ReCALL **29**(2): 138–159. 2017 © European Association for Computer Assisted Language Learning
doi:10.1017/S0958344016000252
First published online 6 February 2017

Evaluating online bilingual dictionaries: The case of popular free English-Polish dictionaries

ROBERT LEW

Adam Mickiewicz University in Poznań, Poland
(email: rlew@amu.edu.pl)

AGNIESZKA SZAROWSKA

Adam Mickiewicz University in Poznań, Poland
(email: agnieszka.szarowska@gmail.com)

Abstract

Language learners today exhibit a strong preference for free online resources. One problem with such resources is that their quality can vary dramatically. Building on related work on monolingual resources for English, we propose an evaluation framework for online bilingual dictionaries, designed to assess lexicographic quality in four major areas: coverage of lexical items, their treatment in the entries, access to lexical information, and presentation of lexicographic data. The proposed framework is applied to a set of six popular bilingual English-Polish dictionaries freely available online, established through an online survey of Polish learners of English.

1 Introduction

Along with rapid technological developments in other fields over the last decades, lexicography has undergone dramatic changes. Most importantly, there has been a clear shift from print to the digital medium. Electronic dictionaries (EDs) are a relatively modern invention, defined in broad terms by Nesi (2000: 839) as:

> any reference material stored in electronic form that gives information about spelling, meaning, or use of words. Thus, a spell checker in a word-processing program, a device that scans and translates printed words, a glossary for on-line teaching materials, or an electronic version of a respected hard-copy dictionary are all EDs of a sort, characterised by the same system of storage and retrieval.

The best-known classification of digital dictionaries is that given in De Schryver (2003), but it pre-dates the current explosion of online dictionaries, which may be characterized as dictionaries that cannot be accessed without an active internet connection (Müller-Spitzer, Koplenig & Töpel, 2012: 430).

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

Dictionaries are an important resource for foreign language learners. With the ongoing digital revolution in lexicography, language learners have largely shifted to digital dictionaries, primarily online (L'Homme & Cormier, 2014; Lew & De Schryver, 2014). According to a recent large-scale survey, digital dictionaries are the single most important resource for language learners (Levy & Steel, 2015).

Online dictionaries come in different shapes and sizes, and many are of dubious quality. Unlike authoritative monolingual dictionaries for learners of English, for which an ad-supported business model is viable given the scale effect of millions of users worldwide, quality bilingual dictionaries for a combination of English with a lesser-used language can rarely afford to make the contents available freely online. As many users of online dictionaries are not willing to pay for their content, they are stuck with the free offerings. However, the average language learner is unable to competently assess the quality of the available lexical resources, many of which are poor lexicographically (Lew, 2014).

Despite the enormous popularity of online dictionaries, there have been very few attempts to evaluate them systematically (Lew, 2011; Meyer & Gurevych, 2012; Pearsons & Nichols, 2013; Yamada, 2010). The few studies that are available seem to be concerned with monolingual English dictionaries, completely disregarding bilingual dictionaries, even though the latter are known to be preferred by learners (Lew & Adamska-Sałaciak, 2015). Lew (2011) provides an overview of selected monolingual online English dictionaries, categorized and discussed in terms of the major lexicographic issues. Meyer and Gurevych (2012) offer a detailed description of Wiktionary as a popular collaborative dictionary. Yamada (2013) is an evaluation of online dictionaries of English using the framework elaborated in Pearsons and Nichols (2013). Yet, as far as we are aware, no comparable analysis has ever been attempted for online bilingual dictionaries.

This study presents a proposal for an evaluation framework for online bilingual dictionaries. The framework can be applied directly in the evaluation of dictionaries with English as the source language (i.e. going from English to another language); for dictionaries with another source language, a different set of test items would have to be used, specific to that language, though it can be created on analogous principles. The proposed framework is then applied to those English-Polish online dictionaries that are most popular among Polish learners of English. In order to establish a list of the dictionaries that learners use most frequently, an online survey was conducted involving Polish learners of English at different levels of language proficiency.

Any evaluation needs criteria. We take as a starting point an existing evaluation framework for online monolingual English dictionaries (Pearsons & Nichols, 2013) and construct a framework for online bilingual dictionaries with English as a source language. The framework addresses four fundamental lexicographic areas: coverage of lexical items, their treatment in the entries, access to lexical information, and presentation of lexicographic data. The first two areas are medium-independent: they apply equally to print and digital dictionaries. By contrast, access and presentation are areas that operate differently for print and digital dictionaries, and have benefited the most from the digital revolution (Lew & De Schryver, 2014). The criteria for these two areas will be developed on the basis of recent findings on the effectiveness and user-friendliness of various features of digital dictionaries.

In the immediately following section, we report on an online survey employed to establish the most popular English-Polish online dictionaries. The top six of these are subjected to a systematic expert evaluation. The framework for this evaluation is given in Section 4 (evaluation framework), followed by the presentation of evaluation results and lexicographic recommendations.

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

2 Selection of dictionaries for evaluation

2.1 Method

In order to establish which English-Polish online dictionaries were the most popular among Polish learners of English, a questionnaire was designed on the online survey platform *SurveyGizmo*. An online format was selected for two reasons. First, it enabled us to use advanced features unavailable in pencil-and-paper format, such as drag-and-drop ranking, question piping and looping. These features helped improve the reliability and validity of the survey. The second reason was that an online survey was the natural choice when trying to reach an audience of *online* dictionary users. As the survey was intended for Polish learners of English, it was drafted in Polish. The questionnaire was piloted on a group of ten volunteers, and, after some improvements, made available to the participants (see Section 2.2). It went live between 2 and 16 March, 2014.

The questionnaire probed several aspects of online dictionary use. The two items that are most relevant in the present context concern the online dictionaries most used by participants, and their ranking of quality of these dictionaries. In addition, the questionnaire collected the following information about every single dictionary indicated by a given participant but not reported here: the purposes for which participants used the dictionary, context of use, format and type of device used for accessing them, and their level of proficiency in English.

A key item in the questionnaire inquired about online dictionaries most often used by the participants. They were presented with a clickable list of eleven dictionaries which were known to the researchers to be popular. This list had been constructed based on information gathered on a number of internet forums with substantial threads (at least ten responses) discussing recommended English-Polish online dictionaries, giving: *Diki, bab.la, MEGA-słownik, DICT, Ling.pl, Getionary, Wiktionary, PONS, Angool.com, WordReference*, and *leksyka.pl*. The above were listed explicitly within this questionnaire item.

Apart from this initial list of eleven dictionaries, the item had three extra boxes for participants to list up to three other dictionaries they used often that were not on the initial list. A participant could select as many dictionaries as they wished, combining the ones from the list with the ones supplied by the participants.

In the next step, the survey system presented a drag-and-drop ranking task with all (and only) the dictionaries ticked or added in the previous item by a given participant. Participants were here instructed to drag the dictionaries across to an empty box, placing the best at the top, and the worst at the bottom. This item allowed dictionaries to be rearranged at will until the participant was satisfied with their choice and confirmed it. There was no obligation to include every single one of the dictionaries they had selected or supplied in the ranking.

2.2 Participants

Participants were recruited online via snowball sampling. A group of initial respondents were asked to nominate, through their social networks (Facebook, email), other participants who met the eligibility criteria: participants had to be Polish learners of English, regardless of their language proficiency, aged between 18 and 25. Most of them were students of different specialisms who studied English either as a major or minor subject.

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

The online survey invitation included a clarification of objectives, explaining that participation was voluntary and it was possible to withdraw at any time. Participants were informed that by returning the questionnaires they were consenting to participation in the study and their responses being analysed and published, and that the survey would be anonymous. They provided written consent online by ticking a relevant box. The survey was fully anonymous and no sensitive personal data were collected (except possibly their IP addresses, HTTP Referrer, and HTTP client data, which were subsequently removed from the data file). Overall, there were 205 participants. Of these, 187 responses were used for the analysis.

The breakdown of the participants' level of proficiency (self-declared in the online questionnaire) was as follows, based on the 180 responses returned for this item (not all participants provided a response on this optional item). Half of the participants were advanced or proficient, with declared CEFR levels C1 (48) or C2 (43). About 40% represented (upper-)intermediate proficiency (B1: 29; B2: 42). Only 10% were beginners (A2: 10; A1: 8). This should not be surprising, as most Polish learners within this age bracket now represent relatively high levels of proficiency in English.

3 Results

3.1 Dictionaries most frequently used

Table 1 presents the distribution of the participants' choices as to which dictionaries they used most often. A dozen others (not in the table) received single nominations: *Collins*, *Cambridge*, *linguee*, *infobot*, *glosbe*, *mtranslator*, *medland*, *translator*, *Oxford*, *Wikipedia*, *tłumacz.onet*, *thefreedictionary*.

Table 1 *Dictionaries most frequently used (multiple answers were possible)*

dictionary	count	percent
MEGAsłownik	82	46%
Diki	65	37%
bab.la	54	31%
Ling.pl	47	27%
PONS	36	20%
Wiktionary	29	16%
WordReference	25	14%
Google translate	24	14%
DICT	17	9.6%
Getionary	7	4.0%
translatica	6	3.4%
Angool.com	4	2.3%
leksyka.pl	4	2.3%

It may be noted that not all of the frequently named resources had been supplied in the initial list of eleven dictionaries. Participants reported frequent use of machine-translation services, notably Google translate, with as many as 24 participants adding it under the *other* category. No doubt, the use of Google translate as a language resource is much higher; it can

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

hardly be classified as an online dictionary, though, so it was not included in the initial list (and likely quite a few other participants did not report its use for this very reason). Other translation services also received single mentions, as did some monolingual dictionaries.

Based on these results, the seven most popular dictionaries were preselected for inclusion in the evaluation.

3.2 Dictionaries most highly ranked

Participants were also asked to rank the dictionaries they named (or a subset thereof) in terms of quality. We had expected the results to largely overlap with the dictionaries most frequently used (as it is rational to tend to use the tools you value highly). This expectation was confirmed, as the set of top seven dictionaries turned out to be the same in both cases. However, there was one notable difference: *MEGAsłownik*, which was by far the most popular dictionary among the participants, was not ranked as the best, but the second best. The number one dictionary according to the participants' evaluation was *Diki*, the second most popular. The ranking of the remaining dictionaries, however, corresponded to their declared frequency of use, i.e. the dictionaries reported to be used more often were ranked more highly.

3.3 Final list of dictionaries selected for evaluation

MEGAsłownik unexpectedly closed down shortly after the survey was conducted. Accordingly, the list of the dictionaries for evaluation was reduced to the remaining six: *Diki* (http://www.diki.pl), *bab.la* (http://bab.la), *Ling.pl* (http://ling.pl),[1] *PONS* (http://pl.pons.com/t%C5%82umaczenie?l=enpl), *Wiktionary* (http://pl.wiktionary.org), and *WordReference* (http://www.wordreference.com/enpl/).

4 Evaluation framework

4.1 Criteria for coverage and treatment

The criteria selected for the evaluation of coverage of lexical items and lexicographic treatment were partially based on the framework proposed in Pearsons and Nichols (2013). As in the original framework, each criterion was marked on an integer scale of 1 to 5, with verbal labels for these being as follows: 1: Unsatisfactory; 2: Partly acceptable; 3: Mostly satisfactory; 4: Satisfactory; 5: Very satisfactory. However, as most of the original criteria were somewhat vague, in an effort to make the evaluation more objective, for most of the original items, we added four additional, more detailed sub-criteria, each worth a single point. Thus, if a dictionary met none of the four sub-criteria, it received the lowest mark of 1 for a given main criterion; when a single detailed criterion was met, a score of 2 was awarded; two criteria earned a score of 3; three criteria, 4; and all criteria, 5.

For the coverage and presence of labels criteria, the proportion of features covered in a given dictionary was computed and converted to scores as follows: a mark of 5 was given to 80–100% coverage or presence of labels; 4 to 60–79%; 3 to 40–59%; 2 to 20–39%;

[1] Ling.pl is a dictionary *aggregator*: it combines content from several independent source dictionaries.

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

1 to 0–19%. The items used in evaluating coverage and labels are given in Sections 4.1.1 and 4.1.2 below.

The list of areas, main criteria, and sub-criteria used in the evaluation of coverage and treatment was as follows:

I Coverage

1. Coverage of entries (whether a certain item was present in a dictionary)
    a) neologisms
    b) specialized information technology (IT) vocabulary
    c) multi-word expressions with frequent verbs

II Treatment

2. Presence of labels for
    a) usage level (e.g. formal, informal, technical, literary, slang)
    b) regional variety (British/American)
    c) part of speech
3. Hyperlinked cross-references from
    a) equivalents
    b) examples
    c) synonyms
    d) other
4. Pronunciation indication (audio, transcription)
5. Provision of example sentences
6. Additional information on
    a) usage
    b) synonyms
    c) antonyms
    d) related words and phrases
7. Grammar and collocation
    a) irregular forms of verbs
    b) noun countability
    c) comparative and superlative forms of irregular adjectives
    d) collocations

4.1.1 Items used in the evaluation of coverage. Thirty vocabulary items were used to assess the coverage of entries. These items were selected from within three categories: neologisms, specialized vocabulary in the field of IT, and multi-word expressions containing frequent verbs.

The rationale behind using neologisms was to have a measure of how up-to-date the headword lists were. Even if some of them would not survive, their inclusion was thought to be a valid measure of an effective inclusion policy for an online dictionary. The ten items were selected randomly from among neologisms added to Oxford Dictionaries Online in 2013: *babymoon, bankster, bitcoin, buzzworthy, dappy, fauxhawk, selfie, twerk, phablet* and *omnishambles*.

The specialized IT items were selected at random from the glossaries on information technology and computer science on the website http://whatis.techtarget.com: *adware, cache, firewall, spreadsheet, router, crowdsourcing, failover, scareware, freemium*, and *interface*.

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

The ten test multi-word expressions were selected from *Longman Dictionary of Contemporary English Online*, *Oxford Dictionaries Online*, and *Macmillan English Dictionary Online*. The idea was to have less frequent expressions containing very frequent verbs. The items selected were: *to take exception to, to make a beeline for, to pass the time of day with sb, to put sth behind you, to pull rank on sb, to go all out to do sth, to bring sb up short, to set the seal on sth, to go by the board* and *to try one's hand at*. Each of the 30 items was looked up manually in the six dictionaries and its presence or absence noted.

4.1.2 Items used in the evaluation of labelling. The ten items for assessing the inclusion of appropriate usage labelling in the dictionaries were adopted from an instructional section in *Oxford Learners Dictionaries Online*. In this section, the use of labels in the dictionaries was explained with representative examples and these examples were adopted for the present analysis. They included: two formal words (*admonish, besmirch*); two informal words (*bonkers, dodgy*); two technical items (*accretion, adipose*); two literary words (*aflame, halcyon*); and two slang words (*dingbat, dosh*).

The ten items for assessing the presence of regional variety labels were adapted from Leaney (2007), and they consisted of five British English words in specific senses: *flat, mark, (bank)note, pavement, petrol*, and their American English counterparts: *apartment, grade, bill, sidewalk, gas*.

The ten items that were used to assess the presence of part of speech labels were also adapted from Leaney (2007): *allow, accent, commit, contract, damage, export, flower, forgive, import, level*. Inflected verb forms of these items were also used in assessing the retrieval of inflected forms. Each of the thirty items was looked up manually in the six dictionaries and its presence or absence noted.

4.1.3 Items for the remaining treatment criteria. The assessment of the presence of the remaining elements in the dictionaries did not call for dedicated item lists. Therefore, for this purpose a joint list of ten words was drawn up, based on the frequency lists supplied with the *Corpus of Contemporary American English* (Davies, 2008–). The most frequent words from the four major syntactic classes were selected as follows:

- verbs: *be, have, do*
- nouns: *time, year, people*
- adjectives: *other, new*
- adverbs: *up, so*

In addition, the following irregular adjectives were used to check for inclusion of their irregular comparative and superlative forms: *good*, *bad*, *many*.

4.2 Criteria for access and presentation

The criteria for the evaluation of access were largely based on Lew (2012) and they included:

III Access
   8. Headword identification

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

a) accessing inflected forms (reducing inflected forms to the lemma)
b) accessing misspelled words (the "did you mean" function)

9. Accessing multi-word units (separate headwords for multi-word expressions)
10. Type-ahead search
11. Entry navigation devices (entry menus or signposts helping users in selecting the relevant sense)

The criteria for the evaluation of presentation were derived from De Schryver (2003), Dziemianko (2011, 2012), Lew (2012), Lew, Grzelak and Leszkowicz (2013), Nielsen (2008), and Pearsons and Nichols (2013), and they consisted of the following items:

IV Presentation

12. Presence of pictorial illustrations, which have been found to be helpful in dictionary consultation (Lew, 2012)
13. Consistent entry form (i.e. inclusion of the same microstructural elements in consistent order and presentation), which has been claimed to make dictionary look-up process easier (Nielsen, 2008: 177)
14. Full names of grammatical codes and symbols given in the user's mother tongue (De Schryver, 2003)
15. Use of bold type other than in the headword or equivalent (Lew, Grzelak & Leszkowicz, 2013)
16. No intrusive advertisements, which were found to impede dictionary consultation (Dziemianko, 2011, 2012); this criterion also featured in Pearsons and Nichols (2013)

In addition, customizability of the interface (De Schryver, 2003) was also considered as a potential criterion, but, as it turned out, none of the dictionaries under evaluation offered such an option, therefore this criterion was dropped. For the same reason, we did not include another potential criterion: the use of colours for distinguishing various elements of the microstructure, as advocated in Dziemianko (2015).

In evaluating access and presentation, to test whether specific elements were present in the dictionaries, generally the same test items were used as in the evaluation of treatment, but with a few variations. In testing the two criteria related to headword identification, in the inflected form search test, inflected forms were used rather than lemma forms; for example, for the verb *do*, the inflected forms *did* and *done* were searched for. In testing for the presence of the "did you mean" function (Lew & Mitton, 2011, 2013), deliberate single-letter misspellings were used, thus *haeve* was entered to test for *have*, etc. In testing whether multi-word expressions could be accessed directly, two well-known idioms were used whose presence in all six dictionaries had previously been verified: *a piece of cake* and *kick the bucket*. Finally, to test for the presence of pictures, frequent concrete nouns were used (*cat, dog, chair*), as such entries were most likely to feature pictures (Stein, 1991) if they were at all offered in the dictionary.

The scoring system applied in the evaluation of access and presentation was similar to that used for coverage and treatment, with a scale ranging from 1 (unsatisfactory) to 5 (satisfactory). One score was given for access and another one for presentation, reflecting the number of features present in a given category: if all five features tested were present in a particular category, a score of 5 was awarded; four features earned a dictionary

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

a score of 4, etc. If only one or no features were found in a dictionary, a score of 1 (unsatisfactory) was entered.

In general, when a given feature was not fully consistent across the test items, the mark reflected the situation for the majority of the items. For example, if at least six out of ten items checked included collocational information in a dictionary, a point was awarded for this feature. These points were awarded even if the quality of the lexicographic information was sometimes in doubt, but such cases are noted.

4.3 Evaluation procedure

The evaluation of the six online English-Polish dictionaries was conducted manually, independently by two experts on lexicography (the two co-authors), with all marks entered in parallel Excel workbooks. On completion, the two datasets were compared. All detailed marks by the two experts turned out to be in complete agreement (an interrater agreement of 100%). The results of the evaluation will be presented below. For data analysis, we used Microsoft Excel 2013 and R version 3.2.3 (R Core Team, 2015).

5 Evaluation results and discussion

5.1 Coverage

The coverage of entries from the three vocabulary categories was examined. The results are presented in the tables below. A +/– marking system was applied, where "–" means that a particular entry was not present in a dictionary and " + " means that an entry was present in a dictionary.

The coverage of neologisms was not comprehensive in any of the dictionaries (Table 2). Four out of ten entries were covered in Diki and Wiktionary, three in bab.la, two in PONS and none at all in Ling.pl and WordReference.

Table 2 *Coverage of neologisms*

item	Diki	bab.la	Ling.pl	PONS	Wiktionary	WordReference
babymoon	–	–	–	–	–	–
bankster	–	–	–	–	+	–
bitcoin	–	–	–	–	+	–
buzzworthy	–	–	–	–	–	–
dappy	–	–	–	–	–	–
fauxhawk	–	–	–	–	–	–
selfie	+	+	–	+	–	–
twerk	+	+	–	–	–	–
phablet	+	–	–	+	+	–
omnishambles	+	+	–	–	+	–
count	4	3	0	2	4	0

Coverage of IT terms was rather better (Table 3). Nine out of ten technical vocabulary items related to information technology were covered in Diki, eight in bab.la, and seven in Wiktionary. Five items were found in WordReference, four in Ling.pl, and three in PONS.

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

Table 3 *Coverage of technical vocabulary*

item	Diki	bab.la	Ling.pl	PONS	Wiktionary	WordReference
adware	+	+	–	–	+	–
cache	+	+	–	+	+	+
firewall	+	+	+	–	+	+
spreadsheet	+	+	+	+	+	+
router	+	+	+	–	+	+
crowdsourcing	+	+	–	–	–	–
failover	+	–	–	–	–	–
scareware	+	+	–	–	+	–
freemium	–	–	–	–	–	–
interface	+	+	+	+	+	+
count	9	8	4	3	7	5

Diki was unbeatable in its coverage of multi-word expressions (Table 4), scoring ten out of ten. Six phrases were found in Ling.pl, five in PONS, while bab.la and WordReference scored four. Wiktionary turned out to be the worst, as not a single one of the multi-word expressions tested was found there.

Table 4 *Coverage of multi-word expressions*

item	Diki	bab.la	Ling.pl	PONS	Wiktionary	WordReference
to take exception to	+	+	+	+	–	+
to make a beeline for	+	+	+	+	–	+
to pass the time of day with sb	+	–	+	–	–	–
to put sth behind you	+	–	–	–	–	–
to pull rank on sb	+	–	+	–	–	–
to go all out to do sth	+	+	+	–	–	+
to bring sb up short	+	–	–	–	–	–
to set the seal on sth	+	–	–	+	–	–
to go by the board	+	–	+	+	–	–
to try one's hand at	+	+	–	+	–	+
count	10	4	6	5	0	4

5.2 Treatment

5.2.1 Labels. The inclusion of labels was checked for three types of labels: usage level, regional variety, and part of speech. We used the same scoring system as in the case of the coverage of entries (Section 5.1).

When it came to checking for usage labels (Table 5), labels in Polish were accepted (they are in fact preferable to the Polish user), as were variations in labelling, such as in marking levels of usage (*informal* versus *slang*, which is not always an easy call), or more domain-specific labels (e.g. *astronomy* rather than *technical*). Diki, bab.la, and Ling.pl got about half of the usage labels right, PONS and WordReference three out of ten, while Wiktionary did not score any marks, although the majority of the testing items were simply missing from the dictionary.

Table 5 *Presence of usage labels (n/f = entry missing)*

item	Diki	bab.la	Ling.pl	PONS	Wiktionary	WordReference
admonish (formal)	+	–	+	–	–	–
besmirch (formal)	–	+	–	+	–	n/f
bonkers (informal)	–	–	+	n/f	n/f	+
dodgy (informal)	+	–	+	+	n/f	–
accretion (technical)	+	+	–	+	n/f	–
adipose (technical)	–	–	n/f	–	n/f	–
aflame (literary)	–	+	+	–	n/f	–
halcyon (literary)	–	+	+	n/f	n/f	–
dingbat (slang)	+	+	–	n/f	n/f	+
dosh (slang)	+	+	n/f	n/f	n/f	+
count	5	6	5	3	0	3

When it comes to regional variety labels (Table 6), Diki and PONS did very well, with 10 and 9 regional labels present, respectively. The other dictionaries scored between 3 and 5; Wiktionary did the worst.

Table 6 *Presence of regional variety labels*

item	Diki	bab.la	Ling.pl	PONS	Wiktionary	WordReference
flat (BrEng)	+	–	–	+	–	+
apartment (AmEng)	+	–	+	+	–	+
mark (BrEng)	+	–	–	–	–	–
grade (AmEng)	+	+	+	+	–	–
(bank)note (BrEng)	+	–	–	+	–	–
bill (AmEng)	+	–	+	+	–	–
pavement (BrEng)	+	+	–	+	+	+
sidewalk (AmEng)	+	+	+	+	+	–
petrol (BrEng)	+	–	–	+	–	+
gas (AmEng)	+	+	+	+	+	–
count	10	4	5	9	3	4

Part of speech labels were present for all ten test items in all six dictionaries. Thus, all dictionaries received perfect marks on this sub-criterion. We omit a detailed table for this.

5.2.2 Hyperlinked cross-references. Inclusion of hyperlinked cross-references in the six dictionaries is summarized in Table 7. Diki included hyperlinked cross-references for related words, idioms and phrases, but not for individual words in senses and examples. Phonetic transcription of each word was hyperlinked to instructions on the International Phonetic Alphabet. Similarly, syntactic labels were hyperlinked to their extensions with explanation. There were also links to verb conjugation patterns.

In bab.la, all Polish equivalents were hyperlinked to their entries in the Polish-English section of the dictionary. Within each entry there was a list of hyperlinked synonyms and

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

Table 7 *Presence of hyperlinked cross-references*

cross-references for	Diki	bab.la	Ling.pl	PONS	Wiktionary	WordReference
equivalents	–	+	–	+	+	+
examples	–	–	–	+	+	+
synonyms/related words and phrases	+	+	–	+	+	+
other	+	+	–	–	+	+
score	2	3	0	3	4	4

related phrases, and the dictionary also supplied links to verb conjugation patterns. In Ling.pl, no cross-references could be found. The only words that were hyperlinked served as advertisements, leading to non-lexical commercial websites such as online shops. In PONS, equivalents, related phrases, as well as all words in examples were hyperlinked. In Wiktionary, each equivalent was hyperlinked, as well as synonyms, antonyms, related phrases and individual words in example sentences. There were also links to expansions of syntactic labels. WordReference included hyperlinks for every single word as well: equivalents, synonyms, related words and phrases, individual words in example sentences, and explanations of abbreviations used in the dictionary.

5.2.3 Indication of pronunciation. A summary of the extent of pronunciation indication in the six dictionaries is given in Table 8. Diki offered both transcription and audio recordings, though only for British English. In bab.la, no transcription was provided, but audio recordings were present for both English and Polish words. There was no clear indication as to which variety of English was presented. In Ling.pl, there were audio recordings (without an indication of the variety, but the recordings sounded like American English), and transcription for both British and American English, yet the location of the transcriptions was inconsistent, given that Ling.pl aggregated content from a number of individual dictionaries. PONS offered both transcriptions and audio recordings for Polish and English, including audio recordings for phrases, with a choice of British or American versions. In Wiktionary, there were both transcriptions and audio recordings. The inclusion of varietal forms was inconsistent across entries: some provided only the American English variety, while others had British, American and even Australian English. Finally, WordReference provided both transcription and audio recordings for both American and British English.

Table 8 *Indication of pronunciation*

criterion	Diki	bab.la	Ling.pl	PONS	Wiktionary	WordReference
transcription						
one variety	+	–	+	+	+	+
more varieties	–	–	+	+	+	+
audio recording						
one variety	+	+	+	+	+	+
more varieties	–	–	–	+	+	+
score	2	1	3	4	4	4

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

5.2.4 Example sentences. Example sentences (Table 9) were present in all six dictionaries, though the details of presentation varied. In Diki, each main sense of an entry had between one and three example sentences, translated in brackets and with audio representation available at a click. In addition, most test entries had an additional substantial set of over twenty examples available on request. These extra examples seemed to be drawn automatically from a corpus, and so not all were very well-suited to the entry. A translation option was offered for these examples, though these were clearly machine translations, of low accuracy and questionable benefit to language learners.

Table 9 *Example sentences*

criterion	Diki	bab.la	Ling.pl	PONS	Wiktionary	WordReference
5+ examples per entry	+	+	–	–	+	+
Polish translation	+	+	+	+	+	+
sources of examples	–	+	–	–	–	–
full-sentence examples	+	+	–	–	+	+
score	3	4	1	1	3	3

In bab.la, each sense was accompanied by a single example sentence, but when this sentence was clicked on, a further four sentences appeared. All the example sentences were translated and their Internet sources were identified.

The presentation of examples in Ling.pl again suffered due to the aggregator status of this resource, being highly inconsistent. The presence and extent of exemplifying material depended on which specific dictionary happened to be drawn on at a particular point. When examples did appear, they were usually translated. In PONS, there were only a few examples under some of the entries. Most were short phrases, with very few full-sentence examples. They were all translated into Polish. Wiktionary included at least one example per sense (thus over five senses in our test polysemous entries). All were translated into Polish, but no source was identified. WordReference, like Wiktionary, typically supplied a single example sentence for each sense, with no source given. Most of the examples were accompanied by Polish translations, except when the examples illustrated multi-word expressions.

5.2.5 Additional information. Additional lexical information evaluated was as follows: usage notes, antonyms, synonyms, and related words and phrases (note that it is the presence of these information categories that is at issue here, while in Section 5.2.2 above it is the presence of the hyperlink that matters.) A summary of the results is given in Table 10.

In Diki, most entries included a list of hyperlinked related words and phrases in a sidebar. Some senses were also accompanied by hyperlinked synonyms and antonyms below the Polish equivalent, but no notes on usage were available. In bab.la, each entry was accompanied by synonyms and related phrases, but there were no antonyms or usage notes. In addition, forum posts with additional information were available. Ling.pl did not list any additional information. A minority of entries (depending on which specific dictionary they came from) included related phrases within an entry, but they were not specifically marked.

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

Table 10 *Provision of additional information*

additional information	Diki	bab.la	Ling.pl	PONS	Wiktionary	WordReference
usage notes	–	–	–	–	+	+
antonyms	+	–	–	–	+	–
synonyms	+	+	–	+	+	+
related words and phrases	+	+	–	+	+	+
score	3	2	0	2	4	3

In PONS, synonyms were included as sense indicators and related phrases were listed. Antonyms and usage notes were not present. Wiktionary was quite rich in additional information. Synonyms, antonyms, related phrases and usage notes were all present for numerous entries. Wiktionary was the only dictionary to score a point on each of the four sub-criteria. In WordReference, related phrases were present and some entries had accompanying usage notes. Synonyms could be called up by clicking on “Angielskie synonimy” (‘English synonyms’) below the entry word, which directed the user to an English thesaurus for a particular word. Antonyms were not available in this dictionary.

5.2.6 Grammar and collocation. A summary of the results for information on grammar and collocation is given in Table 11. Grammatical information in Diki covered irregular verb forms, noun countability, and comparative and superlative forms of irregular adjectives. Some entries supplied collocational information (in collapsed view by default but available on request), which consisted of large sets of collocations classified by part of speech. These sets appeared to be automatically extracted (so they were not always fully accurate), and supplied with machine translations into Polish. In bab.la, only irregular forms of verbs were found. No information on collocation, comparative and superlative forms of irregular adjectives, or noun countability was given. In Ling.pl, irregular forms of verbs, noun countability, and comparative and superlative forms of irregular adjectives were all present, but collocational information was not available. PONS covered all the features under evaluation, i.e. irregular forms of verbs, noun countability, comparative and superlative forms of irregular adjectives as well as collocation. In Wiktionary, grammatical information was quite exhaustive, with irregular forms, noun countability, comparative and superlative forms all given, along with collocational information.

Table 11 *Provision of grammatical and collocational information*

feature	Diki	bab.la	Ling.pl	PONS	Wiktionary	WordReference
irregular forms of verbs	+	+	+	+	+	+
noun countability	+	–	+	+	+	–
comparative and superlative forms of irregular adjectives	+	–	+	+	+	–
collocations	+	–	–	+	+	+
score	4	1	3	4	4	2

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

5.3 Summary of the results for coverage and treatment

The areas of coverage and treatment both relate to the content of dictionaries, as opposed to what is done with that content. Therefore, we present here a summary of the evaluation in these two areas. Table 12 lists the scores for all coverage and treatment criteria (see Section 4.1 for details on how the scores were awarded). The results indicate that the best overall score was received by Diki: 4.0 out of 5. Wiktionary scored 3.5, PONS 3.4, WordReference 3.3, bab.la 3.1, and Ling.pl ended up with the lowest score of 2.6. Of course, we should not be forgetting that these are just the most used and most liked dictionaries: an elite, as it were, of the freely available English-Polish online bilingual dictionaries.

Table 12 *Scores for coverage and treatment*

criterion	Diki	bab.la	Ling.pl	PONS	Wiktionary	WordReference
coverage of neologisms	2	2	1	1	2	1
coverage of technical vocabulary	5	4	2	2	4	3
coverage of multi-word expressions	5	2	3	3	1	2
presence of usage-level labels	3	3	3	2	1	2
presence of variety labels	5	2	3	5	2	2
presence of part of speech labels	5	5	5	5	5	5
presence of cross-references	3	4	1	4	5	5
presence and form of pronunciation	3	2	4	5	5	5
example sentences	4	5	2	2	4	4
additional information	4	3	1	3	5	4
grammar and collocation	5	2	4	5	5	3
overall mean score	4.0	3.1	2.6	3.4	3.5	3.3

The evaluation so far has focused on the traditional lexicographic criteria of coverage and treatment. Let us now proceed to two further aspects that can take advantage of the specific possibilities afforded by the digital medium: access and presentation.

5.4 Access

Access was evaluated with the use of the five detailed criteria listed in Section 4.2; these are repeated in the first column of Table 13, which gives a summary of the results for access. In Diki, inflected forms did lead to their lemmas correctly, though the dictionary did not immediately redirect the user to the full entry. Diki featured a "did you mean" function, displaying a list of suggestions for misspelled searches. Multi-word units were accessible directly and type-ahead search was present. No entry navigation devices were available.

Bab.la recognized inflected forms correctly, with automatic redirection to their parent lemmas. Misspelled words were also identified with a list of suggestions. Multi-word expressions could be accessed directly, and an incremental search function was available. Bab.la did not contain entry navigation devices.

Ling.pl did not recognize inflected forms. This dictionary resembled a print dictionary in that there were traditional cross-references to citation forms (e.g. "gone → zob. go") that were not hyperlinked, so a user had to again type "go" in the search box. Ling.pl did not

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

Table 13 *Access*

criterion	Diki	bab.la	Ling.pl	PONS	Wiktionary	WordReference
accessing inflected forms	+	+	–	+	–	+
accessing misspelled words	+	+	–	+	–	+
accessing multi-word units	+	+	+	+	+	+
type-ahead search	+	+	+	+	+	+
entry navigation devices	–	–	–	+	–	+
score	4	4	2	5	2	5

feature a “did you mean” function, so misspelled words were not found at all. No entry navigation devices were featured. However, a type-ahead search function was available, as were separate entries for multi-word expressions.

In PONS, most inflected forms were immediately displayed as their lemmas, yet, in the case of irregular verbs such as *gone*, there was a note that this was the past participle of the citation form (*go*), but without a hyperlink. Still, a hyperlink to the lemma could be found in the “see also” box, located below the grammatical information section. The dictionary was able to identify misspellings as well as direct multi-word expression searches, and featured an incremental search function. In addition, PONS included entry navigation devices in the form of synonyms or short glosses in brackets next to each sense, thus scoring a maximum of 5.

Wiktionary was highly inconsistent when it comes to identifying inflected forms. It covered mostly inflected forms of irregular verbs and forms representing lemmata of more than one syntactic class, but it usually failed on regular inflected forms. The dictionary did not accommodate misspellings, and no entry navigation devices were present. On the positive side, multi-word expressions could be accessed directly and there was a type-ahead search function.

WordReference recognized all inflected forms and automatically redirected the user to the relevant entries. It accommodated misspellings, multi-word expression searches, and had type-ahead search capability as well as sense signposts as entry navigation devices. With a score of 5, WordReference tied with PONS in respect of access features.

5.5 Presentation

A summary of the results for the presentation aspect of the evaluation is given in Table 14. Diki included pictorial illustrations. Presentation was consistent across entries. Grammatical codes and symbols were given in full and in Polish. Bold type was used for translation equivalents, and there were no distracting advertisements.

Bab.la did not feature pictorial illustrations. Entry form was consistent. Abbreviations were given in Polish, but their expansions were not available. Bold type was used to highlight tokens of headwords within examples. In bab.la there were quite intrusive ads.

In Ling.pl, no pictures were present. The resource could not be credited with entry consistency, as each “entry” brought together original entries from a number of dictionaries, and in each case the number and order of entries presented was different. Abbreviations went unexplained and were mostly given in English. A big banner ad was displayed at the top of the screen. On a positive note, bold type was used for some microstructural elements, although this again was inconsistent due to the aggregator character of the dictionary.

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

Table 14 *Presentation*

criterion	Diki	bab.la	Ling.pl	PONS	Wiktionary	Word Reference
pictures	+	–	–	–	+	–
consistent entry form	+	+	–	+	+	+
full names of grammatical codes and symbols in L1	+	–	–	–	+	–
use of bold type	+	+	+	+	+	–
no intrusive advertisements	+	–	–	–	+	–
score	5	2	1	2	5	1

PONS did not include any pictures, either. Abbreviations were in Polish, but their expansions were not given. Entry form was consistent and there were numerous bold elements in this dictionary. The page included distracting advertisements.

Wiktionary contained pictures in certain cases and each entry looked the same. There were full names of grammatical codes and symbols and they were provided in the users' mother tongue. There were elements in bold and there were no adverts.

WordReference did not include any pictures or provide full names for abbreviations, which were in English. Entry form was consistent, but there were no bold elements other than the entry word, and there were distracting adverts.

5.6 *Overall scores*

Table 15 brings together evaluation scores in all four areas for the six dictionaries, with Figure 1 presenting these results visually. Overall scores reflecting all four areas evaluated are given in the bottom row of the table, assuming equal weight given to each of the four areas. The dictionary that has emerged ahead of the competition is Diki, with an overall score of 4.3. The scores of the next four dictionaries, Wiktionary, PONS, bab.la, and WordReference are quite close to each other (3.3, 3.2, 3.0, and 2.9, respectively). Ling.pl has received the lowest score of the six at 2.0. It has to be stressed that this should not be taken to mean that Ling.pl is an awful dictionary, as the dictionaries selected for evaluation are the top six from a larger set.

Access and presentation are two areas dependent on the digital medium. Still, the distribution of evaluation scores in these two areas turns out to be fairly similar to that in the

Table 15 *Summary of the scores for the four areas evaluated*

area	Diki	bab.la	Ling.pl	PONS	Wiktionary	WordReference
coverage	4.0	2.7	2.0	2.0	2.3	2.0
treatment	4.0	3.3	2.9	3.9	4.0	3.8
access	4.0	4.0	2.0	5.0	2.0	5.0
presentation	5.0	2.0	1.0	2.0	5.0	1.0
overall score	4.3	3.0	2.0	3.2	3.3	2.9

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

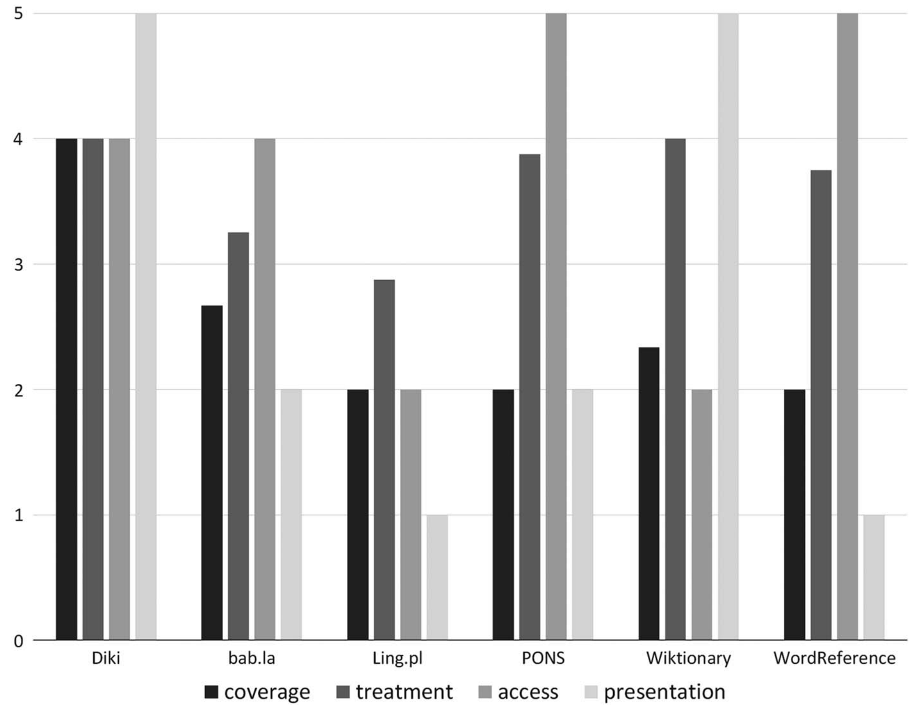

Fig. 1. Summary of the scores for the four areas evaluated

more traditional lexicographic aspects of coverage and treatment. The two dictionaries with perfect scores for presentation are Diki and Wiktionary, whereas access received maximum possible scores in PONS and WordReference.

5.7 Evaluation-based ranks versus user survey ranks

Taking into account overall scores, the final ranking of the dictionaries is as given in the *evaluation rank* column of Table 16. These ranks may be compared with those emerging from the online questionnaire reports by dictionary users so as to assess the degree of overlap between expert evaluation and users' evaluation ranks.

The correlation between evaluation ranks and survey-derived ranks turns out to be low (Spearman's $\rho = 0.31$, $p = 0.54$, n.s.), suggesting that the choices made by participants tend to be motivated by factors other than lexicographic quality. Nevertheless, the top dictionary in the survey overlaps with the best dictionary in the expert evaluation. Perhaps this means that most users are able to recognize clear differences in quality, though it could well be a coincidence: we do not have enough data points to make this kind of determination with confidence.

6 Lexicographic recommendations

The majority of the dictionaries investigated in the present study appear to be quite good. The scores that prevailed were "mostly satisfactory" or "satisfactory". Ling.pl was the exception, with scores gravitating towards "partly acceptable". Diki was a clear winner,

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

Table 16 *Final evaluation scores, evaluation-based ranks, and user survey ranks compared*

dictionary	evaluation score	evaluation rank	survey rank
Diki	4.3	1	1
Wiktionary	3.3	2	5
PONS	3.2	3	4
bab.la	3.0	4	2
WordReference	2.9	5	6
Ling.pl	2.0	6	3

with the average score exceeding the “satisfactory” grade. The four remaining dictionaries received similar scores between 2.9 and 3.3, which corresponds to a label of “satisfactory”.

Diki had problems with its coverage of neologisms (“partly acceptable”), and the editors would do well to extend the coverage in this respect. Minor problems were also observed in its usage-level labelling: this aspect should be revised and the labelling applied more consistently. With regard to cross-referencing, a greater number of microstructural elements should be hyperlinked, such as individual words in example sentences so that users can get help when they run into unknown vocabulary in an example. When it comes to pronunciation, offering transcription and audio recordings in American English along with the current British English would be an improvement. Usage notes would alert Polish learners of English to specific problematic issues. Diki might also consider introducing entry navigation devices for easier navigation between the senses.

Wiktionary was ranked in second place overall, but it failed in the coverage of multi-word expressions and usage-level labelling. This should not be surprising, given that this dictionary is edited collectively by enthusiastic amateurs (Lew, 2014; Meyer & Gurevych, 2012). Its coverage of neologisms could also be improved. Multi-word expressions should be discoverable independently of the component lexemes. Absence of words that should be labelled may indicate that the overall coverage of less central items is poor. Users of Wikipedia who tried to search for inflected forms or misspelled their target items would be at a loss. Finally, entry navigation devices ought to be introduced.

The third best dictionary, PONS, was unsuccessful in its coverage of neologisms and only partly acceptable in the coverage of technical vocabulary and multi-word expressions, presence of usage-level labels, provision of example sentences, and presentation. The scope and accuracy of usage-level labels should be improved. The dictionary should expand its exemplification material. Usage notes would be a welcome addition, as would pictorial illustrations. Expansions of grammatical codes and symbols should be provided, and advertising toned down.

Bab.la’s coverage of neologisms and multi-word expressions could be improved on, along with labelling. Individual words in examples ought to be hyperlinked and the provision of pronunciation information extended by adding transcription and the British variety (to match the British flag symbol displayed in the dictionary). Usage notes and collocational information would be welcome. Grammatical information is only rudimentary, with missing noun countability indication and comparative/superlative forms of irregular adjectives.

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

The dictionary would benefit from pictorial illustrations and entry navigation aids. Advertisements should be less intrusive.

WordReference should improve its coverage of entries in general, as the coverage in all three categories of items examined in the study left much to be desired. As in the other dictionaries, labelling ought to be improved on. Expansions of grammatical codes should be given, along with information on noun countability and comparative and superlative forms of irregular adjectives. Advertisements could be made less intrusive, while pictorial illustrations would be a welcome addition.

Many of the problems evident in Ling.pl are a consequence of its adopted aggregator model, resulting in redundant and inconsistent presentation that is difficult to navigate and lacking in cross-references and example sentences. Even though the resource offers access to several dozen dictionaries, its coverage of vocabulary was not fully satisfactory. The dictionary fails to find inflected or misspelled forms.

To sum up, the most persistent weakness across the dictionaries evaluated was their coverage of neologisms and absence of usage-level labels. The only criterion on which all the dictionaries were very successful was part-of-speech labelling.

Going beyond the strict list of criteria, bilingual online dictionaries would do well to improve the discoverability of multi-word units, since language learners do not always immediately see such units for what they are on encountering them in a text. Doing so would discourage simplistic word-for-word translation and promote the idiom-principle approach. Further, dictionary users themselves could be engaged more fully in commercial projects, by allowing user comments and discussions to enhance entries (such as bab.la’s user forum). These, however, should remain clearly distinguished from the ‘professional’ core content.

7 Conclusion

A general aim of the present study was to propose a comprehensive framework for the evaluation of online bilingual dictionaries, building on the work of Pearsons and Nichols (2013) for monolingual dictionaries of English, supplemented with findings from other relevant publications. Although in this study the framework has been applied in practice to English-Polish dictionaries, it is directly applicable to any online bilingual dictionary with English as the source language, and would require only minor adaptation to serve other language pairs.

A more specific aim of the study was to apply the proposed framework to evaluate selected English-Polish dictionaries freely available online, so as to offer guidance to teachers and learners on the relative quality of cost-free online dictionaries for the growing proportion of younger learners. Of course, there do exist better, professionally edited dictionaries for Polish learners of English (Lew & Adamska-Sałaciak, 2015), but they tend not to be available online, let alone free of charge. Perhaps the growing popularity of dictionary apps for small-screen devices will improve the situation in the near future, but for the moment, given that many young learners exhibit a strong preference for online resources available at no cost, we should at least offer them guidance as to which of the freely available resources are relatively better than others.

Overall, the best dictionary according to our criteria (Diki) turned out to be quite satisfactory. This was also the dictionary that our survey participants identified most frequently and valued most (if we ignore *MEGAsłownik*, which had gone out of business before the study was completed).

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press

We hope that this study will inspire metalexicographers to carry out systematic analyses of online dictionaries, particularly bilingual dictionaries, as these are most often used by language learners. Such analyses could be of use to both language learners and teachers as a reliable indicator of which reference works can be trusted, especially when it comes to non-professional dictionaries.

Before attempting a replication, it may be advisable to reconsider some aspects of the present study with a view to improving them. First, the number of items tested in the evaluation of coverage was rather small, and the selection of categories somewhat arbitrary. A larger wordlist would increase the reliability of the evaluation framework as an instrument, though it might also limit the number of dictionaries that could be evaluated. Items might be sampled from several frequency bands, and the selection of neologisms informed by checking the candidates in a reliable monitor corpus. Running automated queries to test coverage might be considered, though the degree of success would here depend on the features of the specific dictionary interfaces. A more general aspect is the choice of criteria: with a different set (or weighting), the results might differ. It also has to be noted that the quality of equivalents was not evaluated – admittedly an important aspect in gauging the quality of bilingual dictionaries. Finally, the number of criteria used in evaluating coverage and treatment was larger than that for access and presentation. A more balanced distribution across these areas, or else a different approach to weighting the partial scores, might be considered.

Acknowledgements

This work was supported in part by the Polish National Science Centre (Narodowe Centrum Nauki), under grant DEC-2013/09/B/HS2/01125 to the first author.

The authors wish to express their gratitude to all the anonymous participants.

Supplementary material

For supplementary material referred to in this article, please visit https://doi.org/10.1017/S0958344016000252

References

Davies, M. (2008) The Corpus of Contemporary American English. http://corpus.byu.edu/coca/.

De Schryver, G.-M. (2003) Lexicographers' dreams in the electronic-dictionary age. *International Journal of Lexicography*, **16**(2): 143–199.

Dziemianko, A. (2011) Does dictionary form really matter? In: K. Akasu and S. Uchida (eds.), *ASIALEX2011 Proceedings. Lexicography: Theoretical and practical perspectives*. Kyoto: Asian Association for Lexicography, 92–101.

Dziemianko, A. (2012) Why one and two do not make three: Dictionary form revisited. *Lexikos*, **22**: 195–216.

Dziemianko, A. (2015) Colours in dictionaries: A case of functional labels. *International Journal of Lexicography*, **28**(1): 27–61.

L'Homme, M.-C. and Cormier, M. C. (2014) Dictionaries and the digital revolution: A focus on users and lexical databases. *International Journal of Lexicography*, **27**(4): 331–340.

Leaney, C. (2007) *Dictionary activities*. Cambridge: Cambridge University Press.

Levy, M. and Steel, C. (2015) Language learner perspectives on the functionality and use of electronic language dictionaries. *ReCALL*, **27**(2): 177–196.

Lew, R. (2011) Online dictionaries of English. In: P. A. Fuertes-Olivera and H. Bergenholtz (eds.), *e-Lexicography: The Internet, digital initiatives and lexicography*. London/New York: Continuum, 230–250.

Lew, R. (2012) How can we make electronic dictionaries more effective? In: S. Granger and M. Paquot (eds.), *Electronic lexicography*. Oxford: Oxford University Press, 343–361.

Lew, R. (2014) User-generated content (UGC) in online English dictionaries. *OPAL – Online publizierte Arbeiten zur Linguistik*, **2014**(4): 8–26.

Lew, R. and Adamska-Sałaciak, A. (2015) A case for bilingual learners' dictionaries. *ELT Journal*, **69**(1): 47–57.

Lew, R. and De Schryver, G.-M. (2014) Dictionary users in the digital revolution. *International Journal of Lexicography*, **27**(4): 341–359.

Lew, R. and Mitton, R. (2011) Not the word I wanted? How online English learners' dictionaries deal with misspelled words. In: I. Kosem and K. Kosem (eds.), *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011*. Ljubljana: Trojina – Institute for Applied Slovene Studies, 165–174.

Lew, R. and Mitton, R. (2013) Online English learners' dictionaries and misspellings: One year on. *International Journal of Lexicography*, **26**(2): 219–233.

Lew, R., Grzelak, M. and Leszkowicz, M. (2013) How dictionary users choose senses in bilingual dictionary entries: An eye-tracking study. *Lexikos*, **23**: 228–254.

Meyer, C. M. and Gurevych, I. (2012) Wiktionary: A new rival for expert lexicons? Exploring the possibilities of collaborative lexicography. In: S. Granger and M. Paquot (eds.), *Electronic lexicography*. Oxford: Oxford University Press, 259–291.

Müller-Spitzer, C., Koplenig, A. and Töpel, A. (2012) Online dictionary use: Key findings from an empirical research project. In: S. Granger and M. Paquot (eds.), *Electronic lexicography*. Oxford: Oxford University Press, 425–457.

Nesi, H. (2000) Electronic dictionaries in second language vocabulary comprehension and acquisition: The state of the art. In: U. Heid, S. Evert, E. Lehmann and C. Rohrer (eds.), *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000, Stuttgart, Germany*. Stuttgart: Institut for maschinelle Sprachverarbeitung, Universität Stuttgart, 839–841.

Nielsen, S. (2008) The effect of lexicographical information costs on dictionary making and use. *Lexikos*, **18**: 170–189.

Pearsons, E. and Nichols, W. (2013) Toward a framework for reviewing online English dictionaries. *Dictionaries: Journal of the Dictionary Society of North America*, **34**: 201–210.

R Core Team. (2015) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Stein, G. (1991) Illustrations in dictionaries. *International Journal of Lexicography*, **4**(2): 99–127.

Yamada, S. (2010) EFL dictionary evolution: Innovations and drawbacks. In: I. Kernerman and P. Bogaards (eds.), *English learners' dictionaries at the DSNA 2009*. Tel Aviv: K Dictionaries, 147–168.

Yamada, S. (2013) A test of the proposed framework for reviewing online dictionaries: m-w.com, dictionary.com, macmillandictionary.com, dictionary.cambridge.org, and oxforddictictionaries.com. *Dictionaries: Journal of the Dictionary Society of North America*, **34**: 211–224.

https://doi.org/10.1017/S0958344016000252 Published online by Cambridge University Press