can help express relationships between characters (Mandala 2007). Neologisms can also be used in catchphrases, such as Sheldon's use of *bazinga* in *The Big Bang Theory* (Bednarek 2012: 224). And last but not least, some genres, such as science fiction and fantasy, introduce new objects and concepts, for which words need to be created (Munat 2007).

3 Data and method

3.1 TV Corpus

The TV Corpus is a 325-million-word corpus of subtitles from TV series from the 1950s to the 2010s (see Davies 2021). With this size and time-depth, it is arguably the largest diachronic speech-related corpus that is available to date. Subtitles are a special type of transcription of fictional spoken language. They are optimised for space and reading speed, which means that some features – especially orality features, like hesitation phenomena (e.g. *uh, um,* false starts) and discourse markers (e.g. *so, well, I mean*) – are often omitted (Guillot 2019: 37). Since lexical words are less likely to be affected by such omissions, we can assume that they do not have a great impact on our results for emerging lexis.

However, there are a number of other factors related to the nature of the data and the corpus composition that pose problems for our analysis. The TV Corpus includes unmonitored data that originates from a repository of user-generated subtitles, opensubtitles.org. The data on this platform is not curated and, consequently, varies in quality a great deal. It includes subtitle files that were very carefully manually transcribed alongside subtitles that were ripped from DVDs and others that were automatically translated from other languages – including backtranslations of foreign language subtitles of English-language films into English. Such backtranslations of subtitles from older TV series can introduce words into the subtitles that have entered the English language long after the TV series was created. In addition to quality problems of the subtitles, there are also mistakes in metainformation, which originate from assigning the wrong TV series to a subtitle file. For instance, the subtitles of a recent TV series are sometimes wrongly assigned to an older TV series, resulting in apparent antedating of the OED.

These data problems mean that a purely automatic approach is insufficient. Instead, we used automatic data retrieval to compile a list of candidate items, which we then checked manually to verify that the word included in the subtitle was actually used in the TV episode and that the metainformation was correct. In what follows, we provide additional information on both steps of the analysis.

3.2 Comparison with the *Oxford English Dictionary*

To discover emerging vocabulary in the TV Corpus, we compared the years of first attestation of each word in the corpus with those of headwords and senses in the *OED* – an approach that was pioneered in a small historical corpus by Säily et al. (2018, 2021). To do this, we utilised a local version of the corpus, purchased from corpusdata.org, as well as local versions of the *OED* and its *Historical Thesaurus* (*HT*),

Table 1. Metadata included in the spreadsheet of candidate items.

Source	Metadata
TV Corpus	lemma, series title + ID, episode title + ID, episode word count, episode language(s), year, genre, country, Internet Movie Database ID
OED	Word: etymon language, etymology type (e.g. compound), year of first attestation, definition, part of speech, *OED* URL, *OED* reference (e.g. 'ballpoint' in ball, n.1) Sense: year of first attestation, definition, part of speech, *OED* URL
HT	breadcrumb (e.g. society » communication » writing » writing materials » writing instrument » [noun] » pen » ballpoint pen), part of speech, *OED* URL

accessed by agreement with Oxford University Press.[1] We also made use of the automated lemmatisation provided with the TV Corpus, which was not error-free but nevertheless helped with the comparison. We computationally mapped each lemma to headwords and word forms in the *OED*; if a lemma could not be mapped, as in the case of most proper nouns that are not listed in dictionaries, it was not considered for the analysis.[2] The computational mapping involved simple string matching, i.e., if the form of the lemma was identical to an *OED* headword or word form, it was mapped to that headword, and if the year of first attestation in the corpus was earlier than that of one or more of the senses, the lemma was listed as a candidate item. Hyphenated lemmas were also mapped with the parts written together and separately.

As the output, an Excel sheet of candidate items was generated listing the lemmas along with metadata from the corpus, *OED* and *HT* (see Table 1). If there was more than one *OED* sense or *HT* category that was antedated by the corpus instance, each received a row of its own, resulting in a total of 259 rows.

3.3 Manual verification

Our next step was to go through the Excel sheet row by row to verify whether the lemma constituted an actual antedating of the headword and sense in question. To do this, we accessed the online version of the corpus to view the instances in context. If it was obvious that the instance did not represent the headword or sense, we discarded it (e.g. *atas* turned out to be *piñatas* with a character encoding issue: "Never too busy to spend time cutting down your *pi? atas*. Next time, try hanging them a little lower"). If there was a later instance in the corpus that still antedated the *OED*, we checked it as well. For cases that looked promising, we downloaded the subtitle file from the OpenSubtitles website and attempted to locate a video recording of the episode to watch the scene, in order to aurally verify that the word was actually said and to better understand the context. We also did our best to verify that the corpus metadata on the year and episode was correct, and updated them if incorrect when the instance was still an antedating. In some cases, we found that the instance actually belonged to a completely different series that had a similar name, and the year could be off by decades, making the antedating false. These instances were discarded.

The end result was a list of 32 words that qualified as emerging vocabulary for our purposes. Some were missing *HT* metadata; where possible, we added a suitable *HT* category based on similar words in the *OED*. Since our local version of the *OED* was acquired in 2017 and work on the third edition of the dictionary is ongoing, we checked the first attestation dates of the 32 items in *OED Online* in case of any updates; however, all items still antedated the *OED*.[3] When available, we also retrieved the first

1 We chose to use the *OED* because we were interested in items antedating the most comprehensive historical dictionary of English covering multiple varieties. Given that many of the TV series were of American origin, it might also have been of interest to compare the corpus against a dictionary of American English; however, we were unable to gain local access to such a dictionary, which would have been required for the computational comparison of first attestation dates.

2 This means that we do not find neologisms which have never been included in the *OED*. However, the vast majority of lemmas that could not be matched to headwords in the *OED* are nonce formations, proper names and, above all, errors in the corpus data, which is why we decided to employ this restriction.

3 At the time of our analysis in 2022, thirteen of the *OED Online* entries for the items had been fully updated for *OED3*, while the rest had undergone some modifications since *OED2* (published in 1989) but were not regarded as completed by the editors.

attestation dates of the items from the online *Merriam-Webster Dictionary*. This provided additional information concerning the degree of codification of our words, which was especially relevant for terms that may have been better established in the American English context than in British English at the time of their use in TV series. While six of the words were antedated by the *Merriam-Webster Dictionary*, we kept them in the analysis because they still antedated the earliest attestations in the *OED*, which was our chief interest in this paper. We will discuss the role of Americanisms in more detail in Section 5.3.

4 Emerging vocabulary in the TV Corpus: Overview

The emerging words found in the corpus are listed in (1):

(1) *ballpoint, biodome, bleeping, cutesy, direct-dial, ew, fabby, forensics, hang-glide, microlaser, mid-engine, munchy, nanite, ooh-ooh, queso, semi-retired, sheesh, shuriken, sicko, sneakily, spam, steely-eyed, strike-torn, tach, three-hitter, voom, wacko, wowee, yay, yipes, youth-oriented, yucky*

In terms of word class, adjectives (11 items) and nouns (10) form the most frequent groups, followed by interjections (6); there are only a couple of verbs (*hang-glide*, *ooh-ooh*) and one adverb (*sneakily*). The prevalence of interjections is particularly interesting and will be discussed in 5.2.2 below.

Etymologically, only one of the words is classified as a borrowing by the *OED*: *shuriken* 'a Japanese martial arts weapon' (Table 2). Another word with a foreign origin is *queso*, which according to the *OED* was shortened within English from *chile con queso* 'a sauce of melted cheese seasoned with chilli peppers', originally borrowed from Mexican Spanish. The rest are mostly English derivatives (15), compounds (8), and a handful of other etymology types; *ew* and *voom* receive the classification of 'Other sources, imitative'. The derivative category includes some interesting suffixes that will be discussed further in 5.1.1 below, while the compounds will be discussed in 5.1.2 and the foreign words in Section 5.3.

The distribution of the words across the three major *HT* domains is as follows: 'the world' (14 items), 'society' (9) and 'the mind' (8). However, considering the first two levels of the *HT* semantic hierarchy, the most common category is 'the mind » emotion' (4), followed by a variety of categories belonging to each of the three domains, as listed in Table 3. These seem to reflect the topics discussed in the series (e.g. society » leisure), their interactivity and informality (the mind » emotion) and genre (society » law).

The first attestation dates of the words in the corpus range from 1951 to 1989: nine items date from the 1950s, ten from the 1960s, nine from the 1970s, and four from the 1980s (see Table 4). Given that the corpus has more data from the later decades and less from the 1950s, the high number of words from the 1950s and the low number from the 1980s onwards is rather surprising. This could be partly due to variation in the coverage of the *OED*: if the later decades are covered in more detail, finding fresh antedatings to them will be harder; on the other hand, some recent lexis may not yet have been recorded in the *OED* at all, thus escaping our analysis. However, it is also possible that the 1950s were a genuinely creative time in terms of the usage of recent vocabulary in TV series. The words from this period range from informal interjections (*ew*, *sheesh*, *voom*, *yipes*) to words describing people (*sneakily*, *steely-eyed*) and new technology (*ballpoint*, *bleeping*).

As for individual TV series, the greatest number of emerging words (5) is found in *Perry Mason*, an American legal drama from the 1950s–60s, followed by *The Phil Silvers Show* (4), an American military sitcom from the 1950s (see Table 5). Compared to the amount of data we have from each series, *The Phil Silvers Show* ranks surprisingly high and is responsible for nearly half of the 1950s emerging words. Perhaps the comedy genre is particularly conducive to the use of emerging vocabulary with its amusing situations, wordplay and highly informal language use.

The TV Corpus includes genre classifications, which are derived from the classifications in the Internet Movie Database (IMDb). Looking at these genres in more detail, comedy does seem to utilise a very high number of emerging words (16), although it also forms the second largest section of the corpus (see Table 6). This contrasts with drama, which is almost double the size of comedy but yields fewer items (12). Note that series may belong to more than one genre – the metadata often seems to

Table 2. Distribution of emerging lexis across OED etymology types.

N	Etymon language	Etymology type	Words
15	English	Derivative	*bleeping, cutesy, fabby, forensics, hang-glide, microlaser, munchy, nanite, semi-retired, sicko, sneakily, tach, wacko, wowee, yucky*
8	English	Compound	*ballpoint, biodome, direct-dial, mid-engine, steely-eyed, strike-torn, three-hitter, youth-oriented*
3	English	None	*ooh-ooh, yay, yipes*
2	Other sources	Imitative	*ew, voom*
1	English	Conversion	*spam*
1	English	Shortening	*queso*
1	English	Variant	*sheesh*
1	Japanese	Borrowing	*shuriken*

Table 3. Distribution of emerging lexis across HT categories.

N	*HT* category	Words
4	the mind » emotion	*ew, sheesh, yay, yucky*
3	society » leisure	*hang-glide, shuriken, three-hitter*
3	society » occupation and work	*mid-engine, nanite, semi-retired*
3	the mind » mental capacity	*sneakily, wowee, yipes*
3	the world » life	*biodome, steely-eyed, youth-oriented*
3	the world » physical sensation	*bleeping, ooh-ooh, voom*
2	society » communication	*ballpoint, direct-dial*
2	the world » food and drink	*munchy, queso*
2	the world » health and disease	*sicko, wacko*
1	society » law	*forensics*
1	the mind » goodness and badness	*fabby*
1	the world » action or operation	*cutesy*
1	the world » existence and causation	*strike-torn*
1	the world » matter	*microlaser*
1	the world » movement	*tach*

Table 4. Distribution of emerging lexis over time.

N	Decade	Subcorpus size (words)	Subcorpus size (series)	Words
9	1950s	2,033,371	17	*ballpoint, bleeping, ew, ooh-ooh, sheesh, sneakily, steely-eyed, voom, yipes*
10	1960s	8,902,678	55	*cutesy, direct-dial, fabby, forensics, semi-retired, strike-torn, tach, wowee, yay, yucky*
9	1970s	8,781,304	93	*microlaser, munchy, queso, shuriken, sicko, spam, three-hitter, wacko, youth-oriented*
4	1980s	15,011,543	160	*biodome, hang-glide, mid-engine, nanite*

Table 5. Distribution of emerging lexis across TV series (N=1 excluded).

N	Series	Subcorpus size (words)	Words
5	*Perry Mason*	1,219,561	*ballpoint, direct-dial, semi-retired, tach, yay*
4	*The Phil Silvers Show*	152,063	*bleeping, sneakily, steely-eyed, voom*
2	*The Avengers*	646,387	*fabby, forensics*
2	*The Fugitive*	532,457	*strike-torn, yucky*
2	*The Honeymooners*	149,943	*ooh-ooh, sheesh*
2	*Mary Tyler Moore*	85,049	*munchy, wacko*

Table 6. Distribution of emerging lexis across genres.

N	Genre	Subcorpus size (words)	Words
16	Comedy	115,637,132	*bleeping, cutesy, ew, fabby, forensics, munchy, ooh-ooh, queso, sheesh, sicko, sneakily, spam, steely-eyed, voom, wacko, yipes*
12	Drama	201,861,018	*ballpoint, biodome, direct-dial, mid-engine, semi-retired, shuriken, sicko, strike-torn, tach, three-hitter, yay, yucky*
11	Crime	89,702,215	*ballpoint, direct-dial, fabby, forensics, hang-glide, semi-retired, strike-torn, tach, three-hitter, yay, yucky*
8	Family	18,418,700	*bleeping, cutesy, ew, ooh-ooh, sheesh, sneakily, steely-eyed, voom*
7	Adventure	41,066,654	*biodome, hang-glide, microlaser, nanite, shuriken, strike-torn, yucky*
7	Mystery	58,656,691	*ballpoint, direct-dial, nanite, semi-retired, tach, yay, youth-oriented*
6	Action	53,992,106	*fabby, forensics, hang-glide, microlaser, nanite, three-hitter*
2	Romance	37,344,212	*mid-engine, yipes*
2	Sci-Fi	17,506,149	*biodome, microlaser*
2	Western	1,835,536	*shuriken, wowee*
1	Fantasy	24,337,452	*cutesy*
1	Horror	9,976,002	*youth-oriented*
1	Thriller	10,952,025	*youth-oriented*
1	War	4,787,255	*sicko*

list three genres per series – which means that there is significant overlap in the data across genres. One genre in particular stands out with a high number of emerging words compared to the size of the subcorpus: family, which is much smaller than action, adventure, mystery or romance but produces more items (8). All of these words are found in family comedies from the 1950s–60s: *The Phil Silvers Show*, *The Honeymooners*, *I Love Lucy* and *Bewitched*.

5 Emerging vocabulary in the TV Corpus: Recurring patterns

In addition to the distributional patterns discussed in Section 4, we noticed other recurring patterns in our set of 32 emerging words. The words form several overlapping groups, which are based on shared word-formation processes, word meaning and regional origin. In this section, we discuss the most striking patterns and the insight they provide for our study.

5.1 Word-formation processes

Two word-formation processes were especially prominent, suffixation (5.1.1) and compounding (5.1.2). We discuss each of them in turn.

5.1.1 Suffixation

One type of word-formation process that stands out amongst the words comprises informal native suffixes: *-y*/*-sy* (adjectival; *cutesy, fabby, munchy, yucky*), *-o* (nominal/adjectival; *sicko, wacko*) and *-ly* (adverbial; *sneakily*). Apart from *yucky*, all of these formations occur in comedy series, reinforcing the link between comedy and informality. Marked or non-codified uses of the *-y* suffix have also been found to be a significant component of telecinematic language in previous research. Mandala (2007) found that marked *-y* suffixation (e.g. *dangery*) was used to characterise certain group members in *Buffy the Vampire Slayer* (see also Reichelt 2021), whereas Bednarek (2018: 192–193) discovered a number of non-codified *-y* types in the SydTV corpus that were rare or non-existent in large spoken-language or web corpora. While our *-y* words may be less marked than Mandala's, they are non-codified in the sense that they antedate the *OED*, and at least two of them seem to be utilised for the purposes of characterisation.

Firstly, *fabby* 'fabulous' is used repeatedly by a young blonde woman contrasted with the men around her in the 1960s British comedic spy series *The Avengers* (Example (2)). Secondly, the 'nursery form' *-sy* (*OED*, s.v. *-sy*) is employed in *cutesy* by a seasoned photographer who is supposed to be good with children as he attempts to photograph a girl with supernatural powers in the 1960s American sitcom *Bewitched* (Example (3); cf. *bearsy* in the previous sentence). While *cutesy* seems to have been a well-established word in the US at the time (the *Merriam-Webster Dictionary* provides a first attestation date of 1914), the late date given by the *OED* (1968, in an entry first published in 1993 and modified in 2020) suggests that this was not the case in the British context and that TV series like this may have played a role in the dissemination of the word. For more on words classified as Americanisms by the *OED*, including *sicko* and *wacko*, see Section 5.3 below.

(2) Isn't it a terribly clever idea a party on a train like this? Don't you think it's just *fabby*? (TV Corpus, *The Avengers: Dressed to Kill*, 1963)
(3) See that nice bearsy there? All right, now, Tabatha you make *cutesy* for Uncle Diego. Watch the pretty pony. (TV Corpus, *Bewitched: Nobody's Perfect*, 1966)

In terms of the functions of the informal suffixes within the fictional dialogue, they are often used in the context of asking for someone's opinion, as in Example (2) above. Another frequent context is that of expressing a negative stance towards other people, as in Example (4):

(4) Let me tell you something. That Phyllis is still a crazy lady. She's *wacko*. (TV Corpus, *Mary Tyler Moore: Bess, You Is My Daughter Now*, 1970)

5.1.2 Compounding

Out of the 32 words in our dataset, as many as eleven can be classified as compounds: *ballpoint* (which is also a shortening of *ballpoint pen*), *biodome, direct-dial, hang-glide, microlaser, mid-engine, semi-retired, steely-eyed, strike-torn, three-hitter* and *youth-oriented*.[4] Interestingly, only one of the words is used in a comedy series (*steely-eyed*), the most common genres being drama and crime, as in Example (5). This could be because many of the compounds are quite term-like and therefore not as likely to occur in informal comedy series. Some of the more technical terms are used in science fiction series (*biodome*, *microlaser*); see Section 5.2.1 below.

(5) My men have been checking on this box. No dice. You can buy it in any five-and-dime. What about the writing? Done with a cheap *ballpoint*. Perry's going to be disappointed. (TV Corpus, *Perry Mason: The Case of the Fancy Figures*, 1958)

4 The OED classifies *hang-glide*, *microlaser* and *semi-retired* as derivatives rather than compounds because the first part is a combining form rather than an independent word; see Table 2.

By contrast, the non-codified compounds discovered by Bednarek (2018: 193–194) in the SydTV corpus seem to be quite informal and represent "scripted orality, creating realism". This is because she focuses on compounds of three or more parts separated by hyphens, which are often used to modify nouns and are designed to sound like they were created on the fly, as in *spilling-the-fruit-punch type*. Nonce-formations of this kind are not recorded in the *OED*; this explains why they are missing from our data, which focuses on words that become established enough to be included in a dictionary. It should be noted that our method considers only individual lemmas in the TV Corpus as candidates for emerging vocabulary, and hence misses most compounds written separately in the corpus (cf. Section 3.2 above). These would naturally be of interest to account for in future research.

The prevalence of compounds among the emerging lexis, which is here defined as words antedating the *OED*, could reflect the compilation principles of the dictionary: transparent compounds may not be high on the priority list of lexicographers, either in terms of inclusion or in terms of updating. Many of these words do not have entries of their own in the *OED*, instead appearing as part of other entries (e.g. *ballpoint* s.v. *ball*). The first attestation dates of the words in our corpus are quite evenly spread from the 1950s to the 1980s, indicating a continuous need for compound terms in TV series, which could prove to be a good source for lexicographers in this respect.

5.2 Semantic domains

In terms of semantic domains, two opposing trends could be observed. On the one hand, we observed many terms from the domain of technology (5.2.1). On the other hand, emotional exclamations were quite prominent (5.2.2).

5.2.1 Technology

Six of the 32 words come from the domain of technology, namely *biodome, forensics, microlaser, mid-engine, nanite* and *tach*. The strong presence of new technical terms in fiction is not surprising. When starting our analysis, we expected one of the sources of emerging lexis to be science fiction series, which often need to make use of new words to refer to aspects of the fictional world that do not exist in the non-fictional reality. In many cases, these words remain confined to the domain of science fiction, sometimes even to one specific TV series or episode. From our results, *biodome* and *nanite* are examples of this. Both in the TV Corpus and in the *OED* attestations, the two words either appear in science fiction or in the context of discussing futuristic scenarios. The earliest instances found in the TV Corpus antedate the *OED* attestations by one year (*nanite*) and by eight years (*biodome*). One of the two earliest instances of *biodome* in the TV Corpus, which both occur in the same episode, is given in Example (6):

(6) You will board the shuttle alone. It will bring you to my *biodome*. You will carry no arms. (TV Corpus, *Blake's 7: Orbit*, 1981)

While *biodome* and *nanite* remain restricted to science fiction, *microlaser* is an example of a word that has spread to other domains. In the TV Corpus it first appears as a weapon in a 1978 episode of *Battlestar Galactica*, antedating the earliest *OED* attestation by a year (see Example (7)):

(7) Five full sections of flame and nowhere else to fall back to. (Muffitbarks) More hose! Release more hose! (Bleeping) Finite *microlaser*. If that happens when you're near the heart wall... I know. Amplify the laser to operating mode. (TV Corpus, *Battlestar Galactica: Fire in Space*, 1978)

In contrast to this science-fiction use, the *OED* includes more recent attestations of the word from scientific publications and discussions of scientific methods.

Not all words from the domain of technology in the TV Corpus appear in science fiction series. The word *tach*, a shortening of *tachometer*, was found in an episode of the crime series *Perry Mason* and *mid-engine*, used to describe a specific type of car, occurs in the dramedy series *Thirtysomething.* The final word, *forensics,* referring to a forensic science department or laboratory, was found in an episode of the British spy series *The Avengers* (not to be confused with the more recent science fiction series of the same name). The instance is given in Example (8):

(8) I've just come from Miss King's apartment, sir. I called in to give her my findings on the Caspar/Minnow cases. - Mother: Which were? - Negative, sir. Only *forensics* would make a special trip to say that they hadn't found anything. (TV Corpus, *The Avengers: Wish You Were Here*, 1968)

While the adjective *forensic* is attested in the *OED* as early as the mid-seventeenth century and noun uses referring to scientific evidence in the late eighteenth century, the use of *forensic* or *forensics* to refer to (the personnel of) a forensic science department (sense 2 b) is first attested only in 1983. Our instance from the TV Corpus shows that this use of the word must be at least 15 years older.

The strong presence of words from technological domains is thus not restricted to science fiction. Instead, it seems that various genres can present early attestations of words referring to technology. Genres like hospital series and crime series include vocabulary from contexts that are usually only accessible to professionals working in the respective field. By introducing these words to a large audience, they may become known and used more widely, which then justifies their inclusion in dictionaries like the *OED*. This also means that the presence of specialised technical terms in TV series can become an indicator of their subsequent spread and acceptance.

5.2.2 Emotional exclamations and onomatopoeia

Another recurring characteristic of the emerging words we retrieved is that several of them refer to emotional exclamations and/or are instances of onomatopoeia. There are eight words, thus one fourth of our instances, which we can include in this group, namely *bleeping, ew, ooh-ooh, sheesh, voom, wowee, yay,* and *yipes*. These words created some problems for the analysis. For instance, it was sometimes rather difficult to assess whether the word was an accurate representation of what was said in the spoken dialogue. To give an example, in the instance from *I Love Lucy* given in Example (9), the exclamation could have been *Oh!* Instead of *Ew!*:

(9) Is that real mink? No, no, it's mink-dyed imitation skunk. *Ew*! Smells awful, you wouldn't want it. (TV Corpus, *I Love Lucy: The Fur Coat*, 1969)

There are two explanations for the relatively large number of emotional exclamations and onomatopoeia in our data. First, these expressions may not be perceived by lexicographers as the most central items to be included in a dictionary. In contrast to many emerging words that refer to new concepts, emotional exclamations tend to be very transparent in meaning and will not require a great deal of explanation. Second, the codification of emotional exclamations and the representation of sounds is not a straightforward process, which may lead to their exclusion from dictionaries, or their inclusion in different (competing) written representations. While for the other emerging lexis, the question is whether or not the word that was spoken in the TV dialogue was attested before in the *OED*, for this group of words the question could be more about when the written representation was established. This leads to a second problem in the analysis, namely that we do not know when the subtitles were created. If the subtitles of a TV series are created decades after the series' release, then it is possible that we are observing a spelling that was common at the time at which subtitles were added, but which would have had a different written representation at the time of the release of the TV series. Thus, the status of such instances is questionable.

Nevertheless, there are some clearer instances in this group, such as the instance of *yay* given in Example (10):

(10) *Yay*! The pirates are coming! The pirates are coming. (TV Corpus, *Perry Mason: The Case of the Mystified Miner*, 1962)

While we excluded some earlier instances of *yay* in the TV Corpus, due to their occurrence in the context of Spanish, this instance appeared to be a use of *yay* in the sense in which it is attested in the *OED* from 1963 onwards.

Even though the analysis of emotional exclamations and onomatopoetic expressions is not unproblematic, we see a great deal of potential for the study of the codification of such expressions based on subtitle data. In contrast to many other types of written data that are usually consulted by

lexicographers, subtitles are very closely related to spoken language, thus making it possible to study how the written representation and incidence of exclamations that mainly occur in spoken language develop over time.

5.3 Americanisms and foreign languages

The final striking tendency concerns Americanisms and words from foreign languages. Two of the 32 words from our list are explicitly labelled in the *OED* as foreign words, *queso*, and *shuriken.* We discuss these two words in Section 6 below. In addition, eight words are explicitly labelled in the *OED* as Americanisms. These words are *cutesy, munchy, sheesh, sicko, tach, three-hitter, voom* and *wacko*. One of these, *three-hitter*, is a baseball term, meaning that it comes from a domain that is closely connected to US culture (see Example (11)). The other words are American expressions for which alternative terms exist in other English varieties.

(11) The oldest kid, his nickname was Spud. He pitched a *three-hitter* in a school game a couple of months ago. (TV Corpus, *Kojak: Conspiracy of Fear,* 1973)

Our study took the *OED* as a starting point. We treated all words as "emerging" for which we found an instance that antedates the earliest attestation included in the *OED*. Since the TV Corpus includes more data from US series than from any other region, the fact that a fourth of our emerging words are Americanisms is not surprising. Furthermore, for five of the words from this group attestations antedating their use in the TV Corpus can be found in the *Merriam-Webster Dictionary,* which indicates that they were not entirely new in the US context at the time at which we observed them. The remaining three – *munchy, three-hitter* and *voom* – are not included in the *Merriam-Webster Dictionary* at all.

The strong presence of Americanisms in our list of words may partly be due to a bias towards British English in the *Oxford English Dictionary*. It is possible that infrequent Americanisms are either consciously excluded or that they were missed, especially in the early phases of their use. However, it is important to keep in mind that all of these words were still included in the *OED* – we restricted our study to words that have an *OED* entry with the same or a closely-related meaning; see Section 3.2. Thus, the words are considered relevant enough for the English language overall to warrant their inclusion in the dictionary, but their first attestations in the *OED* are later than the instances we observed. This demonstrates two things. First, it shows that resources like the TV Corpus have great potential for lexicographers in identifying early attestations of words, especially of Americanisms and lexemes from other varieties with a strong TV or movie culture; the *OED*, for instance, has been criticised for its patchy coverage of Americanisms in the past (Brewer 2007: 199). Work on the third edition of the dictionary already considers material like "film and radio scripts",[5] but new corpora and methods like ours could significantly facilitate this process. While our study mainly retrieved Americanisms, subtitles of Bollywood and Nollywood films could be explored for Indian English and Nigerian English words. Second, our observation points towards the role of TV in the spread of local variants. The fact that the Americanisms were used in the TV series may have helped their spread geographically, given that the series in which the words were used – e.g. *M*A*S*H*, *Perry Mason*, *Kojak* – had a global audience. We cannot assess this in our study, but further comparisons between the use of Americanisms in TV series and their use in other domains in different geographic regions could shed more light on this question.

6 Introducing emerging vocabulary in TV series

The TV Corpus can be used not only to identify emerging vocabulary, but also to study how it is introduced and how its meaning is made explicit. TV series are a form of unidirectional mass media communication and using terms that have not been established can create problems for comprehensibility. The producers

5 https://public.oed.com/history/oed-editions/preface-to-the-third-edition/.

of TV series do not know their audience in advance and the audience do not have any options of asking for clarification if they do not understand a term. Thus, terms that may not be generally known need to be used in a way that their meaning becomes apparent (cf. the contextual anchoring mentioned in Section 2.1 above).

In several cases, the terms we identified are presented together with (near) synonyms, which explain the less well-established term. For instance, in Example (12), the slang word *sicko* is followed by *mental*, and in Example (13) the word *wacko* is used after *crazy*:

(12) Boy, you really are a *sicko*. Mental. (TV Corpus, *M*A*S*H: Quo Vadis, Captain Chandler,* 1975)
(13) Let me tell you something. That Phyllis is still a crazy lady. She's *wacko*. (TV Corpus, *Mary Tyler Moore: Bess, You Is My Daughter Now,* 1970)

In addition, explicit definitions and explanations can sometimes be found, especially with foreign terms and concepts. For instance, in Example (14) the term *shuriken* is introduced as the name of a type of Japanese throwing knife:

(14) Ever seen that before? It is a *shuriken*. (TV Corpus, *Kung Fu: The Assassin*, 1973)

While studies of contextual anchoring have typically focused on linguistic context (e.g. Kaunisto 2013), in TV series it seems that multimodal information also plays an important role in conveying the meaning of emerging lexis. In the episode of *Kung Fu* cited in Example (14), the Japanese weapon is shown, which helps the audience understand the term. In Example (15), the term *bleeping* is used to describe a sound that was at the centre of the attention in the minute preceding the utterance. Thus, the meaning of the term is immediately clear when it is first used:

(15) Bilko! –
Sir.
What was that *bleeping*?
(stammering): *Bleeping*? –
Yes, that *bleeping*. (TV Corpus, *The Phil Silvers Show: The Big Uranium Strike*, 1956)

The general context, too, is relevant for establishing and explaining potentially unfamiliar terms. In Example (16), the term *queso* is used in the context of ordering Mexican food. The scene involves explicit discussion of other foreign food terms, such as *sangria* and *ceviche*, both of which are attested in the *OED* before 1973.

(16) Julio: For you I think I will order something nice and mild. Some chiles rellenos con *queso*.
Fred: I know I'm gonna have a "queso" after I eat that.
Lamont: Pop, you don't have to eat in here. You can get up and leave. (TV Corpus, *Sanford and Son: Pops 'n' Pals,* 1973)

In this scene, a reluctant character, Fred, is introduced to a Mexican restaurant, which presents opportunities for using foreign terms with explanations for Fred that also help the audience understand the terms. While *queso* is never presented in the scene, *sangria* is served and its ingredients are discussed by the fictional characters.

As a final observation, the use of emerging words can develop within a TV episode. An example of this is *fabby*, which is used in three different passages of an episode of the series *The Avengers* (see Examples (17a–c)):

(17a) Pussy: Isn't it a terribly clever idea a party on a train like this? Don't you think it's just *fabby*?
Steed: Very apt, *fabby*. (TV Corpus, *The Avengers: Dressed to Kill,* 1963)
(17b) Pussy: Absolutely *fabby*, don't you agree?
Policeman: I take you mean fabulous?
Pussy: *Fabby*, fabulous, yes, of course.
Policeman: I hardly think these events will pass into fabled legend. (TV Corpus, *The Avengers: Dressed to Kill,* 1963)

(17c) Steed: I think you're absolutely enchanting.
Pussy: (giggles) You're rather *fabby* yourself.
Steed: And so are you. (TV Corpus, *The Avengers: Dressed to Kill,* 1963)

Except for the second instance in Example (17a), all instances of *fabby* in the episode are used by the same character, a young woman at a fancy-dress party, who is dressed as a pussy cat. The first instance is preceded by a sentence that establishes that *fabby* is used as a positive evaluation (*a terribly clever idea*). This instance is repeated by her scene partner, which is the one time another character uses the term. In the second passage in which *fabby* is used, less than two minutes later, there is explicit negotiation of its meaning (*I take you mean fabulous*). After this exchange, the term seems to be sufficiently established to be used without any further explanations when it is used for the third time towards the end of the episode in Example (17c).

7 Conclusion

Our study addressed two related questions: the suitability of the TV Corpus for studying emerging vocabulary and the role of TV series in introducing and disseminating it. Concerning the first question, our method was able to retrieve 32 words which were used in TV series before their first attestation in the *OED*. This may seem like a small number, given that the TV Corpus is a 325-million-word corpus. However, taking into consideration that neologisms are notoriously difficult to identify in a principled manner, identifying such a number of early attestations can still make a contribution to their study. Moreover, the method would be able to find more items if the retrieval method were adjusted. For instance, our search did not consider compounds that are written as two independent words without a hyphen. Including such compounds would make it possible to identify additional instances of emerging vocabulary. It would also be linguistically interesting to extend the analysis to instances occurring in the same year as the *OED* first attestation, as these too are likely be relatively recent. In addition, we restricted our analysis to words that are included in the *OED*, thus excluding nonce formations as well as neologisms that have not yet been included in the dictionary because they are not yet sufficiently established. Again, removing this restriction would make it possible to retrieve additional items – while, at the same time, also leading to a great deal of unwanted hits. There are different ways in which the retrieval method could be varied and possibly improved. For instance, instead of using the *OED* as a basis for comparison for the automatic retrieval, other dictionaries could be used, such as the *Merriam-Webster Dictionary*.

Despite the automatic retrieval we employed, the identification of emerging lexis in the TV Corpus is not possible without a great deal of manual investigation. The identification of the source of the subtitles and their reliability frequently felt like detective work that involved hunting down descriptions and recordings of TV series. The unreliability of the data in the TV Corpus thus poses serious restrictions on the method's efficiency. Nevertheless, the TV Corpus is an invaluable source for linguistic studies. Despite the amount of manual verification that the corpus data requires, it still makes it possible to study a type of data that, until recently, has not been available for linguistic analysis with a comparable size and time depth. We see great potential in expanding our analysis to related corpora. Most importantly, the Movie Corpus – which includes 200 million words of subtitles from movies released since the 1930s – would make it possible to expand the study of emerging vocabulary even further back in time.

Our method also has certain advantages in comparison to previous methods. Methods that rely on the identification of non-codified expressions by spell-checkers (Bednarek 2019) can only be applied to datasets of limited size and are dependent on the mechanics and lexicon of the spell-checkers, which have not necessarily been designed on linguistic principles. For the TV Corpus with 325 million words, these methods are simply not applicable. Other methods have taken known neologisms as a starting point (e.g. Renouf 2007). While such an approach is able to provide valuable complementary insight on the spread of neologisms, it is not able to identify words that have been used in TV series before they

were fully codified. Furthermore, recent methods that identify neologisms based on frequency change within a dataset (Grieve et al. 2017; Kehoe et al. 2022) are inapplicable to our data as they require even more massive corpora, and they too lack the element of codification provided by the *OED*.

In sum, our study shows that novel corpora like the TV Corpus can help identify early attestations of words that antedate existing first attestations in dictionaries. This insight is relevant for linguists who are interested in studying the spread of new words, as well as for lexicographers who would like to expand the existing citation databases of dictionaries. Corpora like this could be used to study the later stages of emerging vocabulary as well, both quantitatively and qualitatively. Last but not least, applying this insight to subtitle data from present-day TV series may make it possible to identify neologisms in the making. After all, popular TV series address a larger audience than most other text types, and new words that are included in them have a good chance to spread.

This brings us to the second question, the role of TV series in introducing and promoting emerging vocabulary. In addition to their large audience, there are some other aspects of TV series that make them very suitable for spreading emerging lexis. As we mentioned in Section 5.2.1, where we discussed emerging words relating to technology, TV series often present scenes set in highly specialised domains to which ordinary individuals do not have access in real life: high-tech laboratories, detective offices, surgical theatres, and government offices, to name just a few. Words that are specific to such domains may have a higher likelihood to be included in fiction than in generic non-fictional language. At the same time, speakers know many of these terms through TV series and other fictional texts. This makes TV series an interesting resource for identifying lexical items from specialised domains that are about to spread to a wider audience. Given that most TV series address a large, non-specialised audience, specialised vocabulary occurring in such series may gain wider acceptance, especially if terms are used in TV series repeatedly. The same argument can be made for foreign language words that are borrowed, such as the Mexican food terms discussed in Section 6. Once such words are repeatedly used in TV series, they are likely to be at a stage of their development at which they are used in spoken language outside of fiction. Thus, for lexicographers, TV series can be resource for identifying such lexical items at an early stage.

The final point concerning the role of TV series in promoting and studying emerging lexis relates to comedy. Half of the terms we identified occurred in TV series that had Comedy as one of their genre tags – despite the fact that Drama includes 75% more data than Comedy, fewer of our terms came from TV series tagged as Drama. This suggests that humour may play an important role in creating and spreading words. The link between neologisms and humorous word play has been made before (Bednarek 2018: 188–189), but a more comprehensive perspective on comedy and emerging vocabulary is missing so far. Studying how comedy and different genres of fiction in general contribute to the coinage and spread of words is a promising avenue for further research.

Acknowledgements

We would like to thank the OED Advisory Forum for comments on an earlier version of this paper, and the anonymous reviewer and the *ICAME Journal* Editors for helpful feedback. This work was supported in part by the Academy of Finland, grant 323390.

Corpus

TV Corpus, https://www.english-corpora.org/tv/.

References

Baayen, Rolf Harald and Anneke Neijt. 1997. Productivity in context: A case study of a Dutch suffix. *Linguistics* 35 (3): 565–587. doi: 10.1515/ling.1997.35.3.565.

Bednarek, Monika. 2012. Constructing "nerdiness": Characterisation in The Big Bang Theory. *Multilingua* 31 (2/3): 199–229.

Bednarek, Monika. 2018. *Language and television series. A linguistic approach to TV dialogue* (Cambridge Applied Linguistics). Cambridge: Cambridge University Press.

Brewer, Charlotte. 2007. *Treasure-house of the language: The living OED*. New Haven: Yale University Press.

Brewer, Charlotte. 2012. Shakespeare, word-coining and the OED. In Peter Holland (ed.). *Shakespeare survey*. Vol. 65, 345–357. Cambridge: Cambridge University Press.

Busse, Ulrich and Beatrix Busse. 2012. Early Modern English: The language of Shakespeare. In A. Bergs and L. J. Brinton (eds.). *English historical linguistics: An international handbook* (Handbooks of Linguistics and Communication Science 34.1). Vol. 1, 808–826. Berlin/Boston: Walter de Gruyter.

Davies, Mark. 2021. The TV and Movies corpora. *International Journal of Corpus Linguistics* 26 (1): 10–37.

Durkin, Philip. 2014. *Borrowed words*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199574995.001.0001.

Fischer, Roswitha. 1998. *Lexical change in present-day English: A corpus-based study of the motivation, institutionalization, and productivity of creative neologisms*. Tübingen: Gunter Narr.

Grieve, Jack, Andrea Nini and Diansheng Guo. 2017. Analyzing lexical emergence in Modern American English online. *English Language and Linguistics* 21 (1): 99–127.

Grieve, Jack, Andrea Nini and Diansheng Guo. 2018. Mapping lexical innovation on American social media. *Journal of English Linguistics* 46 (4): 293–319. doi:10.1177/0075424218793191.

Guillot, Marie-Noëlle. 2019. Subtitling on the cusp of its futures. In L. Pérez-González (ed.). *Routledge handbook of audiovisual translation*, 31–47. London: Routledge.

Heyd, Theresa. 2010. How you guys doin'? Staged orality and emerging plural address in the television series *Friends*. *American Speech* 85 (1): 33–66.

HT = Historical Thesaurus. Oxford: Oxford University Press. https://www.oed.com/thesaurus.

Kaunisto, Mark. 2013. Scare quotes and glosses: Indicators of lexical innovation with affixed derivatives. In R. W. McConchie, T. Juvonen, M. Kaunisto, M. Nevala and J. Tyrkkö (eds.). *Selected proceedings of the 2012 symposium on New Approaches in English Historical Lexis (HEL-LEX 3)*, 97–106. Somerville, Massachusetts: Cascadilla Proceedings Project. https://www.lingref.com/cpp/hel-lex/2012/abstract2839.html.

Kehoe, Andrew, Matt Gee and Antoinette Renouf. 2022. A data-driven approach to finding significant changes in language use through time series analysis. In S. Flach and M. Hilpert (eds.). *Broadening the spectrum of corpus linguistics: New approaches to variability and change* (Studies in Corpus Linguistics 105), 284–317. Amsterdam: John Benjamins.

Kerremans, Daphné. 2015. *A web of new words: A corpus-based study of the conventionalization process of English neologisms* (English Corpus Linguistics). Frankfurt am Main: Peter Lang.

Körtvélyessy, Lívia, Pavol Štekauer and Pavol Kačmár. 2022. *Creativity in word formation and word interpretation: Creative potential and creative performance*. Cambridge: Cambridge University Press.

Mandala, Susan. 2007. Solidarity and the Scoobies: An analysis of the *-y* suffix in the television series *Buffy the Vampire Slayer*. *Language and Literature* 16 (1): 53–73.

McArthur, Tom, Jacqueline Lam-McArthur and Lise Fontaine (eds.). 2018. Neologism. In *The Oxford companion to the English language*. Oxford: Oxford University Press.

Menzel, Katrin. 2018. Using diachronic corpora of scientific journal articles for complementing English corpus-based dictionaries and lexicographical resources for specialized languages. In J. Čibej, V. Gorjanc, I. Kosem and S. Krek (eds.). *Proceedings of the XVIII EURALEX International Congress: Lexicography in global contexts*, 363–372. Ljubljana: Ljubljana University Press.

Merriam-Webster Dictionary. Encyclopaedia Britannica Company. Online version. https://www.merriam-webster.com/.

Munat, Judith. 2007. Lexical creativity as a marker of style in science fiction and children's literature. In J. Munat (ed.). *Lexical creativity, texts and contexts* (Studies in Functional and Structural Linguistics 58), 163–182. Amsterdam/Philadelphia: John Benjamins.

Nevalainen, Terttu. 1999. Early Modern English lexis and semantics. In R. Lass (ed.). *The Cambridge history of the English language*. *Volume III: 1476–1776*, 332–458. Cambridge: Cambridge University Press.

OED = Oxford English Dictionary. OED Online. Oxford: Oxford University Press. https://www.oed.com.

Palmer, Chris C. 2015. Measuring productivity diachronically: Nominal suffixes in English letters, 1400–1600. *English Language and Linguistics* 19 (1): 107–129. doi:10.1017/S1360674314000264.

Quaglio, Paulo. 2009. *Television dialogue: The sitcom* Friends *vs. natural conversation*. Amsterdam/Philadelphia: John Benjamins.

Reichelt, Susan. 2021. Innovation on screen. Marked affixation as characterization cue in *Buffy the Vampire Slayer*. *International Journal of Corpus Linguistics* 26 (1): 95–126.

Renouf, Antoinette. 2007. Tracing lexical productivity and creativity in the British media: "the chavs and the chav-nots". In J. Munat (ed.). *Lexical creativity, texts and contexts* (Studies in Functional and Structural Linguistics 58), 61–89. Amsterdam/Philadelphia: John Benjamins.

Säily, Tanja, Eetu Mäkelä and Mika Hämäläinen. 2018. Explorations into the social contexts of neologism use in early English correspondence. *Pragmatics & Cognition* 25 (1): 30–49. doi: 10.1075/pc.18001.sai.

Säily, Tanja, Eetu Mäkelä and Mika Hämäläinen. 2021. From *plenipotentiary* to *puddingless*: Users and uses of new words in early English letters. In M. Hämäläinen, N. Partanen and K. Alnajjar (eds.), *Multilingual facilitation*, 153–169. Helsinki: University of Helsinki. doi: 10.31885/9789515150257.15.

Schmid, Hans-Jörg. 2008. New words in the mind: Concept-formation and entrenchment of neologisms. *Anglia – Zeitschrift für Englische Philologie* 126 (1): 1–36. doi: 10.1515/angl.2008.002.

Smyk-Bhattacharjee, Dorota. 2009. *Lexical innovation on the internet – neologisms in blogs*. Ph.D. dissertation, University of Zurich.

Tagliamonte, Sali and Chris Roberts. 2005. So weird; so cool; so innovative: The use of intensifiers in the television series *Friends*. *American Speech* 80 (3): 280–300. doi: 10.1215/00031283-80-3-280.