Are we witnessing the death of dictionaries?

Hilary Nesi
Coventry University (United Kingdom)
h.nesi@coventry.ac.uk

Almost every day now I get news of publications, conferences, seminars and working parties devoted to the influence of Large Language Models (and particularly Chat GPT) on language teaching and learning. These notifications rarely refer to human-readable dictionaries (as opposed to the machine-readable dictionaries used by computational linguists) – the link between corpora and lexicography is no longer really evident. Most of us now search for word meanings via our search engines without specifying a particular dictionary, or even a particular kind of dictionary; the web analytics site *Google Trends* (https://trends.google.com/trends/explore) reveals that worldwide searches for the term 'corpus' have remained fairly steady since 2004, but searches for 'dictionary' have fallen by about 75%, while searches for 'meaning' have taken an opposite trajectory. According to many studies of dictionary user behaviour, language learners have always prioritised meaning over the other types of information a good dictionary provides, for example relating to pronunciation, grammar, collocations and pragmatic use. Perhaps such explanations have now become completely redundant, as we are entering the 'human machine era' (Sayers et al. 2021), and intelligent tools (for well-resourced languages) can correct anything that we write, and translate entire passages of text.

Corpus resources are still created and used by LSP practitioners, of course. Easy-to-use corpus query software packages and ready access to digital texts online have encouraged LSP teachers and learners at all levels of expertise to create their own corpora for specific purposes, and several journals, and special issues of journals, are dedicated to corpus studies with an emphasis on language use in specific domains. The amount of corpus-themed ESP research has increased dramatically over the past 30 years, as measured by Hyland and Jiang's bibliometric analysis of 3,500 articles on ESP topics (2021). On the whole, the corpus studies published in academic journals are

intended for readership within the LSP and Applied Linguistics community, and aim to identify typical lexical or lexicogrammatical features in specific types of texts, or to identify ways of helping learners to explore these features for themselves through some form of data-driven learning. They are explicit about the contents of their corpora, and their findings can often be replicated using the same or similar collections of texts. You could say that these researchers are following in the footsteps of lexicographers and corpus linguists such as John Sinclair, Rosamund Moon and Patrick Hanks, who had one foot in academia and one foot in the world of commercial publishing. Nowadays, however, there seems to be little overlap between the (usually small-scale) corpus-based studies conducted by applied linguists, and the big data analyses conducted by Artificial Intelligence (AI) tools.

Conrad (2000) predicted that corpus studies would affect grammar teaching in the 21st century by blurring the distinction between grammar and vocabulary, and placing more emphasis on register differences and appropriacy in specific contexts. She was thinking about the design of future grammar books (and the part she herself had played in the development of *Longman Grammar of Spoken and Written English*), but her predictions also reflected the changes that had already taken place in lexicography in the 1980s and 1990s, the time of the 'corpus revolution' (Rundell & Stock 1992). In the early 1980s computer memories were just becoming large enough, and computer programming languages sophisticated enough, to process multimillion word corpora; the first corpus-based dictionary, *Collins COBUILD English Dictionary* (1987) drew on a corpus of 7.3 million words, tiny by modern standards but revolutionary in the way it provided lexicographers with information about frequency, collocations, and typical contexts for common words. Corpus lexicography soon spread to other UK publishers of learners' dictionaries, and then to many other dictionary publishers around the world.

As far back as the 1950s Wittgenstein had said that 'the meaning of words lies in their use' and Firth had made the famous pronouncement 'you shall know a word by the company it keeps'. As corpus resources grew, lexicographers were able to demonstrate the truth of these sayings, and show that word meanings are attached to patterns rather than to words in isolation. Early corpora were not large enough to provide adequate contexts for infrequent words, or even to distinguish the rarer (and sometimes technical) meanings of frequent words, but gradually growth in the number of available corpora, and in corpus size, provided lexicographers with more

information relating to syntactic behaviour, domain restrictions, register, pragmatics and common learner errors (drawing on learner corpora, generally made up of scripts from language proficiency exams and language classes). Work in these areas has paved the way for AI generated language support that can be tailored to suit specific purposes; such support goes way beyond the level of the individual word, but does not supply learners with the explicit linguistic information contained in old-style dictionary entries.

Longman Grammar of Spoken and Written English was published as a print edition in 1999, so when Conrad was making her predictions for the 21st century she was not thinking about the possible long-term effects of digitalisation. At that time there were still only about 400 English dictionaries on the World Wide Web, and probably far fewer grammar books. Electronic dictionaries did exist, but they were distributed on CD-ROM or on stand-alone mobile devices, did not contain much more content that their print versions, and did not constitute a threat to the highly lucrative dictionary market. This all changed in the early 2000s, when internet access became faster and more reliable, leading to the mass migration to the web of reference works of all kinds – maps, encyclopedias, grammar books and dictionaries. The print edition of the *Encyclopedia Britannica* ceased production in 2010, and the print version of the Macmillan English Dictionary followed suit in 2013. Other dictionary publishers gradually dropped their print editions, so that today only pockets of hard-copy dictionary use remain around the world – for some neglected languages, amongst the dispossessed, and in under-resourced areas where there is no internet access, for example. Sycz-Opoń (2024) found that translators still use print dictionaries, but even these language professionals are most likely to refer to search engines.

The consequence of moving dictionaries online was that most people started to expect their lexicographical information to be available for free; the loss of income from dictionary sales meant that many lexicographers lost their jobs, and there was little financial support from publishers for further innovation in dictionary design and content. (We may note that everyday users started to create and edit entries in collaborative dictionaries without expert help at about this time). Lexicography experts predicted that the dictionary market would fracture, moving towards more functionally diverse products for many different types of user (Kilgarriff 2005, Rundell 2011). This fracturing may be coming to pass with the development of AI applications, but the migration from print to web immediately resulted in a

blurring of the boundaries between dictionaries for different types of user and for different types of purpose: rather than fragmenting, online dictionaries tended to expand and hybridize, no longer distinguishing between material for different user groups such as LSP learners, children and proficient speakers with a general interest in language. Many more (web-sourced) examples were added to web dictionary entries, and more technical and specialist words were added because the new availability of search logs indicated that users wanted to look up these kinds of words. Print dictionaries had been getting bigger and bigger with each new edition, to the point when they had become difficult for users to carry around, but web dictionaries had no real space constraints. Moreover, for information beyond the scope of a single chosen dictionary, the establishment of 'aggregate' portals enabled simultaneous consultation of several dictionaries at once – monolingual and bilingual, alphabetic and thematic, general, technical and encyclopaedic.

The advantages of web-based reference materials are fairly obvious. They are much easier to consult 'on the go' than reference materials in print, especially if using portable tablets and mobile phones; this suits users, who tend to look up information when they are already busy doing something else. The quicker the consultation, the less it interferes with task flow. Some of the common problems traditionally reported in dictionary user studies have been alleviated because users of dictionaries in electronic form do not need to learn phonetic symbols (they have audio files), and search interfaces have been enhanced so they do not need to know the order of the alphabet, or any other organisational system – they do not even need to spell correctly or to think of the right headword, as searches can now often start from any word or phrase in the entry.

Educators may lament that their LSP students are not using technical dictionaries as much as they should, and are not adequately trained in LSP dictionary use (Glušac & Milić, 2020), but technical terms in the better-resourced languages may be explained quite well in some large general web dictionaries (or via portals that lead to multiple, more specialised, dictionaries), and basic dictionary skills training may no longer be such an issue, given the flexibility of modern web dictionary search routes. Perhaps any training should focus on critical evaluation and awareness raising, as this would be of use to all LSP learners, whether they choose to solve their language problems with the help of dictionary entries, AI tools, or a mixture of both. In East Asia, AI generated language support seems to be becoming

inseparable from dictionary information. Portals such as *NetEase Youdao* in China, *Dr Eye* in Taiwan, and *Naver* in Korea offer access to a range of Western and local dictionaries and attract millions of users because of their additional AI affordances (unavailable on Western web dictionary sites). Some of the information that these portals provide verges on the nonsensical; there are misspellings and misleading examples (taken from local sources such as social media sites), and huge numbers of strange derived forms (such as *linguistician*) which seem to have been generated by applying word formation rules automatically, without regard for actual usage. For the moment these resources are highly problematic (Nesi 2012, Yeung 2022) but the quality may eventually improve as the technology advances. It is noteworthy that they are starting to downplay their dictionary status; for example *Dr Eye* is now promoted as "one of the most popular translation software in the whole Chinese-spoken area" (https://www.dreye.com/en/product/product.php), and *NetEase Youdao* dictionary, which has an enormous following in mainland China, now describes itself as a "translation and language learning app" (https://shared.youdao.com/www/about.html).

AI has for some time been essential for the development of corpus analysis software, automated tagging systems, and other aspects of corpus research (Curry et al. 2024). Currently, lexicographers are debating whether AI can generate dictionary entries (see, for example, de Schryver 2023, Lew 2024), but maybe eventually dictionaries as we know them will no longer exist at all. We have seen dictionary provision progress from pre-corpus days, when data on word behaviour was largely lacking, to pre-web days, when corpora were too small to meet LSP learners' needs, and now to the present day, when the volume of digital data is too immense for humans to make sense of, and might best be analysed using deep learning techniques and sophisticated algorithms. I suspect that, given the choice, most LSP learners would prefer to go straight to an AI generated passage to find out how a word or phrase behaves in context, whatever the task (reading, writing, translating or simply improving their vocabulary knowledge).

Bur bigger is not always better, as we all know. Large Language Models came into widespread use in 2023, and although most of us, both teachers and students, are still relatively ignorant about what they can achieve, the negative implications of their use in academia have already been widely discussed (see, for example, Kuteeva & Andersson 2024). In this paper I have tried to talk about dictionaries and alternatives to dictionaries, so I will ignore related

issues (the environmental impact of AI, and concerns about ethics, copyright, accuracy and bias). Instead I will conclude by considering some of the possible effects of converting from dictionary use to AI applications.

I have two main concerns:

1. that large datasets may skew learners' word choices and writing styles, so that their language comes closer to journalese and less like language for other specific purposes;
2. that AI applications do not require deep processing, and provide solutions to language problems without explaining how or why these solutions have been arrived at.

We usually know at least something about the sources used by major dictionary publishing houses (Cambridge, Oxford and Pearson, for example) thanks to explanations in publications and on the publishers' websites. These sources have been curated to reflect established views about what a corpus is, and does, as a collection of texts that should be selected "to represent, as far as possible, a language or language variety as a source of data for linguistic research" (Sinclair 2004). On the other hand, it is impossible to trace the sources of the huge quantities of internet texts that models such as Chat GPT will draw on. Journalism predominates on the internet, but journalism genres are not the ones that LSP learners are most likely to encounter in the classroom or in their target situations. Some words, such as *commendable*, *intricate* and *meticulous(ly)*, are disproportionately used by chatbots (Stokel-Walker 2024), and I have noticed that they are starting to appear in the writing of my own students. Will this affect their chances of success as scientific communicators? Or, an alternative concern, will the growing amount of chatbot-generated text on the web eventually affect word use in LSP registers?

Also, instead of taking notes in class, many of my students take photos – of lecture slides, textbooks, and any other written material we examine. They go on to translate or paraphrase the text in these images using AI applications, but most instructors would agree that a better way for learners to understand and retain knowledge is to reframe it, using their own words. Taking photos of text seems to be a shallower processing strategy than copying text verbatim, and copying text verbatim is recognised to be less effective than taking notes longhand (cf. Mueller & Oppenheimer 2014). Dictionary use requires a bit more effort; it invites learners to make choices about which word or phrase to use, and to take usage notes and register

restrictions into account alongside definitions and examples. Admittedly, a lot of lexicographical work still needs to be done in order to describe the lexicon completely, but the alternative chatbot approach provides no explanations at all.

I am not sure what answers there are to these questions, but I do suspect that dictionary use is dying. Perhaps students need more training, as we always say when a disruptive technology appears on the scene. Certainly LSP practitioners need to think about the dictionary advice they give their students – maybe it will not be followed, or maybe it will lead learners to 'dictionaries' that are little more than AI applications in disguise.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly*, *34*(3), 548-560.

Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics*, *4*(1), 100082. https://doi.org/10.1016/j.acorp.2023.100082

de Schryver, G. M. (2023). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, *36*(4), 355-387. https://doi.org/10.1093/ijl/ecad021

Glušac, T., & Milić, M. (2020). How university teachers of English for Specific Purposes and their students employ dictionaries in teaching and learning. *Annual Review of the Faculty of Philosophy, Novi Sad*, *45*(5). http://dx.doi.org/10.19090/gff.2020.5.281-295

Hyland, K., & Jiang, F. (2021). Delivering relevance: The emergence of ESP as a discipline. *English for Specific Purposes*, *64*, 13-25. https://doi.org/10.1016/j.esp.2021.06.002.

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

Kuteeva, M., & Andersson, M. (2024). Diversity and standards in writing for publication in the age of AI—Between a rock and a hard place. *Applied Linguistics*. https://doi.org/10.1093/applin/amae025

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, *11*, 426. https://doi.org/10.1057/s41599-024-02889-7

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science*, *25*(6), 1159-1168. https://doi.org/10.1177/0956797614524581

Nesi, H. (2012). Alternative e-dictionaries: uncovering dark practices. In S. Granger & M. Paquot (Eds.), *Electronic Lexicography* (pp. 357-372). Oxford University Press.

Rundell, M. (2011). Many a mickle maks a muckle. Will there still be dictionaries in 2020? Round table talk at *Electronic lexicography in the 21st century: new applications for new users (eLex2011)*. http://videolectures.net/elex2011_rundell_table/

Rundell, M., & Stock, P. (1992). The corpus revolution. *English Today*, *8*(3) 21-32.

Sayers, D. et al. (2021). *The dawn of the human-machine era: A forecast of new and emerging language technologies*. Report for EU COST Action CA19102 'Language in the Human-Machine Era'. COST: European Cooperation in Science & Technology, University of Jyväskylä. https://doi.org/10.17011/jyx/reports/20210518/1

Sinclair, J. (2004). Corpus and text — Basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice*. https://users.ox.ac.uk/~martinw/dlc/

Stokel-Walker, C. (2024). AI chatbots have thoroughly infiltrated scientific publishing. *Scientific American*. https://www.scientificamerican. com/

article/chatbots-have-thoroughly-infiltrated-scientific-publishing/

Sycz-Opoń, J. (2024). Print dictionaries are still in use: A survey of source preferences by Polish translators. *International Journal of Lexicography*, *37*(2). https://doi.org/10.1093/ijl/ecae004

Yeung, Y. (2022). *Supporting Chinese EFL learners' dictionary preferences*. Report on the A.S. Hornby Dictionary Research Award Project. https://www.hornby-trust.org.uk/projects#ASHDRADictionaryResearchAwards