**QUEEN'S
UNIVERSITY
BELFAST**

An industry perspective. Dealing with language variation in Collins Dictionaries

Brookes, I., O'Neill, M., & Davies-Deacon, M. (2021). An industry perspective. Dealing with language variation in Collins Dictionaries. In W. Ayres-Bennett, & J. Bellamy (Eds.), *The Cambridge Handbook of Language Standardization* (pp. 294–312). (Cambridge Handbooks in Language and Linguistics). Cambridge University Press. <https://doi.org/10.1017/9781108559249.011>

Published in:

The Cambridge Handbook of Language Standardization

Document Version:

Version created as part of publication process; publisher's layout; not normally made publicly available

Queen's University Belfast - Research Portal:

[Link to publication record in Queen's University Belfast Research Portal](#)

Publisher rights

Copyright 2021 Cambridge University Press.

General rights

Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

10

An Industry Perspective

Dealing with Language Variation in Collins Dictionaries

Ian Brookes, Mary O'Neill and Merryn Davies-Deacon

10.1 Introduction

Language variation presents a problem for a commercial dictionary publisher. The perceived aim of a dictionary is to provide a clear and definitive spelling and a definition or translation of any word that the user might look up, but a dictionary's ability to provide unequivocal explanations is compromised when a word can exist in several different forms or be pronounced in several different ways. Publishers are therefore forced to make decisions about how far a dictionary entry acknowledges language variation and how far it attempts to mask variations by offering a standardized form.

This chapter will illustrate how some Collins dictionaries deal with some aspects of language variation. After a brief discussion of the context in which these dictionaries are produced, it provides three case studies: the first looks at how dictionaries portray languages where multiple norms enjoy a similar level of prestige, using the example of Irish; the second looks at how a large monolingual dictionary of English acknowledges the existence of variation within the language; and the third investigates how variation in spelling, pronunciation and lexis are dealt with in two English dictionaries within the Collins range designed specifically for learners of English. Finally, the focus will shift to the impact of technology on the process of publishing dictionaries, showing how different considerations come into play when creating materials for online access and suggesting how this may bring about opportunities for different approaches to variation and standardization.

10.2 The Collins Dictionary Range

Collins, an imprint of HarperCollins Publishers, is a well-known and long-established dictionary publisher based in the UK, producing a range of dictionaries including its flagship product, the *Collins English Dictionary (CED)*,

the thirteenth edition of which was published in 2018. While the *CED* itself dates back only to 1979, the tradition of dictionary publishing at Collins is far more extensive, going back to the early nineteenth century.¹ The volumes published at this time were typically inexpensive illustrated dictionaries; they reflected the values of the company's founder, William Collins (1789–1853), who established it with the professed hope of widening access to knowledge, intending in this spirit 'to publish a dictionary that everyone could afford' (Brookes & O'Neill 2018 (hereafter *CED* 2018): ii).

The values of affordability and accessibility remain evident in the publisher's current range of products, which cater for a wide range of users: dictionaries are available in various sizes, ranging from the large *CED* and equivalently sized bilingual dictionaries to the small, inexpensive *Gem* dictionaries that provide essential vocabulary in a range of languages. As dictionaries become smaller and cheaper, decisions have to be made about how many words are included and in what detail, and so there is less opportunity to reflect the full range of variation within a language.

Collins publishes a very wide range of monolingual and bilingual dictionaries, as well as language-learning courses and phrase books. Monolingual English dictionaries are available for native speakers as well as learners of English as a foreign language. The bilingual dictionaries typically pivot around English, although other language pairs, such as French–Italian and Spanish–German, are available.

10.3 Treatment of Languages with Multiple Norms

As a language with multiple norms, Irish is a particularly relevant case study for questions of standardization. There are three traditional dialects, Ulster, Connacht and Munster, as well as an official standard that focuses on spelling and grammar. Marked phonological differences among the dialects persist (Ó Baoill 2010: 173; cf. also Chapter 28 in this volume), to the extent that in the Gaeltacht areas, 'each dialect area still maintains its own fairly well-defined speech forms' (Ó Dochartaigh 1992: 29). The acquisition of a particular dialect is especially important for new speakers, who may feel under pressure to adopt a dialect that resembles traditional Irish in order to assuage concerns over authenticity (O'Rourke & Walsh 2015: 64); the dialects are thus more prestigious than the standard in certain contexts. The use of dictionaries is nonetheless crucial for new speakers, who will often be acquiring Irish in a context of formal instruction. While dialectal differences most obviously concern matters of pronunciation, differences in vocabulary can also be observed, and they are clearly of more immediate relevance to lexicography: for example, the Ulster dialect uses *bomaite* ('minute') rather than *nóiméad* and *fosta* ('also') rather than *freisin*.

¹ See www.collinsdictionary.com/about (accessed 17 July 2019).

Another crucial matter for Irish is the issue of borrowed versus coined vocabulary. In the context of minoritized languages, where a word for a new or technical concept must be adopted, borrowings are often considered concessions to the dominant language and avoided by more militant speakers, despite typically being favoured by certain groups of traditional speakers. Neologisms, on the other hand, can be condemned for an apparent purism in their avoidance of borrowed elements: Rottet (2014: 214) outlines the situation in the case of Breton, noting that the variety typically used by certain militant speakers 'is criticized for making liberal use of artificially created coinages' (cf. also Chapter 12 in this volume). In many cases, a neologism and a borrowing coexist, each bearing different connotations and favoured by a different group of speakers. Examples in Irish include *carr/gluaisteán* ('car') and *fón/guthán* ('phone').

How does a dictionary deal with these matters? Answers can be sought in two places: firstly, investigating the front matter can determine the extent to which the dictionary makes these issues explicit and how it proposes to deal with them; secondly, the body of entries itself shows how these decisions play out in practice. The decisions made in the case of Irish can be illustrated with reference to three Collins Irish dictionaries: the *Easy Learning* title (Beattie 2016), intended for school pupils, and the *Pocket* and *Gem* editions (Comer et al. 2019a; Comer et al. 2019b), aimed at adults. The *Easy Learning* dictionary covers 16,000 entries, with a more explanatory style and numerous full example sentences demonstrating words in use. The *Pocket* and *Gem* dictionaries give more concise treatments of entries but cover a larger number of headwords (27,000).

Beginning with pronunciation, the *Pocket* edition is the only one of the three to discuss this subject explicitly. In the case of the *Easy Learning* dictionary, pronunciation is therefore left in the hands of the teacher, who can provide dialectally appropriate pronunciations; the *Gem* does not offer the requisite amount of page space to cover the subject. The pronunciation guide in the *Pocket* dictionary begins by foregrounding the coexistence of the three main dialects and notes throughout that there are differences among the three, using footnotes and separate sections to discuss specific varieties as necessary. For example, with regard to the letter combinations bh/mh at the end of words and after long vowels and diphthongs, a footnote explains that broad (or velarized) bh and mh are pronounced *v* in southern dialects but *oo/w* in northern dialects.

At times, the *Pocket* dictionary acknowledges further variations within the main dialects: for example, one group of pronunciations is shown as being used in 'parts of Munster and Connacht' (Comer et al. 2019b: ix). While this shows that there is a large variety of dialectal pronunciations, it does not go so far as to explain which are the specific areas in question; however, indicating that these differences exist means that users are able to seek more information should they require it. The body of the dictionary, like in the *Gem* and *Easy Learning* titles, gives no information on the

pronunciation of individual words, avoiding the prioritization of a single variety, the inclusion of excessive complex phonological information for this level of user and the significant loss of space that would occur if multiple pronunciations were indicated.

The examples of *bomaite/nóiméad* and *fosta/freisin* noted above can be used to investigate the dictionaries' treatment of dialectal variation with regard to vocabulary. While all three dictionaries are small, they all contain both *bomaite* and *nóiméad* as headwords in the Irish–English section, and both the *Pocket* and *Gem* editions give both words as Irish equivalents for the English headword 'minute', although the *Easy Learning* edition gives only *nóiméad*. This may at first sight suggest a slight bias away from the Ulster dialect, but the case of *fosta* and *freisin* attests the opposite: again, both are listed as headwords in the *Pocket* and *Gem* dictionaries, but both sides of the *Easy Learning* dictionary give only *fosta*, the Ulster word. This indicates that decisions about vocabulary are not based on the perceived status of the dialects, but result instead from other considerations, most notably the need for simplicity for the user at this level. There is no deliberate emphasis on one particular dialect, and at no point is the dialect associated with a particular word specified.

Overall, these examples show that Collins Irish dictionaries typically take an inclusive approach, including all of the major dialects of Irish on an equal footing, and thereby catering for users of multiple varieties. The exception to this is the *Easy Learning* dictionary, particularly in its English–Irish side, which provides a single Irish word. This is a policy decision, made on pedagogical grounds, that *Easy Learning* titles should give a single translation so as not to confuse learners. However, in the Irish–English side of this dictionary, both possible words are present in most of the examples examined, which allows learners to understand texts from a variety of sources. In general, therefore, the focus is on describing all forms in use – notwithstanding limitations of space – without attributing prestige to any particular variety.

10.4 Variation in Dictionaries of English

Conflicts among multiple linguistic norms have specific consequences in minoritized language contexts, but are also present, with different effects, in more widely spoken languages. English is an obvious example, where different norms have become established in different jurisdictions due to the language being spoken all around the world. Given that this is the language for which the largest number of Collins dictionaries, and the largest Collins dictionaries, are produced, their treatment of such variation can be examined extensively.

10.4.1 Treatment of Linguistic Variation in *Collins English Dictionary*

CED is a synchronic dictionary (Hanks 2010: 582), focusing on English as it is currently spoken rather than on the historical development of the language. The current online edition contains over 185,000 entries with more than 260,000 definitions. It does not restrict itself to describing English as spoken in the UK, and it includes variants from across the English-speaking world. Earlier editions of the dictionary contained an essay on 'The Development of English as a World Language' (Aitken et al. 1979), which gave a survey of the regional dialects of England, Scottish and Irish English, American English and English as spoken in Canada, the Caribbean, Australia and New Zealand, Africa and Asia, as well as creoles and pidgins. This essay provided a historical account of the development of English in different parts of the world and offered examples of regional variation in both pronunciation and vocabulary. Although the essay is not included in more recent editions, the front matter notes that 'it has always been a concern of *Collins English Dictionary* to reflect English's status as a global language' (*CED* 2018: vii).

How far this concern with the different varieties of English is reflected in the main text of the dictionary can be examined with regard to lexis, phonology and orthography.

10.4.1.1 Lexis

The dictionary's front matter states that 'words or senses restricted to or associated with a particular country or region are labelled accordingly' (*CED* 2018: xvi), and that the most frequent labels used are to indicate variations restricted to Australia, Britain, Canada, the Caribbean, Ireland, New Zealand, Scotland, South Africa and the USA. These labels appear in the dictionary in italics:

brak² (brak) *n* *South African* a mongrel dog

A second type of lexical variation is acknowledged by the presence of labels which mark the register of a word, indicate items that are regarded as slang, informal, taboo, offensive, derogatory, not standard, formal, archaic and so on. These are indicated in the same way:

peke (pi:k) *n* *informal* a Pekingese dog

These two types of label can also be combined to indicate words that differ from the standard form in terms of both region and register:

bitser ('bitsə) *n* *Austral informal* a mongrel dog

² Statistics used here and elsewhere in this chapter are taken from the database used to create the online version of the dictionary on CollinsDictionary.com. This version is slightly larger than the most recent printed edition of the dictionary and contains material added since the publication of that volume.

It should be noted that not all variations are indicated by an italic label. A second way of showing variation occurs when a word is cross-referred to another entry and regional or register information is incorporated into the wording that introduces the cross-reference, as in these cases:

dug³ (dʌg) *n* a Scot word for **dog**

pooch (pu:tʃ) *n* a slang word for **dog**

Considering only the example of words that indicate dogs, the dictionary offers regional variations from India (*kuta*, *kuti*), Ireland (*bowler*) and New Zealand (*kuri*, *tripehound*), besides those noted above, while there are several further informal and slang variations (e.g. *mutt*, *tyke*).

The dictionary's coverage of the main regional variants is indicated in Table 10.1, which lists the number of occurrences of different regional labels:²

While arguments could be made about the extent to which the different regional forms of English are covered – although many regional usages are marked, as are Scottish, Irish and Welsh – it is clear that the dictionary compilers have actively sought to include words from all of these major English-speaking areas.

Before leaving the subject of lexis, it is worth noting that the dictionary has a further strategy for alerting users to regional variation. This is by the inclusion of regional forms as variants at an equivalent entry, as in this pair of entries:

court card *n* (in a pack of playing cards) a king, queen, or jack of any suit.

US equivalent: **face card**

face card *n* (in a pack of playing cards) a king, queen, or jack of any suit.

Also called (in Britain and certain other countries): **court card**

Table 10.1 *Frequency of different regional labels in Collins English Dictionary (online edition).*

Region	Occurrences of regional label	Occurrences in introductions to cross-references
Australia	1,696	14
Canada	1,704	47
Caribbean	75	0
India	81	4
Ireland	237	20
New Zealand	1,204	4
Scotland	1,611	174
South Africa	421	8
USA	6,311	597

Here, the dictionary not only covers regional variation, but actively draws the reader's attention to the fact that in American English only *face card* is used, whereas in other places both forms are accepted.³

10.4.1.2 Phonology

The front matter of *CED* announces that 'pronunciations of words in this dictionary represent those that are common in Standard English',⁴ going on to state that 'when a headword has an acceptable variant pronunciation or stress pattern, the variant is given by repeating only the syllable or syllables that change' (*CED* 2018: xiv).

However, the vast majority of *CED* entries contain only one pronunciation form. Of the 165,000 places where pronunciation is shown, over 155,000 (94 per cent) contain a single form, while fewer than 500 (0.3 per cent) have more than two forms shown. The highest number of different pronunciations for a word is five, which occurs twice, at *Byzantine* and at the plural form *causes célèbres*:

Byzantine (bɪ'zæn, taɪn, -ti:n, baɪ-, 'bɪzən, ti:n, -taɪn)
cause célèbre ('kɔ:z sə'leɪbrə, -'leɪb, French kɔz se'leɪbrə) *n, pl causes célèbres*
 ('kɔ:z sə'leɪbrəz, -'leɪb, 'kɔ:zɪz sə'leɪbrə, -'leɪbz, French kɔz se'leɪbrə)

By representing pronunciations in what it calls the 'Standard English' form, the dictionary commits itself to a single phonetic system, and regional and national variations are regarded as being largely beyond its scope. However, as the example of *causes célèbre* above shows, loanwords often have variants that indicate a pronunciation used in the source language.

There are also occasional variants indicated from English-speaking regions:

bream¹ (bri:m, Austral brɪm)
tomato (tə'mɑ:təʊ, US tə'meɪtəʊ)

However, there are only forty-six occurrences of this sort of regional pronunciation variant listed in the entire dictionary, suggesting that only especially striking and noteworthy instances are included.

In contrast to its policy regarding lexis, where variation appears to be embraced, the policy of the dictionary with regard to pronunciation is thus one of standardization, with regional and non-standard forms only occasionally being acknowledged. There may also be practical considerations behind this policy: lexical variants are relatively easy to identify and categorize, while variation in pronunciation is more difficult to determine and the variants are potentially limitless, making a complete analysis beyond

³ There are 163 occurrences of regional variants that are presented as 'equivalents', as well as further variations on this theme in other entries.

⁴ As a dictionary of British English, the entry for 'Standard English' defines it as 'English which is characterized by idiom, vocabulary, etc. that is regarded as correct and acceptable by educated native speakers'. The term is more neutral than 'Received Pronunciation'.

the scope of the dictionary. Giving precedence to (usually) one variant has certain advantages in terms of presenting information neatly – the example of *causes célèbres* above shows how cluttered a dictionary entry can become when multiple variants are shown – but it obviously does not give the full story of variations in pronunciation within English.⁵

10.4.1.3 Orthography

CED acknowledges that a word can have more than one acceptable spelling from, noting in the front matter that ‘common acceptable variant spellings of English words are given as alternative forms of the headword’ and that ‘where different, US spellings are also recorded in the headword’ (*CED* 2018: xiii). It is notable that while the dictionary acknowledges orthographic variation, it restricts its coverage of variant forms not only to those that are ‘acceptable’, but also to those that are ‘common’. This suggests the presence of barriers to inclusion based both on adherence to linguistic norms and on the frequency with which forms occur in the corpora used by lexicographers as evidence for linguistic use.⁶ The lexicographers reserve the right to exclude forms that are linguistically acceptable but infrequent and also those that are frequent but deviate from normal standards.

The large majority of words listed in the dictionary have no variant shown. There are more than 212,000 primary spelling forms in the data, and over 92 per cent of these have no variant shown.⁷ Only sixteen entries in the entire dictionary have more than four variant forms shown, although one remarkable entry, for the adjective *paradisal*, shows nine different forms:

paradisal (ˌpærəˈdaɪsəl), **paradisial** (ˌpærəˈdaɪsiəl), **paradisian** (ˌpærəˈdaɪsiən),
paradisic (ˌpærəˈdaɪsɪk), **paradisical** (ˌpærəˈdaɪsɪkəl), **paradisiac** (ˌpærəˈdaɪsɪæk), **paradisiacal** (ˌpærəˈdaɪsɪəkəl), **paradisaic** (ˌpærəˈdaɪsɪk) or **paradisaical** (ˌpærəˈdaɪsɪkəl) adj of, relating to, or resembling paradise

Of the 16,000 spelling forms that are presented as variants, about 10 per cent are labelled as regional forms, most of these being US spellings:

aesthesia or **US esthesia** (iːsˈθiːziə)
cosy or **US cozy** (ˈkɒʊzɪ)

remove artificial line breaks so that the quotation is continuous, with all lines except the first indented equally

⁵ The dictionary’s policy with regard to phonological variation was, of course, originally influenced by the limitations imposed by print dictionaries. The digital format opens up the possibility of covering a much wider range of variation in pronunciation, as is noted in Section 10.6.

⁶ Work on updating the *CED* is now driven by the Collins Corpus, which continuously gathers examples of English in use and contains over 9 billion tokens.

⁷ This standardization to a single form is typical of the practice of single-volume dictionaries of English, but is markedly different from the practice of larger historical dictionaries such as the *Oxford English Dictionary* or the *Scottish National Dictionary*.

A further 10 per cent are accounted for by the dictionary's policy of showing variant forms ending in *-ise* for verbs whose primary spelling is *-ize*:

globalize or globalise ('gləʊbəˌlaɪz)

This set of words presents an interesting case, as it is commonly perceived that the *-ise* form is British and *-ize* an Americanism. However, the choice to use *-ize* as the primary form can be justified from an etymological perspective as the suffix goes back to the classical Greek *-ίζειν* (*-izein*; Ishikawa 2011: 390), and in practice both forms are used in British English. Thus, *-ize* can be considered the more globally acceptable term, and it is consequently adopted as the primary form in most dictionaries published in the UK, including the *Oxford English Dictionary*.⁸

While *CED* acknowledges a certain amount of orthographic variation, it does so within a structured framework. It generally offers a single approved spelling form and restricts the number of acceptable variants to a minimum, often following regular patterns (as in the case of words ending in *-ize*) or regional forms (as in words formed from the prefix *haem-*, where the US variant *hem-* is routinely shown). It includes very few spellings regarded as non-standard, although it does include entries for a handful of very common non-standard forms (such as *yoof* and *nite*), always taking care to label these as such.

10.4.2 Treatment of Linguistic Variation in Two English Dictionaries Used in Education

Section 10.3 showed how learners' dictionaries may follow different principles from more traditional titles, as in the example of the *Easy Learning* dictionary of Irish, which was seen to offer a single equivalent for a number of headwords in the English–Irish side for which other dictionaries gave two. Some other particularities of learner dictionaries will be highlighted here with reference to two educational dictionaries of English: *COBUILD Advanced Learner's Dictionary* (Sinclair 2018) and the *Collins English School Dictionary* (Brookes et al. 2018). In both cases, the editions examined here are those produced for the UK market, which default to British English where necessary: the effects of this will be seen below.

These dictionaries are aimed at two different user bases: the *COBUILD* dictionary is aimed at adult non-native speakers of English in the advanced stages of learning (in the case of this particular edition), while the *School* dictionary targets native speakers of English in the first few years of secondary school and consequently assumes a certain degree of native-speaker knowledge of English. Users of both dictionaries are therefore in a sense learning English, but this must be interpreted quite differently in each

⁸ See <https://web.archive.org/web/20190115111516/https://blog.oxforddictionaries.com/2011/03/28/ize-or-ise> (accessed 25 July 2019).

case: *COBUILD* dictionary users are acquiring English as a second language, but as adults and at an advanced level, and they are more likely to require explicit grammatical rules and explanations of nuances between near-synonymous terms. Users of the *School* dictionary, on the other hand, are more likely to be using dictionaries to check spellings and for looking up the more unusual or difficult words that they would come across in a school setting. The *COBUILD* dictionary is also more likely to be used in self-study, while the *School* dictionary might be kept in schools and used under the supervision of a teacher.

10.4.2.1 Pronunciation

The front matter and supplements of the *School* dictionary offer no information on pronunciation. Some entries in the body of the dictionary, typically less assimilated borrowings that do not follow the usual rules of English pronunciation, include a pronunciation spelling (thus we find ‘said laz-zan-ya’ under the headword *lasagne*). Such pronunciation spellings appeal to native speakers’ intuitions as to the standard sounds represented by English vowels and might not be intelligible to non-native speakers, but this is not an issue for this dictionary, as it is intended for contexts where English is used as a first language.

As a dictionary not aimed at native speakers, the *COBUILD* dictionary gives much more detail about pronunciation. Each individual entry in the dictionary includes a pronunciation written using a slightly modified form of IPA.⁹ The front matter advises its readers that its pronunciations are ‘based on the two most widely taught accents of English’ (Sinclair 2018: xxv), which it states are Received Pronunciation (RP) and General American (GA).

It is perhaps significant that the dictionary uses the phrase ‘widely taught’ rather than ‘widely spoken’, as RP is known for its lack of resemblance to the varieties used by most British English speakers, having been characterized as a ‘minority accent’ (Trudgill 2001: 171). Moreover, RP ‘appears to be changing quite rapidly’ (Kerswill 2006: 48), raising questions about which specific pronunciations should be considered part of RP, or indeed whether the very concept of RP is still relevant: a number of other names have been proposed for the contemporary ‘standard variety of English spoken in south-eastern England’ (Saito 2015: 98).

The pronunciation style advocated in *COBUILD* on the whole conforms to a more conservative variety of RP: the monophthongization of the diphthongs /ʊə/ and /eə/ as /ɔ:/ and /ɛ:/, respectively (leading in the former case to the loss of some minimal pairs in most forms of contemporary RP; e.g.

⁹ There is a single system of IPA because it is necessary to approximate the pronunciations in a general dictionary like this, which does not claim to be a specialized phonetics text. So, for instance, the vowel sound in the word ‘poke’ is shown as /oʊ/ as a compromise between the RP /əʊ/ and the GA /o/. Some significant regional differences are noted at individual entries, so the entry for ‘bath’ has both RP /bɑːθ/ and the form /bæθ/ which is used in many other accents.

sure and *shore*; Saito 2015: 99), is not reflected in the *COBUILD* dictionary. Less important in terms of the effect on minimal pairs, but still notable, is the use of /u:/ (e.g. in *goose*) and /æ/ (e.g. in *trap*) rather than the newer realizations /ʉ:/ and /a/ noted by Kerswill (2006: 49). It could be argued that this conservative approach perpetuates the attribution of prestige to a variety spoken only by the elite and the concomitant devaluing of other types of English. On the other hand, a standard of some kind has to be nominated for these purposes, and RP is a viable option despite its shortcomings, as it ensures that learners will be widely understood.

10.4.2.2 Lexis and Orthography

In the matter of lexical and orthographic variation, words can be grouped into a number of categories. Where a small orthographic difference exists between British and American forms, the *COBUILD* dictionary's typical approach is to include both forms as headwords, but for the American variant, instead of providing a definition, to redirect the user to the British version. Hence, *center* appears with the instruction 'see *centre*', *cozy* with 'see *cosy*', and so on. In the British English entry, the user is correspondingly given a warning in the form 'in AM, use *center*'. In the US edition of this dictionary, the process is reversed, with the American spellings receiving full definitions and the British ones redirecting the user.

Where the difference is more substantial than orthographic variation, both headwords receive full definitions: both *nappy* and *diaper* appear, as do *pavement* and *sidewalk*, and so on. Keeping these entries separate means that a single appropriate pronunciation can be given in each case – it might be confusing to see the pronunciation /daɪəpər/ under the headword *nappy*, for example – as well as allowing for any subtle semantic differences, such as between *fresher* and *freshman*. Again, pairs of entries of this type cross-reference each other in the same way as the spelling variants. The instruction to 'use X' (e.g. 'in British English, use *nappy*') is a reminder that this is a learners' dictionary, for users seeking instruction on how English should be used.

While this approach insists on the equal validity of both variants, it can be seen, in the case of the UK edition of the *COBUILD* dictionary, to place slightly greater priority on British forms, in cases of orthographic variation redirecting users from American to British headwords and printing the full definition in only the British entry. As noted, the converse approach is taken in the US edition. There are, however, also a few cases in the UK edition where vocabulary that may be commonly thought of as American takes priority. Thus, in cases where a distinction is widely perceived as mapping onto these two geographical varieties, but the 'American' variant is also used in British English, it is this variant that takes the definition, while its 'British' counterpart shows merely 'see ...'. *Enquire* and *yoghurt* are two examples of this. Similarly, *-ise* spellings are not included in headwords, but

at the end of the corresponding entry the user is informed that they may ‘also use [e.g.] *fertilise*’ in British English.

The *COBUILD* dictionary thus provides users with comprehensive and systematic information on the differences between British and American spelling and vocabulary. The *School* dictionary takes a different approach. As mentioned previously, this is targeted specifically at children using English within the UK, and British variants thus take precedence, although there remains the need to explain American vocabulary to children when it occurs, as it is likely that they will encounter texts of American origin. Variation that concerns only orthography is not indicated: there is no entry for *fetus*, for example, and the *foetus* entry gives no indication that the alternative spelling exists. In cases of lexical variation, the American word is listed as an entry, but this typically makes explicit reference to the British equivalent: for example, the definition of *diaper* is, ‘In American English, a diaper is a nappy’. This is more explicit than using ‘see ...’, and consequently perhaps more suitable for children with little experience of using dictionaries. For *-ise/-ize* spellings, the system found in the *COBUILD* dictionary is reversed: *-ise* spellings are used in headwords, with the note ‘also spelt [e.g.] *fertilize*’ included at the end of relevant entries. This again reflects the fact that the dictionary is aimed at a UK audience, for use in education: *-ise* spellings are advocated in UK curriculum guidelines.

This section has shown how different audiences can determine how variation and standardization are treated across dictionaries. Nonetheless, traditional print books are no longer the main focus of much of the work of the Collins dictionary team, particularly in the case of English dictionaries. The impact of technology and particularly the Internet has caused major changes in how dictionaries are produced and distributed, leading to the potential for significant changes in how variation and standardization are handled. The following section will discuss this with regard to the use of digital tools and the production of online dictionaries at Collins.

10.5 Technological Innovations: *CollinsDictionary.com*

Collins was an innovator in the use of corpora for dictionary production. Since the creation of its first corpus in the 1970s, reliance on corpora has become a standard feature of dictionary production, and as the capabilities of relevant technologies have increased, so have corpus sizes: Atkins and Rundell (2008: 58) note that ‘corpora have increased in size by roughly one order of magnitude in each decade since the 1970s’. This has allowed a more expansive approach to corpus collection, allowing the identification of rare and newer vocabulary. The large Collins monitor corpus, to which new examples of language are added on a monthly basis, helps lexicographers to describe the most up-to-date forms of the language. Querying the corpus using the commercial Sketch Engine program (first developed in

2003 and used by many dictionary publishers) allows the easy identification of frequent collocations, an important feature of COBUILD dictionaries in particular.

While these technological innovations make certain in-house processes easier and allow the use of more sophisticated methods than was previously possible, significant technological development has equally occurred on the client-facing side. The growth of the Internet and World Wide Web for consumer use has of course been instrumental in changing how members of the public interact with services, notably in the development of free online encyclopaedia facilities – Wikipedia being the best known – as well as dictionaries with a variety of remits and operational models, such as Wiktionary, dictionary.com, Urban Dictionary and others. To compete in the contemporary market, traditional dictionary publishers must also provide Internet-based services and hope to compete with these online-only providers by harnessing their prestige as well-known companies. As a result, most large dictionary publishers now operate online dictionaries: indeed, in the case of Macmillan, this is now the only version available of its principal English dictionary, the print version of which was discontinued in 2012.¹⁰

Since late 2011, Collins has operated a stand-alone website at <https://collinsdictionary.com>, initially providing dictionaries of English, French, German and Spanish. As of mid-2019, the site offers monolingual dictionary content in English, along with bilingual content allowing translation of words between English and seven other languages (Chinese, French, German, Hindi, Italian, Portuguese and Spanish); thesaurus content for British and American English; English word lists; a grammar section concerning English, Spanish, French, German and Italian; additional grammatical tools for COBUILD dictionary users; and a machine translation service. It also contains a quiz bank where users can test themselves on the spellings and definitions of words, a Scrabble section (Collins being the official publisher of all Scrabble-related books) and a blog comprising posts that typically highlight words on a particular, often topical, theme, with a few of these uploaded each month.

Online dictionaries are able to implement a number of features that offer a more user-friendly experience than print dictionaries. For example, the Collins site contains a predictive search function that suggests possible entries as the user types in a word to look up: this can aid those who are unsure about spelling. If the user nonetheless proceeds to look up a word that is not among the suggestions, a list of possible correct versions is displayed, with a link to each entry. Entry pages also contain links that cross-reference other entries, allowing users to look up similar words without having to type them in and to explore related content such as synonyms, collocations and idioms. All of these features make the online version of

¹⁰ See www.macmillandictionaryblog.com/bye-print-dictionary (accessed 26 July 2019).

a dictionary easier to use and potentially more informative than its print equivalent, especially for users who begin with a lower level of competence in English: in this respect, the online dictionary is particularly useful for learners.

Certain obvious limitations of space and media in print dictionaries can also be surmounted by the online format. Entries in the online dictionaries include audio and video content demonstrating pronunciations, high-quality images where appropriate and illustrative quotations from the Collins Corpus. Some of these features would be impossible to include in a print dictionary, while the inclusion of others would be limited by the constraints of space. The online format is also able to avoid some of the more limiting features of the print version of *CED*. Thus, parts of speech are given in full rather than represented by abbreviations, while etymological content is introduced by the phrase 'word origin' rather than being placed in square brackets.

Maintaining an online dictionary also allows Collins to track trends in how the facility is used by monitoring traffic to the website: visitors' geographical provenance can indicate which specific varieties of world English need more representation, while information on which words are looked up gives an indication of the semantic fields that need particular development.

Maintaining an online dictionary also requires editors to be constantly aware of lexical and semantic change. On the Internet, data must be kept up to date, reflecting not only the coining of new terms, but also changes in the meaning of existing vocabulary, and such updates must be applied across the board. For example, changes of name such as that from Swaziland to eSwatini in 2018 must be reflected not only in that particular entry but anywhere on the site that the word appears, including in word lists and quizzes, which use different data sets from the main body of entries. Certain words relating to health and wellness are also in flux as older terms have acquired negative connotations and lost their clinical status. This is indeed reflective of a general trend towards more inclusive, less stigmatizing language in a number of contexts, such as gender and sexuality, disability and ethnicity. In other fields, terms currently in widespread use could be rendered irrelevant by future political or legal changes. A balance must be struck between reflecting language currently in use and not according undue legitimacy to words that may turn out to be in use only for a very short time.

Another issue heightened by the existence of the online dictionary is that of accountability and interaction with the public. In the past, those who took issue with the contents or functionality of a dictionary would have to take the trouble to contact the editors by letter. Now, on the other hand, most communication from the public occurs online, and the volume of correspondence has increased. This means that a significant amount of editors' time is spent addressing queries and issues raised by users.

The advent of social media has also changed how consumers interact with companies in the sense that such interactions may now take place on

a public platform. Collins maintains a Twitter account (@CollinsDict) with just over 13,000 followers as of July 2019, as well as a Facebook page with over 12,000 followers. Both of these accounts are mostly used for sharing 'Word of the Day' content that links to specific entries on the dictionary website, as well as new blogposts. Having a presence on these public platforms, where users are able to comment on posts and openly share their opinions, further increases accountability: such communications may be seen and amplified by users' social media contacts, and responses from Collins are visible and subject to scrutiny. For this reason, Collins carefully monitors what it puts online.

Interaction with the public also comes in the form of a crowdsourcing function on the Collins dictionary website. As with sites such as Wiktionary and Urban Dictionary, any user may propose a word along with a definition, with the intent that these be added to the dictionary. This feature aims to make users of the site feel personally involved in the dictionary-creation process and also taps into the public as a source for drawing attention to recently coined words. Unlike some sites with similar functions, however, Collins undertakes a rigorous evaluation process for each submitted word, screening every submission so that offensive or irrelevant material is not displayed in the site and seeking to remove duplicate suggestions. Any eventual inclusions are published only with definitions written by lexicographers, as there is a need to maintain the lexicographical standards traditionally employed in the production of its dictionaries.

Crowdsourcing dictionary content aims to break down the boundary between creator and consumer, emphasizing that dictionaries should not be seen as prescriptive authorities. Submitters' specific fields of knowledge may help expand the content of the dictionary in areas where editors might not have realized the need for such development, while contributions from users around the world help broaden the dictionary's repertoire of terms used in varieties of English other than those spoken in the UK and USA. This process can also identify new and relevant vocabulary: the term *Brexit* was suggested by a user as early as mid-2013.¹¹ The words suggested are a major source of expansion not only for the dictionary website, but also for printed editions of the *CED*; in its twelfth edition in 2014, nearly 600 out of 800 new words and senses were derived from user-generated content.

Some limitations of the system can be illustrated with a brief analysis of the sixty-six words submitted to the dictionary in April 2018. Of these words, fifteen were submitted by administrators (i.e. members of the editorial staff and associated freelance lexicographers), while the remaining fifty-one (i.e. just over three-quarters of submissions) came from ordinary users. The submissions came from various countries:¹² there were fifteen contributions each from the UK and USA and five from India, as well as

¹¹ See www.collinsdictionary.com/submission/10549/Brexit (accessed 29 July 2019).

¹² As indicated by users' own specification of their location in the site's member registration form.

eight more from other English-speaking countries, making up about 90 per cent of submissions in total. The remaining seven submissions came from countries where English is not an official language.

While users are not required to provide a great deal of information about themselves, the details that can be gleaned from account statistics are still of interest. Half of the submissions in April 2018 from ordinary users came from users who had created their account within the previous week, presumably with the specific aim of submitting that particular term. Some, on the other hand, are prolific and long-standing contributors: nine submissions in April 2018 came from users who had registered over 2000 days earlier. These longer-registered users appear to be the chief contributors of publishable content: all of the submitted words from the sample that have since been published in the online dictionary came from accounts that at the time were at least 900 days old.

It can also be noted that the action of submitting words appears to appeal disproportionately to men. While users are asked for no information on their gender, they typically provide their first name upon registration, which means that the gender profile of participants can be estimated. Thus, forty-five of the fifty-one submissions in April 2018 that came from users other than administrators allow the tentative assignation of a gender to the contributor, and out of these, only eleven appear to have come from women. This significant gender gap echoes that found in other participative online contexts, most notably Wikipedia, where studies consistently show that fewer than 20 per cent of editors are women (16.1 per cent according to Hill & Shaw 2013: 3). The underrepresentation of women in this domain may also have reinforced gender biases in corpora that sample language: nearly twenty years ago, Hidalgo Tenorio (2000: 228) concluded that COBUILD's efforts to reflect 'language in use' at the time in fact attested to a 'discriminatory, prescriptive, outdated use'. Today, lexicographers at Collins are well aware of the biases inherent in the evidence about language that they receive through both corpora and user-generated submissions.

10.6 Variation, Standardization and Future Directions

The case studies presented in this chapter showed how issues around variation and standardization are dealt with in different print dictionary products. The perceived readership of the dictionary can lead to different approaches: dictionaries aimed at school pupils tend to prioritize a single form of each word, but nonetheless equip users to understand other language varieties; dictionaries aimed at the mainstream are more likely to allow for variant forms, although the number of variants shown is restricted and excludes rare and non-standard forms. This is in keeping with the concept of a dictionary as an object of standardization, appealing

to norms accepted by educated native speakers. The dictionaries also present standardized systems of pronunciation, although these too are tailored to perceived readerships. In acknowledging a degree of variation while still presenting established forms as standard, Collins attempts to create dictionaries that are on the one hand accessible and easy to use, and on the hand reflect the variation in language use around the world.

Some of the advantages brought about by the online dictionary format have the potential to affect questions of variation and standardization, and there are certain significant differences in how such matters are handled on the CollinsDictionary.com website. Instead of combining content from British and American English, the website presents British and American material as two separate blocks within each entry page, along with a third block containing the COBUILD dictionary definition if one exists. This means that there is often near-duplication of material between British and American versions, particularly if the word has the same form in both varieties. It does, however, allow for subtle differences of meaning and cultural references, as well as pronunciation, to an extent that would not be possible in print dictionaries, again for reasons of space. In fact, the American entries displayed on the website are licensed from two American dictionaries, *Webster's New World College Dictionary* and *Random House Webster's Unabridged Dictionary*, meaning that this content is more firmly tailored to users of American English. The order in which content from these different sources is displayed is determined by the user's location: if a user accesses the website from the USA, content from American dictionaries is higher on the page, and content from the US edition of COBUILD is displayed rather than from the UK edition.

With these multiple sources available, users can choose the appropriate dictionary for their needs by clicking on the relevant anchor link in a bar displayed at the top of individual entry pages: typically, the choice is between the COBUILD, CED and American definitions. This provision offers a solution to the question of competing international standards that works well in the online context, allowing each entry to match the needs of different types of user more closely. Similarly, vocabulary that takes variant forms can be included in a way that allows users to look up whichever variant they choose and be redirected to the definition by means of hyperlinks, or by automatic redirection: for example, entering *advisor* into the search bar takes the user to the page for *adviser*.

The website also promotes different varieties of English in its inclusion of pronunciation videos spoken by real people with a range of regional accents. The inclusion of these videos helps legitimize varieties of spoken English that deviate from the standard forms represented in the dictionary texts, as well as giving learners a sample of the type of variation in pronunciation that exists within the UK.

Future work on the development of the online dictionary aims to develop these concessions to users' individual language varieties further as part of

a drive to allow them to customize their experience. Users are already presented with information in a specific order depending on their location, as noted above, but it is intended that future development of the website will further exploit the ability to detect users' preferences in a way that enables additional customization. The development of the online dictionary in this direction increasingly promises to enhance what is offered in print format, and may eventually allow for a dictionary concept that reflects the full variety of language in use rather than being orientated towards a single prestigious standard variety.

References

- Atkins, B. T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Aitken, A. J., Allsopp, S. R. R., Bansal, R. K., et al. (1979). The development of English as a world language. In P. Hanks, ed., *Collins Dictionary of the English Language*, 1st ed. Glasgow: Collins, pp. xxii–xxx.
- Beattie, S, ed. (2016). *Collins Easy Learning Irish Dictionary*, 2nd ed. Glasgow: Collins.
- Brookes, I. & O'Neill, M., eds. (2018). *Collins English Dictionary*, 13th ed. Glasgow: Collins.
- Brookes, I., Grandison, A., Hollingworth, L., Hucker, H., Munro, M. & O'Neill, M., eds. (2018). *Collins English School Dictionary*, 6th ed. Glasgow: Collins.
- Comer, N., Uí Dhuifinn, G., Mac Mathúna, S., Nic Mhaoláin, M., Ó Corráin, A. & Ó Mianáin, P., eds. (2019a). *Collins Gem Irish Dictionary*, 5th ed. Glasgow: Collins.
- Comer, N., Uí Dhuifinn, G., Mac Mathúna, S., Nic Mhaoláin, M., Ó Corráin, A. & Ó Mianáin, P., eds. (2019b). *Collins Irish Dictionary: Pocket edition*, 5th ed. Glasgow: Collins.
- Hanks, P. (2010). Compiling a monolingual dictionary for native speakers. *Lexikos*, **20**, 580–98.
- Hidalgo Tenorio, E. (2000). Gender, sex and stereotyping in the Collins COBUILD English language dictionary. *Australian Journal of Linguistics*, **20**(2), 211–30.
- Hill, B. M. & Shaw, A. (2013). The Wikipedia gender gap revisited: characterizing survey response bias with propensity score estimation. *PLoS ONE*, **8**(6), e65782.
- Ishikawa, S. (2011). Duality in the spelling of English verb suffixes -ize and -ise: a corpus-based study. *International Proceedings of Economics Development and Research*, **26**, 390–6.
- Kerswill, P. (2006). Standard English, RP and the standard–non-standard relationship. In D. Britain, ed., *Language in the British Isles*, 2nd ed. Cambridge, UK: Cambridge University Press, pp. 34–51.

- Ó Baoill, D. P. (2010). Irish. In M. J. Ball & N. Müller, eds. *The Celtic Languages*, 2nd ed. London: Routledge, pp. 163–229.
- Ó Dochartaigh, C. (1992). The Irish language. In D. MacAulay, ed., *The Celtic Languages*. Cambridge, UK: Cambridge University Press, pp. 11–99.
- O'Rourke, B. & Walsh, J. (2015). New speakers of Irish: shifting boundaries across time and space. *International Journal of the Sociology of Language*, **231**, 63–83.
- Rottet, K. J. (2014). Neology, competing authenticities, and the lexicography of regional languages: the case of Breton. *Journal of the Dictionary Society of North America*, **35**, 208–47.
- Saito, H. (2015). Vowel shifts of English. *グローバル・コミュニケーション研究*, **2**, 93–102.
- Sinclair, J., ed. (2018). *Collins COBUILD Advanced Learners' Dictionary*, 9th ed. Glasgow: Collins.
- Trudgill, P. (2001). *Sociolinguistic Variation and Change*. Edinburgh: Edinburgh University Press.