Getting the measure of derivational morphology in adult speech a corpus analysis using MorphoQuantics

Vol. 6 (2014) 3-17
University of Reading
LANGUAGE STUDIES WORKING PAPERS
ISSN 2040-3461
Editors: C. Ciarlo and D.S. Giannoni

Getting the Measure of Derivational Morphology in Adult Speech: A Corpus Analysis Using *MorphoQuantics*

Jacqueline Laws and Chris Ryder

This paper describes the methodology used to compile a corpus called *MorphoQuantics* that contains a comprehensive set of 17,943 complex word types extracted from the spoken component of the British National Corpus (BNC). The categorisation of these complex words was derived primarily from the classification of Prefixes, Suffixes and Combining Forms proposed by Stein (2007). The *MorphoQuantics* corpus has been made available on a website of the same name; it lists 554 word-initial and 281 word-final morphemes in English, their etymology and meaning, and records the type and token frequencies of all the associated complex words containing these morphemes from the spoken element of the BNC, together with their Part of Speech. The results show that, although the number of word-initial affixes is nearly double that of word-final affixes, the relative number of each observed in the BNC is very similar; however, word-final affixes are more productive in that, on average, the frequency with which they attach to different bases is three times that of word-initial affixes. Finally, this paper considers how linguists, psycholinguists and psychologists may use *MorphoQuantics* to support their empirical work in first and second language acquisition, and clinical and educational research.

1. Introduction

The term *selfie* was recently named by Oxford Dictionaries as the word of 2013. This complex word is an excellent example of one of the mechanisms we use to fill a lexical gap in our language. By adding suffixes like -*ie*/-*y* to the noun base *self*, we create a new word which conveys information that otherwise could only be expressed in a much longer expression: *a photograph of oneself*. Complex words such as *selfie* and *cheerful* are formed from a base and a suffix; others, such as *unkind* and *replay* are formed from a base and a prefix. There are around 850 derivational affixes in English (Stein 2007) and, given that they generally modify the meaning of the base word, it is not surprising to learn that in English there are more complex words than there are simple words, such as *self*, *cheer*, *kind* and *play*.

Over recent years, the interest in complex words in English and other languages has focused on two inter-related areas of study. From a purely linguistic perspective, the development of various productivity measures of derivational affixes, based on affix type and token frequency, has provided valuable insights into the factors that determine how easily derivational morphemes produce new words (Baayen 1993; Hay & Baayen 2002; Baayen 2009 inter alia). From a psycholinguistic perspective, the investigation of a variety of frequency effects on lexical decision reaction times (Taft 2004; De Jong et al. 2000; Silva & Clahsen 2008 inter alia) has advanced our understanding of language processing. Being able to isolate the type and token frequencies of the components of complex words is therefore key to developing these fields of research. The aim of developing the *MorphoQuantics* electronic resource was to make such data available to users from both research and applied disciplines. Thus the central purpose of the work presented in this paper has been to provide an empirical benchmark against which theories in many fields can be developed and tested.

In this introduction, we define the types of derivational morphemes that occur in English and provide the rationale for developing *MorphoQuantics*, a corpus of complex words from adult spoken language. The main body of the paper sets out the methodology used to compile *MorphoQuantics* and provides summary statistics of its contents, along with example extracts to show the type of information that is provided for each affix. We conclude by summarising some of the contributions that the data contained in *MorphoQuantics* can provide to various research endeavours discussed throughout the paper, and future planned developments.

1.1. Prefixes, suffixes and combining forms

Prefixes are word-initial and include examples such as *un-* in *un*-kind and *re-* in *re*-play; examples of word-initial combining forms are *demo*-crat and *stetho*-scope. Suffixes are word-final and include morphemes such as *-ful* in help-*ful* and *-ure* in clos-*ure*; examples of word-final combining forms are mon-*arch* and tri-*pod*. These affixes may change the meaning of a word, e.g., from the positive connotation of the adjective *kind* to the negative *un*-kind, and may also change word class, e.g., the verb *close* can be converted to the noun form clos-*ure*.

From an etymological perspective, derivational morphemes fall into two major classes: those derived from Germanic roots (neutral), e.g., *-ful*, *-less*, *-ness* and *-ly*, which can be added to free bases without any change in stress; and Latinate forms (non-neutral), e.g., *-ity*, *-ic*, *-ate* and *-ous*, which are mostly added to bound bases and often create a word stress shift, e.g., as seen in the transition from *átom* to *atómic*. Combining forms, also known as neo-classical elements, are derived from Greek and Latin lexemes.

It is important, at this point, to note the distinction between a root and a stem. Following Payne (2006: 18), a root expresses the most basic meaning of a word and cannot be further broken down into smaller units, e.g., *help* in *helpful*, and *-duce* in *reduce*, i.e., roots can be free or bound, as in these examples. In the case of the bound root *-duce*, it forms the most basic meaning of *reduce* and other words (de-*duce*, in-*duce*) from the Latin verb *ducere* ('to lead'), but does not constitute a word in itself. Stems, on the other hand, may be constructed of smaller units but can be understood in isolation. Thus, *reduce* is not a root, since it can be broken down into *re*- and *-duce*, but it is a stem because it can be independently understood in discourse and takes regular inflectional morphology (reduce-*s*, reduc-*ing*) and further derivational morphemes (*ir*-reduc-*ible*). Therefore, some morphemes can be both roots and stems, e.g. *dog*, *help*: they are roots because they cannot be broken down into smaller units (free roots), and they are also stems because they are independent items to which both inflectional and derivational morphemes can be attached.

The distinction between the classification of a word-initial affix, or word-initial combining form (and likewise that of a word-final affix, or word-final combining form) has been the subject of debate in recent years, the details of which are outside the scope of this paper, but the issues are reviewed in Bauer et al. (2013). The following sections will set out the properties of prefixes, suffixes and combining forms, and how issues relating to their classification were resolved for the purposes of this study.

1.1.1. Prefixes

Prefixes are attached to the beginning of a word. They are normally bound elements and their function is to add additional information that mainly refers to the dimensions of space (*by*-stander), time (*post*-war), degree (*infra*-red), quantification (*octo*-genarian), repetition (*re*-play) and negation (*un*-kind). Prefixes do not change the word class of the base word (Stein 2007). There are allomorphic forms of prefixes which arise depending on the first letter of the base to which the prefix is attached, e.g., the prefixes *il-* in *illegal*, *im-* in *imbalance* and *ir-* in *irrational* are all allomorphs of the prefix *in-*, as in *inequality*.

1.1.2. Suffixes

Derivational suffixes are attached to the end of a word. They are normally bound elements and their function is to add additional information and/or change the word class of the base word (Stein 2007). A few of their multifarious meanings include: nominalisation by conferring agentivity (paint-*er*, advis-*or*), creating abstract nouns from adjectives (kind-*ness*, simplic-*ity*), forming diminutives from proper nouns and nouns (Ann-*ie*, dogg-*y*), generating a causative state (memor-*ise*, liber-*ate,* sharp-*en*), denoting resemblance (pictur-*esque*, wool-*en*), a specific quality (poison-*ous*, dirt-*y*) and many more. As with prefixes, there are allomorphs of some suffixes, e.g., the -*ance* in *performance* is an allomorph of -*ence* in *dependence*.

1.1.3. Combining forms

Since neoclassical combining forms are derived from Greek and Latin lexemes, there is much debate as to whether they should be considered affixes, in the way that prefixes and suffixes are, or bound roots (Lehrer 1998; Prćić 2005, 2008; Stein 2007). Because of their lexical origins, they provide a wide variety of meanings and may be considered to be more like elements of compounds than complex words, since the first element can be said to modify the second (Bauer et al. 2013: 441); in the following examples, the left-hand element modifies the right, whether it is a combining form or a base: *photo*-cell and bacteri-*ology*.

These forms present a problem for classification because the same neoclassical element may be considered an affix in one context, but a combining form in another, and may be either word-initial or word-final, e.g., *arch*-angel, mon-*arch*. The classification of a morpheme as a pre/suffix or combining form is governed by the element to which it is attached and whether or not this in itself is an existing affix in English. While -*logue* would be considered a suffix in *travelogue*, it is a combining form in *monologue*, in which it is paired with another element that is not a free stem; similarly, *mono*- is a prefix in *monorail* but a combining form in *monologue*.

In essence, it is the presence of two roots, neither of which are stems, that classifies the two morphemes in *monologue* as combining forms. Both these morphemes are bound roots, from Greek *μονο*- (‘one’, ‘having one’) and -*λογος* (‘a type of discourse’), but neither exists as an independent stem in English; whereas in *travelogue* and *monorail*, the suffix -*logue* and prefix *mono*- have been attached to the stems *travel* and *rail*.

On the one hand, classifying such morphemes as ‘both’ affix and combining form does not tell us which is being used in a given word; on the other, treating the final elements in *travelogue* and *monologue* as two different morphemes seems counter-intuitive given that their meanings and origins are identical, and that it is unconventional to classify units based not on their own characteristics but on those of the elements with which they appear at the time. In order to overcome these classificational difficulties with combining forms, it was decided to adopt the system used by Stein (2007), which means that some affixes, be they word-initial or word-final, are classified as both pre/suffixes and combining forms.

1.1.4. Sources of affixes in English

The seminal volume that has served as the richest source of information on common forms of English affixation is Marchand (1969). In a corpus study, Hay and Baayen (2002) explored the productivity characteristics of a subset of 80 common affixes in English. Both these sources did not include combining forms and, as Table 1 shows, the number of suffixes from these sources is around double the number of prefixes. However, Table 1 also demonstrates that when combining forms are also included, as in the volume produced by Stein (2007), the number of word-initial affixes is nearly double that of word-final affixes. This is because there is a vast quantity of word-initial morphemes, classified as both combining forms and prefixes (251 in Table 1), that are specific to specialised genres, such as medicine, chemistry and other scientific subjects, and are not shared by more accessible domains.

Table 1 presents the number of distinct affixes reported by Stein (2007); if part of speech variants (e.g., *evolution-ary* [Noun] and *bound-ary* [Adjective]) and variant spellings (e.g., *co-agulate* and *col-laborate*) are added to these, the total number of word-initial and word-final affixes rises to 583 and 396 respectively.

	Word-Initial Affixes	*Word-Final Affixes*	*Totals*
Marchand (1969)	65	104	169
Hay & Baayen (2002)	26	54	80
Stein (2007)	547	296	843
breakdown:	prefixes 171	suffixes 164	
	combining forms 125	combining forms 107	
	both 251	both 25	

Table 1. Number of distinct affixes reported in frequently cited sources.

The aim of Stein's (2007) volume was to provide non-native learners of English with a set of the 'chief bound elements' of complex words in present-day English, in order that they may become familiar with the function and form of those derivational morphemes which are likely to occur in an academic context. For this reason, the affix set provided by Stein was used by the current authors as the basis for producing a master list of the complex words to be investigated in the spoken component of the BNC.

1.2. Why focus on complex words in adult spoken language?

Spoken language constitutes the largest proportion of language produced on a day-to-day basis and an understanding of the usage profile of complex words tells us a great deal about the characteristics of word formation and derivational morpheme productivity in English, from a theoretical perspective, as well as providing the means for supporting literacy programmes in educational and clinical settings.

Spoken language furnishes the primary linguistic data set that infants are exposed to and it is therefore crucial to vocabulary acquisition. The development of word-formation rules (Clark 1993) is influenced by Transparency (of word bases and affixes), Simplicity (the number of morpho-phonological changes that are involved) and Productivity (morphemes and word-formation processes that are more productive are acquired earlier than less productive ones). Furthermore, the acquisition of derivational morphology is challenging to investigate because of its interdependence with lexical development. On the one hand, productive use of derivational processes increases with vocabulary development (Clark 1981; Derwing & Baker 1986; Anglin 1993); and on the other, the awareness of derivational processes has been shown to enhance the learning of new vocabulary (Freyd & Baron 1982), spelling performance (Carlisle 1988) and the efficacy of general language instruction (Moats & Smith 1992). In addition, there are important educational implications relating to children's knowledge of non-neutral morphemes: Latinate forms feature predominantly in the language of the academic register (Schleppegrell 2001, 2004), both in the classroom and in school texts, and early exposure to these low-frequency word types has been shown to facilitate academic success (Corson 1985; Dickinson & Tabors 2001).

Therefore, having a description of the frequency distribution characteristics of derivational usage patterns in adult speech can provide a useful tool for mapping the development of word-formation rules in pre-school children, and can assist in identifying whether their language is developing normally. Furthermore, the same distributional norms can also be used to gauge linguistic awareness of second language learners of English. The following section provides further information on these and other related lines of research.

1.3. Filling the gap: the requirement for *MorphoQuantics*

The development of *MorphoQuantics* was motivated by the requirement to obtain type and token frequencies of a wide range of derivational morphemes that could provide normative data for adult speech against which other speech sources could be compared. Currently, no such source of frequency data exists. Complex words in any online version of the BNC can be identified by searching on specific word-initial or word-final affixes, but the results produced include all words, complex and simple, containing the search string and the researcher must resort to checking which words from the resulting list are of relevance, by checking the etymology of each item. Therefore, *MorphoQuantics* was developed to provide the type and token frequencies of a comprehensive list of English derivational morphemes. The complex words from which these frequencies were derived have been checked against the *Oxford English Dictionary*, and thus constitute a validated set of normative data of adult speech for use by researchers in a variety of fields.

The *MorphoQuantics* corpus is available free of charge via the website of the same name (Laws & Ryder 2014). The data are provided in electronic format so that researchers can download the corpus material and perform further searches to suit their research objectives.

MorphoQuantics has wide applicability to various fields of applied linguistics. From a developmental perspective, as mentioned in the previous section, this resource can greatly enhance our understanding of the acquisition of word-formation rules in young children. A project of this type is currently being conducted by Laws (a. In preparation): the CHILDES database (MacWhinney 2000) has been used to create a corpus of speech data from children aged 2 to 5, and a separate corpus containing the concurrent speech of adults also participating in the interactions. The acquisition of derivational morphology in this age group will be gauged against *MorphoQuantics* data, on the basis that the latter provides norms relating to the ‘general’ linguistic environment which exists in parallel with the immediate Child-Directed Speech obtainable from the corresponding adult speech data extracted from CHILDES files. The Laws (a. In preparation) project identifies the order of derivational morpheme acquisition in pre-school children, and the relative speed with which that acquisition takes place.

The results of this forthcoming paper will produce strategies for helping teachers enhance literacy in young schoolchildren. In addition, it can input directly to methods for exploring the factors that affect vocabulary development in children with Specific Language Impairment; there is evidence that these individuals encounter difficulties producing complex words compared with their language-matched peers (Marshall & van der Lely 2007). *MorphoQuantics* provides a dataset from which experimental materials can be produced to support these explorations.

From the perspective of second language acquisition, the frequency distribution patterns of derivational morphemes provided by *MorphoQuantics* can be used to compile stimulus material to investigate linguistic processing. The lexical decision paradigm has already been employed extensively in this area of research (Taft 2004; Silva & Clahsen 2008 inter alia), using word frequencies based on written sources; given the different distributional profiles of derivational morphemes across written and spoken data (Plag et al. 1999), experimentation employing the frequency norms derived from the latter would seem to provide a fruitful direction to explore.

Finally, the data from *MorphoQuantics* can be used to compare the characteristics of derivational usage patterns of spoken language (from the BNC) with other genres such as texts, emails, blogs and Twitter, where the style of English used is closer to the spoken than written form.

1.4. Selection of the corpus of spoken English

The task of obtaining electronic corpora of spoken language is, of course, an order of magnitude more costly in time and effort than the process of compiling a written corpus, which is why the spoken elements of many electronic corpora of British and American English are very limited, e.g., 10% of the British National Corpus (BNC) is spoken (approximately 10 million tokens) and 5% of the Bank of English (approximately 2.275 million tokens).

The BNC provides the largest grammatically tagged 'reference' corpus of present day British English (Leech et al. 2001). The spoken element of the BNC was recorded between 1991 and 1994; this means that some vocabulary usage will be out-of-date, and more recent coinages will not be available. Nevertheless, the BNC is generally considered the 'standard' that researchers use to investigate word frequency in contemporary English and it is often used as a baseline against which other corpora are compared, e.g., Montero-Fleta (2011). It is important to note that the CHILDES database, which is considered a reference standard for child language research, was also recorded some time ago: the files used by Laws (a. In preparation) were recorded between 1962 and 2004. Clearly, the ability of these large corpora to reflect 'present day English' is to some degree limited, but their contribution as reference standards is invaluable for corpus linguists, particularly in studies that employ a diachronic approach.

For these reasons, the BNC was deemed a most appropriate dataset for the corpus analysis reported here, both because it provides a rich source of spoken data and because it is frequently used as a baseline for diachronic analyses in research.

2. Method

2.1. The Spoken Component of the BNC

The spoken component of the BNC is made up of about 10 million tokens obtained from two main sources called the Demographically Sampled component (DS) and the Context-Governed component (CG). The DS component is derived from spontaneous speech data of 124 British speakers ranging in age from under 15 to over 60, from the social class categories DE to AB, and from various regions of the UK; it constitutes about 40% of the spoken BNC. The CG component contains speech from various domains such as education, business, institutions and leisure environments; it constitutes about 60% of the spoken BNC. The data set provided by *MorphoQuantics* combines both these components.

The BNC is tagged for parts of speech (PoS). The version of the BNC used to compile *MorphoQuantics* was Davies (2012); the grammar tagger employed was CLAWS5. A grammar tagger provides invaluable assistance to the researcher by assigning PoS to each token in the corpus; however, there are times when the context, especially in spoken language which is typically fragmented, is insufficient for the software to disambiguate between two possible PoS assignments. Details about how such ambiguities were resolved are discussed in Section 2.5 below.

2.2. Selection of derivational morphemes

As stated in 1.1.4, the main source of word-initial and word-final derivational morphemes was Stein (2007). To these a further three word-initial and six word-final entries were added from Marchand (1969) and Hay and Baayen (2002), and from consulting the *Oxford English Dictionary* (OED online), for purposes of clarification. These additional entries were: the variant *ab-* (*abdicate*); *oxy-* (*oxymoron*); *spiro-* (*spirochete*); *-ad*, (*nomad*); *-centric*, (*egocentric*); *-fic*, (*prolific*); *-like*, (*childlike*); *-ose*, (*cellulose*); *-ulent*, (*fraudulent*). Some morphemes are expressed in more than one variant form, for example the prefix *in-*, as in *inappropriate*, has the same etymon, (Latin *in-* meaning 'not', 'without', 'lacking'), as the three variant forms *il*-legal, *im*-proper and *ir*-rational, depending on the initial letter of the stem. All allomorphic variants of word-initial and word-final affixes were included. Altogether, *MorphoQuantics* contains the complex words derived from 554 word-initial and 281 word-final classifiable affixes and the type and token frequencies for each of these; these totals are smaller than Stein's (Table 1) because some variant types were combined where the meaning and etymology were found to be the same.

2.3. Extraction of complex words from the BNC

All word forms pertaining to each complex word were identified, e.g., all possible inflections were added to the affix search string to capture the plural -*s*, apostrophe -'*s* for both singular and plural forms, third person singular present simple -*s*, the present participle -*ing*, the past simple and participle -*ed*, and the comparative -*er* and superlative -*est* on adjectives. Frequency values of these inflected forms, or 'duplicate forms', were recorded and a total frequency value was also assigned to each entry.

2.4. Classification of complex words

Each item was checked against the *Oxford English Dictionary* (OED online) to ensure that it was a true complex word and the affix was classified accordingly. There were a few instances in which entries in the OED online were not transparent, in that the etymology of a given word was shown not to be an example of a particular affix, while the entry for the affix itself listed the given word as an example of its usage. An example of this is *entry*: the OED provides no classification for the -*y* suffix (of which there are six sub-categories, -*y*^{1} to -*y*^{6}), but by consulting the characteristics of the -*y*^{5} suffix, the OED provides *entry* as an example of this sub-category. In instances where the affix was not clearly defined by the OED, judgements were made based on the individual circumstances and with reference to Stein (2007); notes highlighting any ambiguities in classification were attached to the relevant headword and/or affix. Additionally, some affixes, in particular highly specialized combining forms, were not given a separate entry in the OED, for example *tacho*-, the etymon for which had to be extracted from the definition of the first element of *tachometer*. These were relatively few, however, and it was normally clear from the etymological information whether or not the morpheme corresponded to the relevant Greek or Latin lexeme.

Each complex word was assigned to an affix 'category', depending on the grammatical and semantic properties of the affix. From the grammatical perspective, a simple affix string can generate several different PoS 'categories', e.g., the suffix -*ly* can be attached to adjectives to form adverbs (slow-*ly*) and to nouns to form adjectives (friend-*ly*), thus the former variant was coded -*ly*^{1} and the latter -*ly*^{2}. Alternatively, an affix string may be homonymous: the suffix -*y* has several different semantic functions, e.g., 'having' or 'full of' (*brainy*), 'to cause to have' (*dirty*), nominalisation of a verb (*recovery*), 'office or domain' (*millinery*), diminutive forms (*doggy*, *Billy*) and many more. Some of these examples share etymological roots, as illustrated by the example of *in*- in Section 2.2, thus rendering the different variants polysemous, but many do not.

Appendices A and B provide the full list of affixes used in this study. All of these are included in *MorphoQuantic*s, regardless of whether they occurred in the BNC; those affixes that were observed are presented in bold font. All suffixes were classified by the PoS or range of the PoS categories licensed by the suffix. Since prefixes, on the other hand, do not influence the PoS of the complex word they attach to, no PoS category was included in their classification.

2.5. Resolving ambiguities in part of speech

The ability of a grammar tagger to assign PoS accurately in every context is inevitably limited: the accuracy is estimated by the developers of the CLAWS software to be about 1.7% for the whole 100-million token BNC corpus. In addition, about 4.7% of the time, there is insufficient information for the grammar-tagger to assign an unambiguous PoS category. For example, there are contexts in which it may be ambiguous whether the headword ANTIDEPRESSANT is being used as a Noun or an Adjective: it may be considered a noun in the expression: *...just an ordinary antidepressant*, but an Adjective (pre-modifier) in the sentence: *...we have excellent anti-depressant drugs*. In these cases, the software may assign both classifications of singular noun and adjective (NN1-AJ0).

All ambiguities of this type have been resolved in the corpus listed on *MorphoQuantics*: each ambiguously coded headword was located in the original BNC transcript and the PoS recorded in accordance with the context in which it was found. Individual token frequencies are listed for each PoS. In some cases, an ambiguous PoS was assigned to a headword which was discovered upon investigation to be an instance of a speech error, such as a repetition or hesitation (e.g. ... *this has great significant--- significance*); these incorrect forms were disregarded in the data). Where no ambiguity was flagged by the grammar-tagger, the PoS was only checked if the word class assigned seemed unusual or unlikely; for example *mauvey* was assigned the PoS (NN1), singular noun, whereas on inspection of the context, it was found, unsurprisingly, to be (AJ0), an adjective, and was therefore corrected. *MorphoQuantics* lists the original PoS provided by CLAWS5 as well as any 'resolved' PoS assignment, so that users can identify the actual instance of the item in the BNC, should that be required.

3. Results and discussion

3.1. Type and token frequency data

A total of 835 distinct derivational morphemes were analysed. *MorphoQuantics* contains 17,943 unique complex word types amounting to 1,008,280 tokens; the breakdown in terms of the number of prefixes, suffixes and combining words is presented in Table 2. Given that the spoken component of the BNC contains a total of 9,963,663 tokens, it appears that around 10% of the spoken BNC is composed of complex words, although this number is a little inflated since a few headwords are included in both word-initial and word-final token counts. The numbers presented here are correct at the time of writing; building the *MorphoQuantics* database is an iterative process and, as additional affixes are added to the corpus from other sources, the totals will be adjusted accordingly.

	Affixes in BNC	*Sum of types*	*Sum of tokens*	*Types per affix class*	*Affixes NOT in BNC*
Word-initial					
▪ Prefixes: 177	96	4,067	315,051	42.36	81 (46%)
▪ Combining forms: 125	41	56	1,328	1.37	84 (67%)
▪ Both: 252	131	615	12,959	4.69	121 (48%)
Totals = 554 (66%)	268	4,738	329,338	17.68	286 (52%)
Word-final					
▪ Suffixes: 163	141	12,822	671,389	90.94	22 (13%)
▪ Combining forms: 96	61	191	3,837	3.13	35 (36%)
▪ Both: 22	20	192	3,716	9.60	2 (09%)
Totals = 281 (34%)	222	13,205	678,942	59.48	59 (21%)
Total Affixes = 835	490	17,943	1,008,280	36.62	345 (41%)

Table 2. Distributional characteristics of derivational morphemes found in the spoken component of the BNC.

As the affix totals in Table 2 illustrate, the overall number of word-initial affixes in English greatly outweighs the number of word-final morphemes when combining forms are included in the count (554/835=66%); however, as discussed in 1.1.4, combining forms have bolstered the number of word-initial types that occur in the language, particularly those which function as 'both' prefixes and word-initial combining forms. The additive contribution of combining forms and 'both' prefixes and combining forms to the total number of word-initial affixes is 45% ((125+252)/835), compared with their additive contribution of 14% to the total of word-final affixes ((96+22)/835).

When examining which categories of affixes actually occurred in the spoken component of the BNC, it appears that the numbers of word-initial and word-final affixes are more equitably represented (268:222). Half the word-initial affixes (286/554=52%) did not occur in the corpus, whereas only 21% of the word-final affixes (59/222) were not observed. This finding can be explained by the observation reported in the previous paragraph that nearly half (45%) the affixes are 'both' prefixes and word-initial neoclassical combining forms; these typically relate to medical and scientific terms and as the data show here, they are less likely to be mentioned in informal speech ((41+131=172)/(125+252=377)=46%) than in written sources. A very much smaller percentage of word-final affixes, on the other hand, contain such specialist terms (14%), and 66% of these occurred in the BNC ((61+20=81)/(96+22=118)). A more detailed account of the asymmetry between the functions and distributional characteristics of word-initial and word-final affixes can be found in Laws (b. In preparation).

The ability of prefixes and suffixes to attach to different bases (affix type size) is considerably greater than that for combining forms: on average, prefixes attach to around 42 different bases (4,067/177) and suffixes around 90 (12,822/141), as opposed to 3.9 ((56+615)/(41+131)) for word-initial and 4.7 ((191+192)/(61+20)) for word-final combining forms. This result again reflects the specificity of function of combining forms compared to the more productive prefixes and suffixes. Yet, taken as a whole, the average type size of word-final affixes is more than three times larger, at nearly 60 (13,205/222), than that of word-initial affixes, at around 18 (4,738/268), reflecting the greater ease with which affixes are attached to the end rather than the beginning of bases (Laws b. In preparation).

The token representation of word-final affixes in the spoken element of the BNC exceeds that of word-initial morphemes by more than a factor of two, but because there are three times as many word-final affix types, the overall Type-Token Ratios (TTR) are very similar: 0.014 (word-final) and 0.019 (word-initial). Therefore, although word-final affixes produce three times as many types as word-initial affixes, the comparable TTRs show that token frequency is relatively independent of type size.

3.2. Structure of *MorphoQuantics* and its associated website

The *MorphoQuantics* website provides two levels of information relating to derivational morphemes. The first of these presents a summary of all 835 derivational morphemes employed in this study (Appendices A and B) in alphabetical order on two separate pages, one for word-initial and one for word-final affixes. The type and token frequencies of each are also recorded in the summary information. Examples of those affixes not occurring in the BNC were taken from Stein (2007) or the OED. Table 3 presents an example of the table headings and relevant entries for the suffix *–ness*:

Affix & variant:	*-ness*
PoS formed by the affix:	*Ns*
PoS of the base to which the affix attaches:	*Adjs & Ns*
Real examples from BNC, Stein (2007) or OED:	*awareness, nervousness, willingness*
Language of origin:	*Old Frisian*
Etymology:	*'-nisse'*
Meaning of the affix in the language from which it is derived:	*the state, quality or condition of being* *the description denoted, or an instance of this*
Classification in Stein (2007) as to whether the affix is a prefix, a combining form, or can be used as both:	*suffix*
Type frequency:	*311*
Token frequency:	*5,064*

Table 3. Summary table of derivational morphemes in *MorphoQuantics*: example for *-ness*.

By selecting the morpheme of interest, say *-ness* in this case, the second level of information in *MorphoQuantics* is presented: here the user can view all 311 complex word types containing this suffix, as well as all the duplicate word forms of each type, e.g., *uneasiness,* *sicknesses*, with their associated token frequencies. An example of the layout of an affix-related webpage is illustrated in Table 4.

Complex Word	*BNC PoS*	*Resolved PoS*	*BNC Tokens*	*Summed Tokens*
ABJECTNESS	(NN1)	(NN1)	1	1
ADROITNESS	(NN1)	(NN1)	1	1
AGGRESSIVENESS	(NN1)	(NN1)	2	2
ALERTNESS	(NN1)	(NN1)	2	2
…..	…..	…..	…..	…..
APPROPRIATENESS	(NN1)	(NN1)	4	5
ASSERTIVENESS	(NN1)	(NN1)	19	19
ATTRACTIVENESS	(NN1)	(NN1)	5	5
AWARENESS	(NN1)	(NN1)	88	90
…..	…..	…..	…..	…..
CLEARNESS	(NN1-NP0)	(NN1)	1	1
CLEVERNESS	(NN1)	(NN1)	4	4
CLOSENESS	(NN1)	(NN1)	8	8
CLUMSINESS	(NN1)	(NN1)	1	1
…..	…..	…..	…..	…..
COMPETITIVENESS	(NN1)	(NN1)	12	12
COMPLETENESS	(NN1)	(NN1)	6	7
CONCISENESS	(NN1)	(NN1)	1	1
CONSCIOUSNESS	(NN1)	(NN1)	39	45
…..	…..	…..	…..	…..

Table 4. Example from *MorphoQuantics* showing a selection of complex words with the suffix *-ness.*

From the selection of complex words bearing the suffix *-ness* in Table 4, it can be seen that the headword CLEARNESS was ambiguously tagged by CLAWS5 as either a singular noun (NN1), or a Proper Noun (NP0); checking the occurrence of this item in the BNC transcript allowed this ambiguity to be resolved to a singular noun (NN1).

Specific examples of duplicate forms found included APPROPRIATENESS: 4 occurrences of *appropriateness* and one of *inappropriateness*; AWARENESS: 88 occurrences of *awareness* and 2 of *self-awareness*; COMPLETENESS: 6 occurrences of *completeness* and 1 of *incompleteness*; CONSCIOUSNESS: 39 occurrences of *consciousness*, 2 of *self-consciousness* and 4 of *unconsciousness*.

The researcher can download the information on each affix in the format shown in Table 4, in order to conduct further analyses on the data.

4. Concluding remarks

The searchable dataset provided by the *MorphoQuantics* website is totally unique. Frequency data of the full set of 835 derivational morphemes analysed in this study have not been recorded elsewhere, either for spoken or for written English. This electronic resource provides researchers with many useful measures, such as:

- a comprehensive set of 17,943 complex words in spoken English, from a corpus size of 1,008,280 tokens;
- a further breakdown of semantic categories of affixes that extend beyond those provided in Stein (2007);
- accurate part of speech assignments to each token;
- low frequency complex words, some of which are neologisms;
- a set of frequency norms for controlling stimulus material for use in empirical studies, such as the lexical decision task and vocabulary items for language elicitation tasks;
- baseline frequency measures for evaluating lexical growth in children, individuals with language impairments and second language learners;

- data for the comparison of speech, computer-mediated and more formal written sources of language to explore the relative usage patterns of derivational morphemes.

The website (Laws & Ryder 2014) will be released in November 2014. Future developments of *MorphoQuantics* will include the further breakdown of multi-morphemic complex words, such as *un-surpris-ing-ly*, and the coding of complex words in terms of morpheme parsability, e.g., the suffix *-age* is noticeably more salient in the derived form *orphanage* than it is in *carriage*. It is also planned to separate the affix data into the two sub-components of the BNC (DS and CG), and to add the equivalent data for the written component of the BNC, as well as spoken and written data from other sources, such as the Corpus of Contemporary American English (COCA) and the corpus of Global Web-Based English (GloWbE).

References

Anglin, J. M. (1993). *Vocabulary Development: A Morphological Analysis*. Monographs of the Society for Research in Child Development, Vol. 58. Chicago: University of Chicago Press.

Baayen, R.H. (1993). On frequency, transparency, and productivity. In Booij, G.E., & van Marle, J. (eds) *Yearbook of Morphology 1992*. Dordrecht: Kluwer Academic Publishers, 181-208.

Baayen, R.H. (2009). Corpus linguistics in morphology: morphological productivity. In Luedeling, A., & Kyto, M. (eds) *Corpus Linguistics. An International Handbook*. Berlin: Mouton De Gruyter, 900-919.

Bauer, L., Lieber, R., & Plag, I. (2013). *The Oxford Reference Guide to English Morphology*. Oxford: Oxford University Press.

Carlisle, F. (1988). Knowledge of derivational morphology and spelling ability in fourth, sixth and eighth graders. *Applied Psycholinguistics 9*, 247-266.

Clark, E.V. (1981). Lexical innovations: how children learn to create new words. In Deutsch, W. (ed.) *The Child's Construction of Language*. London: Academic Press, 299-328.

Clark, E.V. (1993). *The Lexicon in Acquisition*. Cambridge: Cambridge University Press.

Corson, D. (1985). *The Lexical Bar*. Oxford: Pergamon.

Davies, M. (2012). *BYU-BNC* [Based on the British National Corpus from Oxford University Press]. Available at <http://corpus.byu.edu/bnc>.

De Jong, N.H., Schreuder, R., & Baayen, R.H. (2000). The morphological family size effect and morphology. *Language and Cognitive Processes 15*, 329-365.

Derwing, B.L., & Baker, W.J. (1986). Assessing morphological development. In Fletcher, P., & Garman, M. (eds) *Language Acquisition: Studies in First Language Development*. 2nd ed. Cambridge: Cambridge University Press, 326-338.

Dickinson, D.K., & Tabors, P.O. (eds) (2001). *Beginning Literacy and Language: Young Children Learning at Home and at School*. Baltimore, MD: Brookes Publishing.

Freyd, P., & Baron, J. (1982). Individual differences in acquisition of derivational morphology. *Journal of Verbal Learning and Verbal Behavior 21*, 282-295.

Hay, J., & Baayen, R.H. (2002). Parsing and productivity. In Booij, G., & van Marle, J. (eds) *Yearbook of Morphology 2001*. Dordrecht: Kluwer Academic, 203-235.

Laws, J.V. (a. In preparation). The factors that determine the order of acquisition of derivational morphology in children aged between 2 and 5.

Laws, J.V. (b. In preparation). Prefixes, suffixes and combining forms: similarities and differences in function, meaning and productivity.

Laws, J.V., & Ryder, C. (2014). *MorphoQuantics*: *A Corpus of Derivational Morphology in Adult Spoken English*. Accessible at <http://MorphoQuantics.co.uk> from November 2014.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English*. Harlow: Pearson Education.

Lehrer, A. (1998). Scapes, holics, and thons: the semantics of English combining forms. *American Speech 73*, 3-28.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. 3rd ed. Mahwah, NJ: Lawrence Erlbaum.

Marchand, H. (1969). *The Categories and Types of Present-Day English Word-Formation*. 2nd ed. Munich: C.H. Beck.

Marshall, C.R., & van der Lely, H.K.J. (2007). Derivational morphology in children with grammatical-specific language impairment. *Clinical Linguistics & Phonology 21*, 71-91.

Moats, L.C., & Smith, C. (1992). Derivational morphology: why it should be included in language assessment and instruction. *Language, Speech and Hearing in Schools 23*, 312-319.

Montero-Fleta, B. (2011). Suffixes in word-formation processes in scientific English. *LSP Journal* 2, 4-14.

OED (2013). *Oxford English Dictionary – OED online*. Oxford: Oxford University Press. Available at <http://www.oed.com>.

Payne, T. (2006). *Exploring Language Structure: A Student's Guide*. Cambridge: Cambridge University Press.

Plag, I., Dalton-Puffer, C., & Baayen, R.H. (1999). Morphological productivity across speech and writing. *English Language and Linguistics 3*, 209-228.

Prćić, T. (2005). Prefixes vs initial combining forms in English: a lexicographic perspective. *International Journal of Lexicography 18*, 313-334.

Prćić, T. (2008). Suffixes vs final combining forms in English: a lexicographic perspective. *International Journal of Lexicography 21*, 1-22.

Schleppegrell, M.J. (2001). Linguistic features of the language of schooling. *Linguistics and Education 12*, 431-59.

Schleppegrell, M.J. (2004). *The Language of Schooling: A Functional Linguistics Perspective*. Mahwah, NJ: Lawrence Erlbaum.

Silva, R., & Clahsen, H. (2008). Morphologically complex words in L1 and L2 processing: evidence from masked priming experiments in English. *Bilingualism: Language and Cognition 11*, 245-260.

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

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *Quarterly Journal of Experimental Psychology 57*, 745-765.

Appendix A. List of word-initial affixes included in *MorphoQuantics*

a-$^{1-3}$
ab-$^{1-2}$
acantho-
acari-
acousto-
aero-
Afro-
agamo-
agro-
all-
allo-$^{1-2}$
alti-
alumino-
ambi-
Americo-
Amero-
amphi-
analytico-
anarcho-
andro-
angio-
Anglo-
aniso-
anomalo-
ante-
anthropo-
anti-
api-
apico-
aqua-
arch-
archaeo-
argento-
aristo-
arterio-
arthro-
astro-
audio-
Austro-$^{1-2}$
auto-
bacci-
bacterio-
balneo-
baro-
bathy-
be-$^{1-2}$
bi-
biblio-
bio-
blasto-
brito-
bromo-
bronchio-
broncho-
bronto-
bryo-
by-
caco-
caeco-
calci-
calori-
carbo-
carcino-
cardio-
carpo-
caseo-
caudo-
centi-
centro-
cephalo-
cerebro-
cero-
cervico-
chalko-
cheiro-
chemico-
chemo-
Chino-
chloro-
chole-
chondro-
choreo-
Christo-
chromato-
chromo-
chrono-
chryso-
cine-
circum-
cirro-
cis-
clerico-
co-
col-
colpo-
com-
comico-
con-
concavo-
concho-
contra-
copro-
cor-
cortico-
cosmo-
costo-
counter-
cranio-
cross-
cryo-
crypto-
crystallo-
cumulo-
cupro-
cyclo-
cylindro-
cysto-
cyto-
dactylo-
Dano-
de-$^{1-2}$
deca-
deci-
demi-
demo-
dendro-
denti-
derma-
dermato-
deutero-
dextro-
di-
dia-
dicho-
dis-$^{1-2}$
dodeca-
dorso-
duo-
dynamo-
dys-
e-$^{1-2}$
eco-
ecto-
Egypto-
eigen-
elasto-
electro-
em-
empirico-
empirio-
en-
encephalo-
endo-
ennea-
entero-
ento-
entomo-
epi-
epilepto-
equi-
eroto-
ethico-
ethno-
Etrusco-
eu-
Euro-
ex-$^{1-2}$
exo-
extra-
facio-
febri-
femino-
femoro-
Fenno-
ferri-
ferro-
fibro-
Finno-
fissi-
flori-
fluoro-
fluvio-
foeti-
fore-
Franco-
fronto-
fructi-
galacto-
gallo-$^{1-2}$
galvano-
gameto-
gamo-
ganglio-
gastro-
genito-
geno-
geo-
Germano-
geronto-
giga-
glacio-
glosso-
glotto-
gluco-
glyco-
glypto-
gono-
Graeco-
grand-
grano-
granulo-
grapho-
great-
gymno-
gynaeco-
gyno-
gyro-
haemato-
haemo-
hagio-
half-$^{1-2}$
halo-
he-
hecto-
helio-
Helleno-
hemi-
hendeca-
hepato-
hepta-
hetero-
hexa-
Hiberno-
hiero-
hippo-
Hispano-
histo-
historico-
historio-
holo-
homeo-
homini-
homo-
hyalo-
hydro-
hyper-
hypo-
hystero-
ibero-
ichthyo-
icono-
ictero-
ideo-
idio-
il-$^{1-2}$
im-$^{1-2}$
immuno-
in-$^{1-3}$
Indo-
infra-
inter-
intra
iodo-
ir-$^{1-2}$
iso-
Italo-
Japano-
Judaeo-
juxta-
kerato-
kilo-
kineto-
klepto-
labio-
lacto-
laevo-
lano-
laryngo-
Latino-
Letto-
leuko-
lexico-
ligno-
lipo-
litho-
logico-
logo-
luci-
lumino-
luteo-
lympho-
macro-
magico-
magneto-
mal-
Malayo-
masto-
matri-
maxi-
mechanico-
mechano-
medico-
medio-
Medo-
mega-
megalo-
melano-
membrano-
meningo-
meno-
mer-
mero-
meso-
meta-
metro-
micro-
mid-
milli-
mini-
mis-
miso-
mixo-
mono-
morpho-
muco-
multi-
musculo-
museo-
musico-
myceto-
myco-
myelo-
myo-
mystico-
mythico-
mytho-
myxo-
nano-
narco-
naso-
necro-
neo-
nephro-
nervo-
neuro-
Nilo-
nocti-
nomo-
non-
nona-
normo-
noso-
nucleo-
octo-
oculo-
odonto-
oeno-
off-
oleo-
oligo-
omni-
on-
onco-
onto-
oo-
opto-
organo-
ori-
ornitho-
oro-$^{1-2}$
ortho-
osmo-$^{1-2}$
osse-
osteo-
oto-
out-$^{1-2}$
over-$^{1-2}$
ovo-
oxy-$^{1-2}$
paedo-
palaeo-
palato-
pan-
pancreato-
panti-
panto-
para-$^{1-2}$
pari-
patho-
patri-
pedi-
pedo-
penta-
petro-$^{1-2}$
phago-
pharmaco-
pharyngo-
philo-
phlebo-
phono-
phospho-
photo-
phreno-
phyllo-
physico-
physio-
phyto-
picto-
pilo-
pisci-
plano-
plasmo-
pleuro-
pluri-
pluto-
pluvio-
pneumato-
pneumo-
politico-
poly-
post-
pre-
pro-$^{1-3}$
procto-
proto-
pseudo-
psycho-
pyelo-
pyo-
pyro-
quadri-
quasi-
quinque-
radio-$^{1-2}$
re-
recti-
reno-
retro-
rheo-
rhino-
rhizo-
Romano-
Russo-
saccharo-
sal-
sapon-
sarco-
scato-
schizo-
seismo-
self-
semi-
septi-
serio-
sesqui-
sexi-
she-
sinistro-
Sino-
socio-
somato-
somni-
spectro-
spermato-
sphygmo-
spiro-$^{1-2}$
spleno-
steno-
step-
stereo-
sterno-
stetho-
stomato-
stone-
sub-
super-
supra-
Syro-
tacho-
tachy-
tauro-
tauto-
taxo-
techno-
tele-$^{1-2}$
teleo-
temporo-
teno-
ter-
terato-
tetra-
Teuto-
thanato-
theo-
theologico--
theoretico-
thermo-$^{1-2}$
thigmo-
thoraco-
thrombo-
Tibeto-
tono-
topo-
toti-
toxico-
tracheo-
trans-
tri-
tribo-
tricho-
tropho-
tropo-
tuberculo-
turbo-
Turko-
Uugro-
ultra-
un-$^{1-2}$
under-$^{1-2}$
uni-
uro-
varico-
vaso-
veno-
ventri-
vermi-
vibro-
vice-
vini-
visco-
viti-
vitro-
vivi-
xantho-
xeno-
xero-
xylo-
zoo-
zygo-
zymo-

Appendix B. List of word-final affixes included in *MorphoQuantics*

- ***-a***
- ***-able***$^{1-2}$
- ***-acea***
- *-aceae*
- ***-aceous***
- ***-ad***$^{1-2}$
- ***-ade***$^{1-2}$
- ***-aemia***
- ***-age***
- ***-aholic***
- ***-al***$^{1-3}$
- ***-algia***
- ***-amine***
- ***-an***$^{1-2}$
- *-ana*
- ***-ance***$^{1-3}$
- ***-ancy***
- ***-and***
- ***-andry***
- ***-ane***
- ***-ant***$^{1-2}$
- ***-ar***
- ***-arch***
- ***-archy***
- ***-ard***
- ***-arium***
- ***-ary***$^{1-2}$
- ***-ase***
- ***-asis***
- *-ass*
- ***-aster***
- ***-ate***$^{1-4}$
- ***-athon***
- *-ati*
- ***-ation***
- ***-biosis***
- *-biotic*
- ***-burger***
- ***-cade***
- *-carpous*
- ***-centric***
- ***-cide***
- ***-cosm***
- ***-cracy***
- ***-crat***
- ***-cy***
- ***-cyte***
- ***-der***
- ***-derm***
- ***-dom***
- ***-drome***
- *-ectome*
- ***-ectomy***
- ***-ee***$^{1-2}$
- *-een*
- ***-eer***$^{1-2}$
- *-eme*
- ***-en***$^{1-4}$
- ***-ene***$^{1-2}$
- ***-ence***$^{1-3}$
- ***-ency***$^{1-2}$
- ***-end***
- ***-ene***$^{1-2}$
- ***-ennium***
- ***-ent***$^{1-2}$
- ***-er***$^{1-4}$
- *-eria*
- ***-ern***
- *-eroo*
- ***-ery***
- ***-ese***$^{1-2}$
- ***-esque***
- ***-ess***
- ***-et***
- ***-eth***
- ***-ette***$^{1-4}$
- ***-etum***
- ***-fer***
- *-fest*
- ***-fic***
- *-fid*
- ***-fold***
- ***-form***
- ***-free***
- *-fuge*
- ***-ful***$^{1-2}$
- ***-gamy***
- *-gate*
- ***-gen***
- *-geny*
- *-glot*
- *-gnomy*
- ***-gon***
- *-gony*
- ***-gram***
- ***-graph***
- ***-graphy***
- ***-gyny***
- *-hedron*
- ***-hood***
- ***-i***$^{1-2}$
- ***-ian***$^{1-2}$
- ***-iasis***
- ***-iatry***
- ***-ible***$^{1-2}$
- ***-ic***$^{1-2}$
- ***-ice***
- ***-icle***
- ***-ics***
- ***-ide***
- ***-ie***$^{1-13}$
- ***-ier***
- ***-ify***
- ***-ile***
- *-im*
- ***-in***
- ***-ine***$^{1-4}$
- ***-in-law***
- ***-ion***
- ***-ish***
- ***-ism***
- *-ismus*
- ***-ist***$^{1-2}$
- ***-ite***$^{1-3}$
- ***-ition***
- ***-it is***
- ***-itude***
- ***-ity***
- ***-ive***$^{1-2}$
- *-ivore*
- ***-ize***
- ***-kin***
- *-kins*
- *-lalia*
- ***-latry***
- ***-le***
- *-lect*
- ***-lepsy***
- ***-less***
- ***-let***
- ***-lexia***
- ***-like***
- ***-ling***$^{1-2}$
- *-lite*
- *-lith*
- ***-logue***$^{1-2}$
- ***-logy***
- ***-loquy***
- ***-ly***$^{1-2}$
- ***-lysis***
- ***-mas***
- ***-ment***
- ***-meter***$^{1-3}$
- ***-metry***
- *-mo*
- *-mobile*
- ***-monger***
- ***-mony***
- *-morph*
- ***-most***
- ***-ness***
- *-nik*
- ***-nomy***
- ***-nym***
- ***-o***$^{1-3}$
- ***-ock***
- ***-ode***
- ***-oid***$^{1-2}$
- *-olater*
- ***-oma***
- ***-on***
- *-ope*
- ***-opsy***
- *-opy*
- ***-or***$^{1-2}$
- ***-orexia***
- ***-ory***$^{1-2}$
- ***-ose***$^{1-2}$
- ***-osis***
- ***-our***
- ***-ous***
- *-para*
- *-parous*
- ***-path***
- ***-pathy***
- ***-ped***
- *-phage*
- *-phagy*
- ***-phane***
- *-phany*
- *-phasia*
- ***-phile***
- ***-philia***
- *-phobe*
- ***-phobia***
- ***-phone***$^{1-2}$
- ***-phony***
- ***-phore***
- ***-phrenia***
- ***-phyll***
- *-phyte*
- *-plasia*
- ***-plasm***
- *-plast*
- *-plasty*
- ***-plegia***
- ***-pod***
- ***-poly***
- ***-proof***
- ***-ridden***
- *-rrhagia*
- ***-scape***
- ***-scope***
- ***-scopy***
- ***-sect***
- ***-ship***
- ***-sick***
- ***-some***$^{1-2}$
- ***-sophy***
- ***-speak***
- ***-stat***
- ***-ster***
- *-stome*
- ***-stomy***
- ***-stricken***
- *-ta*
- *-taxis*
- ***-teen***
- ***-th***$^{1-3}$
- ***-thermia***
- ***-thermy***
- ***-to-be***
- *-tome*
- ***-tomy***
- ***-trix***
- ***-trophy***
- ***-ty***$^{1-2}$
- ***-ule***
- ***-ulent***
- ***-uple***
- ***-urgy***
- ***-ure***
- *-ville*
- ***-ward***$^{1-3}$
- ***-ways***
- ***-wise***
- ***-worthy***
- ***-wright***
- ***-y***$^{1-21}$

Superscripted numbers ($^{1-2,\ 1-4}$) refer to the number of variant forms an affix may have, e.g., the suffix *-ly*$^{1-2}$ attaches (1) to adjectives to form adverbs (*slowly*), and (2) to nouns to form adjectives (*friendly*). Bold affixes were observed in the spoken element of the BNC, those not in bold were not.

Jacqueline Laws is an Associate Professor of Linguistics in the Department of English Language & Applied Linguistics at the University of Reading. She lectures in grammar at all undergraduate and postgraduate levels and has also lectured in first language acquisition and theoretical syntax. Her main research interests include English derivational morphology, corpus linguistics, argument structure in Italian and Mandarin, and motion event cognition. Email: j.v.laws@reading.ac.uk.

Chris Ryder has recently completed a Master's Research degree in Applied Linguistics at the University of Reading and is preparing an application for a PhD relating to his work on the *MorphoQuantics* corpus. He is currently tutoring undergraduate seminars, having also had experience in lecturing at this level. His research interests include English derivational morphology, corpus linguistics, phonetics and phonology, and forensic linguistics. Email: c.s.ryder@reading.ac.uk.