Dictionaries and Technology

Robert Lew

Adam Mickiewicz University in Poznań

1. The Corpus Revolution in Lexicography

Lexicographers have always understood the importance of working with authentic language data in describing language. Before the advent of computers, serious dictionary-making involved an arduous process of manually collecting millions of citations from literature. Dictionary-makers were sometimes assisted in this task by the educated public through special reading programs. The resulting citations were placed on citation slips and painstakingly arranged in voluminous files. This method was laborious in the extreme, and it also had a major methodological flaw: Human readers naturally focus on the unusual. As a result, any database of manual citations tends to emphasize instances of creative use of language, but the uninspiring everyday uses of common words remain unnoted, as those seem too trivial to be worth recording.

In the 1980s, dictionary-making underwent a major revolution thanks to the pioneering COBUILD project (Sinclair, 1987). This was the first lexicographic project to make systematic use of text corpora, and the corpus revolution was thus initiated with learners' dictionaries. From there, it gradually spread to other types of lexicography as well as kick-started the development of corpus linguistics.

The COBUILD team had assembled an electronic collection of 7.3 million words of text for the compilation of the dictionary, and this number grew with the addition of further text, to reach 18 million words for the final editing phase. The corpus — initially known as the Birmingham Collection of English Text and later renamed the Bank of English — seemed huge at the time, but compared to today's corpora holding billions of words, it is very small.

Within dictionary-making, corpora are useful in a number of ways. They form the material basis for selecting potential headwords and identifying the senses and uses to be covered. They provide objective data for the description of the morphological and syntactic behavior of words, as well as the relative frequency of alternative spelling forms. Identification of collocational behavior and significant multi-word units are among the more advanced applications, requiring corpora of larger size and language engineering tools of greater sophistication. Text corpora offer lexicographers ready access to large numbers of potential examples of authentic use of language, and indeed COBUILD's original selling point was that it dealt with *real* language. However, corpus lexicographers realize today that authenticity alone is not in itself a guarantee that an example is suitable for inclusion in a dictionary entry.

COBUILD describes its methodology as *corpus-driven*. This term refers to a predominantly inductive approach, where one starts with the evidence itself. The approach is sometimes opposed to one characterized as *corpus-based*, in which the corpus plays a less central, more complementary role, mostly as a source of (post-hoc) evidence for pre-existing ideas. In the realm of lexicography, a project might rely partially on a corpus, but also incorporate other independent considerations. For example, in ordering senses within an entry, lexicographers might be guided not just by the objective frequency of specific identifiable uses, but consider which sense is semantically more basic. Thus, *summit* in the sense 'top of the mountain' might be listed first, before the 'important political meeting' sense, as the latter sense is

This is a preprint version of:

Lew, Robert. 2013. 'Dictionaries and Technology' In Chapelle, Carol (ed.), *The Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell. (http://onlinelibrary.wiley.com/book/10.1002/9781405198431)

derived from the first. But in COBUILD, the more textually frequent political sense is listed first, as dictated by the primacy of the corpus principle.

Lexicographers producing dictionaries for language learners may also utilize learner corpora. These are collections of non-native texts written (or, less usually, spoken) by language learners at various levels of proficiency. A learner corpus can be explored lexicographically to identify specific problems that learners of a language experience, so that attention can be drawn to the problematic points in the relevant entries, either by the appropriate selection of examples, or by explicitly stating the problem in a usage box accompanying the entry.

2. Corpus Query Systems

A corpus, to be useful, needs to be equipped with a front-end interface through which corpus users can easily interrogate the text collection. The standard way of presenting corpus data has been through concordance lines displaying the target word (keyword) in a textual context, usually a single line of text. In the original COBUILD project, concordances for individual words had been printed off on paper, as computers were then too crude to generate concordances in real time.

Today's interfaces to text corpora provide ever more sophisticated ways of assisting lexicographers in getting to the usable entry as efficiently as possible, and with a minimum of effort. Amongst the most innovative is word-profiling software, designed to generate structured views of search items, as by grouping patterns of use or collocates. A free resource of this type for English is the Just the Word service (http://www.just-the-word.com). The SketchEngine (Kilgarriff & Tugwell, 2002), a leading commercial system, has even greater flexibility, allowing different types of presentation: concordances, collocates, synonyms, synonym comparisons, and is available for a growing number of languages (e.g. Radziszewski, Kilgarriff, & Lew, 2011). The system is equipped with built-in corpora; in addition, users can build their own corpora. A most useful feature of the SketchEngine are word sketches: one-page summaries of a word's grammatical and collocational behavior (see Figure 1).

One consequence of the growing size of corpora is that textual evidence may become overwhelming and impossible to examine in detail. To remedy the problem, language technology is applied to extract the best example sentences from the many potential ones in a corpus. One system that does this is GDEX (Kilgarriff, Husak, McAdam, Rundell, & Rychlý, 2008). The quality of its output can be tested at http://forbetterenglish.com/. Efforts like these aim at relieving a human lexicographer of as much of the drudgery as possible, so that a maximum of tasks are automated (Kilgarriff, Kovář, & Rychlý, 2010).

3. Corpora in Bilingual Lexicography

In a bilingual dictionary project, two single-language corpora may be used, much as in compiling a monolingual dictionary. Another approach involves the collection of comparable or parallel corpora of the two languages involved. Comparable corpora contain texts in the two languages that are functionally equivalent and in similar proportions. Parallel corpora consist of text pairs, with one (at least) usually being a translation. Parallel corpora may be aligned (synchronized), so that the lexicographer can quickly assess what words and structures are typically used in analogous contexts. With the help of specialized NLP tools, one can attempt to extract typical word-to-word equivalents between the languages. This works best for (scientific) terms, and is useful in the compilation of specialized dictionaries.

This is a preprint version of:
Lew, Robert. 2013. 'Dictionaries and Technology' In Chapelle, Carol (ed.), *The Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell. (http://onlinelibrary.wiley.com/book/10.1002/9781405198431)

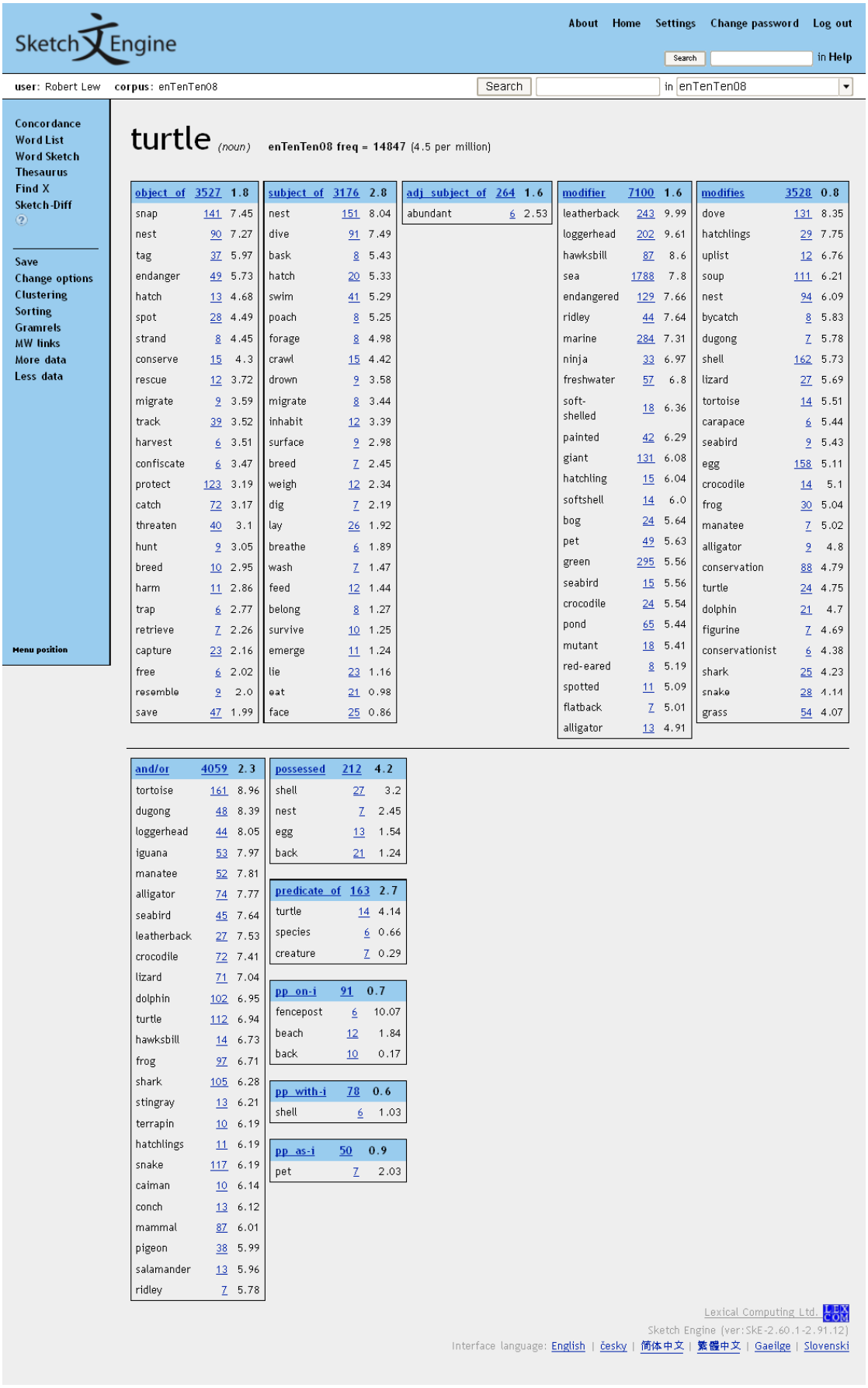

Figure 1: Word Sketch for the English noun **TURTLE**

This is a preprint version of:
Lew, Robert. 2013. 'Dictionaries and Technology' In Chapelle, Carol (ed.), *The Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell. (http://onlinelibrary.wiley.com/book/10.1002/9781405198431)

4. Dictionary Writing Systems

It is entirely possible to assemble a lexicographer's workbench from stand-alone tools such as a text editor, database application, format conversion tools, workflow and planning software, etc. However, many editors and publishers now believe that there is an advantage to be gained from integrating all or most of the functionality needed for a dictionary project within a single piece of software, a Dictionary Writing System (or DWS, for short). Publishers or teams may undertake to build a customized DWS to suit their own purposes, or they can adapt one that is already available. Several DWS suites are now available, the best-known being: IDM DPS, TLex (see Figure 2), iLEX, ABBY Lingvo Content (commercial packages); and Lexique Pro, Léacslann (non-commercial systems).

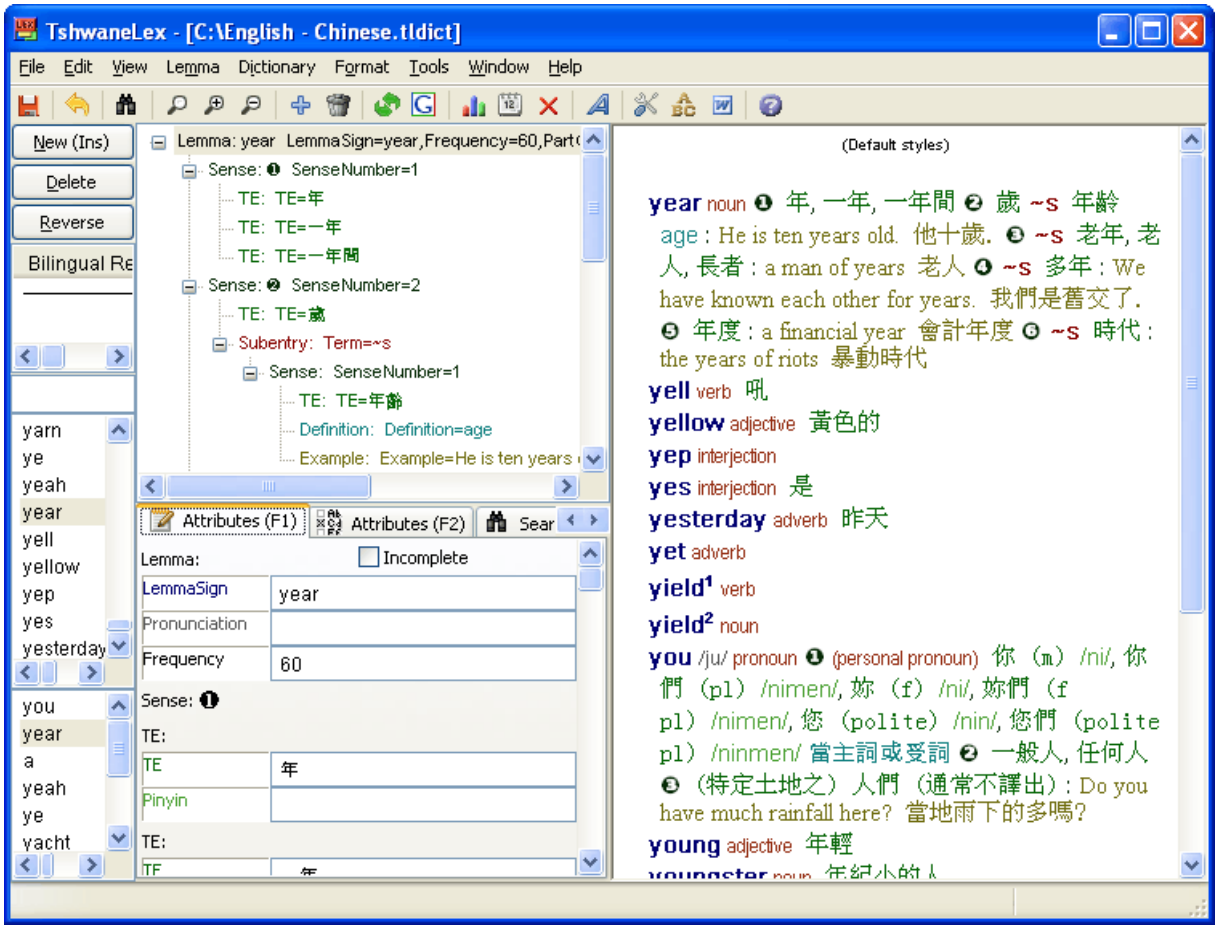

Figure 2: TLex as an Example of Dictionary Writing Software

A Dictionary Writing System will assist in the many tasks involved in carrying out dictionary projects. It facilitates the editing of entries, typically in XML. Lexicographic content (or data) is now normally separated from presentation, which is handled separately, so that decisions affecting how a specific data field is displayed can be made at any point, and will automatically affect all data of a given type (such as numbering and indenting senses, displaying example sentences in italics, separating them with bullets, etc.). A state-of-the-art system will simulate the final presentation in real time. For multi-word expressions, cross-checks can be kept so that lexicographic data are only entered once, but can be displayed under all the relevant headwords in a uniform fashion. Team work may be supported by storing all data on a central server, and locking files in use. Headwords may be grouped using various criteria and assigned to specific experts. Finally, workflow can be organized and progress monitored with statistical reports.

This is a preprint version of:
Lew, Robert. 2013. 'Dictionaries and Technology' In Chapelle, Carol (ed.), *The Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell. (http://onlinelibrary.wiley.com/book/10.1002/9781405198431)

5. New Workflow for Online Dictionaries

Technology has revolutionized not just the ways in which dictionaries are compiled, but also how they are packaged; at this time we are witnessing a transition from the traditional print medium to the electronic medium. More and more language learners around the world are starting to use electronic dictionaries, such as PC-based applications, handheld stand-alone products (particularly popular in Asia), online dictionaries, mobile-phone applications, or dictionaries on e-book readers.

Online dictionaries, in particular, call for a new approach to dictionary-making. The traditional stages of dictionary compilation no longer obtain as they did for printed titles. Since online dictionaries can be incrementally upgraded as often as needed (even on a daily basis), the development cycle may now include simultaneous feedback from dictionary users (De Schryver & Prinsloo, 2001). For example, if many users begin to search for a particular item, it may be added to the dictionary promptly without waiting for a traditional new edition.

6. How Technology Serves Dictionary Users

Well-designed electronic dictionaries offer a number of advantages over the traditional print format, although not all e-dictionaries will actually incorporate the features afforded by the electronic medium. Besides the traditional phonetic transcription, dictionary users (particularly language learners) may be able to listen to the headword (possibly also example sentences) being pronounced by a native speaker. Spelling the searched item incorrectly no longer needs to result in failure, as electronic dictionaries will try to guess at the item actually meant (Lew & Mitton, 2011, 2013).

A known problem of paper dictionaries (for languages with alphabetic writing systems) is that they are traditionally organized around single orthographic words. This creates specific difficulties for dictionary users trying to find multi-word lexical units (such as idioms). A well-designed interface to an electronic dictionary will make it possible to locate a multi-word unit without having to know under which headword the expression is nested, even in cases when the dictionary user is not in fact aware that a multi-word unit is involved, as frequently happens with language learners struggling to understand a text in the foreign language.

When reading texts in electronic format, dictionary consultation can be facilitated once the software for reading and the dictionary can "talk" to each other. For example, on an e-book reading device all that the user should have to do is tap on the word which they believe is problematic, and this word should be looked up automatically. Ideally, the smart dictionary should then examine the textual context for evidence of multi-word expressions and for clues to the particular sense that the search word is likely to be used in. Then, the entry presented should reflect the outcome of these findings by suppressing information which is likely to be irrelevant, and selectively presenting the data that may be of value in this particular comprehension problem.

Electronic dictionaries do not need to be restricted to a single static view, as is the case with printed dictionaries. Presentation of lexicographic data may be adjusted depending on uses and users. In the simplest case, detailed information on grammatical complementation, collocational behavior, or synonyms is superfluous for someone consulting a dictionary while reading a text in a foreign language. But the same detailed data would be very useful for someone writing an essay.

The storage space of electronic media makes it possible to include more lexicographic data than has been possible in printed volumes. This does not mean that all the data available should be presented to the user at all times, as doing so will often result in information

This is a preprint version of:
Lew, Robert. 2013. 'Dictionaries and Technology' In Chapelle, Carol (ed.), *The Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell. (http://onlinelibrary.wiley.com/book/10.1002/9781405198431)

overload, making the entry less helpful. However, with the right kind of control over what data is presented at which time, having access to richer stores of data may offer significant improvements over printed dictionaries. For instance, extra example sentences might be included. For more advanced dictionary users, such an *example bank* could take the form of a full-featured text corpus.

Electronic displays can accommodate media content beyond the usual text (and perhaps the occasional picture) inherited from printed dictionaries. Routine inclusion of graphics such as full-color photographs no longer translates into excessive publication costs, and as long as these are offered as an optional component, they no longer consume valuable space. Video and animation are another possibility, although the available evidence suggests that neither benefit the user (Chun & Plass, 1996; Lew & Doroszewska, 2009). If the electronic device has sound capability, then the dictionary can use audio material, such as pre-recorded or synthesized representation of headwords, and possibly also example sentences and definitions. Some words, notably those representing musical instruments, are associated with characteristic sounds, and recordings of these can enhance the users' comprehension, and, as a long-term consequence, retention of these vocabulary items.

Modern internet technology makes it particularly easy to link and embed content. While this holds potential for building innovative lexical resources, including ones for multiple languages (e.g. http://dict.cc/), the option is sometimes abused to produce so-called aggregator sites (Lew 2011), which merely pull content from several sources with no real concern for quality. Non-expert users, unable to assess their actual value without proper guidance, may be attracted by the generic-sounding domain names or inflated claims of the number of words or languages covered. In contrast, the more active users are taking lexicography into their own hands, taking advantage of Web 2.0 technology. Such user-generated content is most fruitful in the area of specialized lexicography (Lew 2013).

Modern technology blurs the traditional distinctions between different types of dictionaries, and dictionaries versus other lexically-based tools. To refer to two examples already given above: ForbetterEnglish.com is an automatically generated dictionary of collocations, illustrated with examples selected from a corpus — also automatically; Just the Word (http://www.just-the-word.com/) is an interesting lexical profiling tool capable of correcting non-native-like collocational choices based on corpus evidence. Today, dictionaries may be integrated as part of larger software suites for language learners or translators, such as writing assistants (software designed to provide support in writing tasks). It is likely that this trend will continue, with dictionaries of the future being rather different from the ones we know.

References

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *Modern Language Journal, 80*(2), 183-198.

De Schryver, G.-M., & Prinsloo, D. J. (2001). Fuzzy SF: Towards the ultimate customised dictionary. *Studies in Lexicography, 11*(1), 97–111.

Kilgarriff, A., Husak, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal & J. DeCesaris (Eds.), *Proceedings of the XIII EURALEX International Congress* (pp. 425-432). Barcelona: Universitat Pompeu Fabra.

Kilgarriff, A., Kovář, V., & Rychlý, P. (2010). Tickbox lexicography. In S. Granger & M. Paquot (Eds.), *eLexicography in the 21st century: New challenges, new applications* (pp. 411-418). Louvain-la-Neuve: Cahiers du CENTAL.

This is a preprint version of:

Lew, Robert. 2013. 'Dictionaries and Technology' In Chapelle, Carol (ed.), *The Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell. (http://onlinelibrary.wiley.com/book/10.1002/9781405198431)

Kilgarriff, A., & Tugwell, D. (2002). Sketching words. In M.-H. Corréard (Ed.), *Lexicography and natural language processing. A festschrift in honour of B.T.S. Atkins* (pp. 125-137): EURALEX.

Lew, R. (2011). Online dictionaries of English. In P. A.Fuertes-Olivera, & H. Bergenholtz (Eds.), *e-Lexicography: The Internet, Digital Initiatives and Lexicography* (pp. 230-250). London/New York: Continuum.

Lew, R. (2013). User-generated content (UGC) in online English dictionaries, *OPAL - Online publizierte Arbeiten zur Linguistik*. 9-29.

Lew, R., & Doroszewska, J. (2009). Electronic dictionary entries with animated pictures: Lookup preferences and word retention. *International Journal of Lexicography, 22*(3), 239-257.

Lew, R., & Mitton, R. (2011). Not the word I wanted? How online English learners' dictionaries deal with misspelled words. In I. Kosem & K. Kosem (Eds.), *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011, Bled, 10-12 November 2011* (pp. 165-174). Ljubljana: Trojina – Institute for Applied Slovene Studies.

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

Radziszewski, A., Kilgarriff, A., & Lew, R. (2011). Polish word sketches. In Z. Vetulani (Ed.), Human language technologies as a challenge for computer science and linguistics. Proceedings of the 5th Language & Technology Conference (pp. 237-242). Poznań: Fundacja Uniwersytetu im. A. Mickiewicza.

Sinclair, J. (Ed.). (1987). *Looking up: An account of the COBUILD project in lexical computing*. London - Glasgow: Collins.

Suggested Readings

Abel, A. (2012). Dictionary writing systems and beyond. In S. Granger & M. Paquot (Eds.), *Electronic lexicography* (pp. 83-105). Oxford: Oxford University Press.

Abel, A., & Klosa, A. (2012). Der lexikographische Arbeitsplatz – Theorie und Praxis. In R. V. Fjeld & J. M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress* (pp. 413-421). Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

Atkins, B. T. S., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Hanks, P. (2012). The corpus revolution in lexicography. *International Journal of Lexicography, 25*, 398–436.

Kilgarriff, A. (2005). If dictionaries are free, who will buy them? *Kernerman Dictionary News, 13*, 17-19.

Lew, R. (2010). Multimodal lexicography: The representation of meaning in electronic dictionaries. *Lexikos, 20*, 290-306.

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

Rundell, M. (2012). 'It works in practice but will it work in theory?' The uneasy relationship between lexicography and matters theoretical. In R. V. Fjeld & J. M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress* (pp. 47-92). Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

Trap-Jensen, L. (2010). One, two, many: Customization and user profiles in internet dictionaries. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV Euralex International Congress* (pp. 1133-1143). Ljouwert: Afûk.

Verlinde, S., Leroyer, P., & Binon, J. (2010). Search and you will find. From stand-alone lexicographic tools to user driven task and problem-oriented multifunctional leximats. *International Journal of Lexicography, 23*(1), 1-17. doi: 10.1093/ijl/ecp029.

This is a preprint version of:

Lew, Robert. 2013. 'Dictionaries and Technology' In Chapelle, Carol (ed.), *The Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell. (http://onlinelibrary.wiley.com/book/10.1002/9781405198431)