Lexicom at 25: reflections on the changing world of lexicography and language technology

Michael Rundell[1], Miloš Jakubíček[1,2], Vojtěch Kovář[1,2], Ondřej Matuška[1], Michal Cukr[1]

[1]Lexical Computing, Czechia & United Kingdom
[2]Natural Language Processing Centre, Masaryk University, Brno, Czechia
michael.rundell@gmail.com, firstname.lastname@sketchengine.eu

Abstract

In this paper we show how the academic content and computational tools featured in Lexicom form a parallel history of the last 25 years of innovation in lexicography. Lexicom is a 5-day intensive workshop offering hands-on training in corpus-based dictionary creation, from collecting and annotating language data to publishing the final product. Since it was launched in 2001, by Sue Atkins, Adam Kilgarriff, and Michael Rundell, Lexicom has adapted (sometimes incrementally, sometimes substantially), to reflect ongoing developments in linguistic theory, corpus tools, and NLP. Lexicom's curriculum integrates theoretical grounding with practical tasks such as corpus analysis, regular expressions, word sense disambiguation, and definition-writing. It provides an introduction to all of the key components of dictionary-creation and to the current state of the art in our field. The lexicographic landscape has seen transformative changes during Lexicom's 25-year lifetime. In 2001, corpora were relatively small even for well-resourced languages and non-existent for others; querying tools were quite basic; and the end-product was almost invariably a printed book. We now use billion-word corpora and sophisticated software to produce mainly digital dictionaries. Lexicom has mirrored these shifts, most recently incorporating AI and large language models. Amid all these dramatic changes, some constants in the dictionary-making process remain, and Lexicom continues to serve as both a reflection of and a guide through this ongoing evolution.

Keywords: dictionary; lexicography; Lexicom workshop; NLP, Sketch Engine; Post-editing lexicography; Large Language Models; teaching lexicography

1. Introduction

In September of this year (2025), a Lexicom workshop was held in Bari (Italy), attended by 25 participants from 11 countries and four continents. This marks the 25th anniversary of a course which first ran at the University of Brighton (UK) in 2001. Lexicom is a five-day intensive workshop which – through a mix of lectures, discussion, and practical work – guides participants through the process of creating corpus-based dictionaries and similar resources, all the way from collecting and annotating language data to publishing the finished article.

Lexicom was set up by Sue Atkins, Adam Kilgarriff, and Michael Rundell at a pivotal moment. A golden age of lexicography in the UK was beginning to draw to a close. The 1980s and 1990s had seen a lexicographic boom for English, with the publication of numerous dictionaries of every type, many if not most of which of had been created

'from scratch': think of the Oxford Dictionary of English (first published in 1998); Collins' innovative range of bilinguals with English as a source language (and similar offerings from Oxford); and the 'Big Five' English learner's dictionaries[1] – the last of which, the Macmillan English Dictionary, was published just a few months after the first Lexicom. For those two decades, the UK was the dictionary world's centre of gravity, and the source of some of the most significant innovations in lexicographic practice. The Euralex association and its first conference were also launched in the UK during this period, as was the dictionary community's premier journal, the IJL. Paradoxically, the incipient slow decline of lexicographic activity in the UK coincided with rapid developments in computer technology and language engineering. These offered the potential for radical improvements in the way languages could be analysed and described, taking forward the corpus revolution that began with the COBUILD project in the 1980s. This burst of activity during the last two decades of the 20th century created a large cohort of skilled and well-trained lexicographers in the UK, but a question the Lexicom founders asked themselves was: where would the next generation of lexicographers get their training? The workshop was thus set up, inter alia, to address a gap in the market.

There were (and still are) several options for anyone with an academic interest in lexicography and especially lexicographic theory – the European Master in Lexicography (EMLex) being an obvious example. But, as was noted at the time, there is an important distinction 'between teaching people about lexicography, and training people to be lexicographers' (Rundell, 2001). Lexicom is unique in that it combines a grounding in all of the 'inputs' to dictionary-making (corpus creation and annotation, corpus-query languages, lexical semantics, neology, and so on) with practical exercises in lexicographic tasks (building corpora and analysing them, using regular expressions, word sense disambiguation, definition-writing, and much more). This hands-on element was a key feature of the course right from the outset. And while Lexicom does not claim to provide a complete training package for aspiring lexicographers, it does offer a useful and practical introduction to all of the main components of dictionary-creation.

The content of the course has changed each year – sometimes incrementally, sometimes very substantially – in response to developments in the field and in the wider technological and linguistic environment. So this is a good moment to reflect on changes over 25 years in the lexicographic process, and in the theoretical ideas and computational tools that support it.

2. The lexicographic landscape in 2001 – and today

The landscape has changed dramatically during the quarter-century of Lexicom's existence. It is easy to forget that many technologies we now take for granted were still in their infancy in 2001: the Web was relatively new; broadband Internet access was far from widespread (an AI summary on Google says that broadband 'became widely popular in the late 2000s, with significant growth … starting around 2008'); and mobile phones were used only for making calls or sending SMS messages (the first iPhone was launched in 2007).

[1] by the order of appearance, that means Oxford Advanced Learner's Dictionary (1948), Longman Dictionary of Contemporary English (1978), Collins COBUILD Advanced Learner's Dictionary (1987), Cambridge Advanced Learner's Dictionary (1995) and Macmillan English Dictionary for Advanced Learners (2002)

When the inaugural Lexicom was held at the University of Brighton in 2001, one of the challenges we faced was how to provide participants with corpus data and corpus-querying tools for use in practical tasks. To address this, we manually loaded a 10-million-word subset of the British National Corpus (BNC) onto each desktop machine in the university's computer lab (the full 100 million-word version would have been too large for the hard drives of the time), along with an early version of Wordsmith Tools (Scott, 1999) for generating concordances. The BNC was one of the largest corpora around at that time, and already in use as an evidence base for several UK-based dictionaries. Globally, however, the use of corpora as a basis for dictionaries was far from widespread. This was down to a combination of slow processing speeds, small storage space on personal computers, a paucity of data and NLP tools for most languages, and corpus-querying software which was, by today's standards, sluggish and primitive.

So in 2001, there could be no automatic assumption that every participant in the workshop would be familiar with the basics of corpus linguistics; in fact, for many attendees, this was their first experience of working with corpus data. Nor were there any off-the-shelf dictionary-writing packages at that time,[2] and most publishers used their own homegrown systems. And of course, the finished dictionary would in almost all cases be a printed book rather than an online resource. (For example, the first digital version of the Macmillan dictionary appeared on a CD-ROM in 2007.)

This was the environment into which Lexicom landed in 2001. Fast-forward to 2025, and the situation has changed out of all recognition. Corpus-based lexicography is the norm. The costs of corpus development (in time and money) are a fraction of what they were in 2001. Even low-resourced languages may boast corpora bigger than the original BNC; and for many of the world's 'larger' languages, multi-billion word corpora are commonplace. Corpus-querying software is capable of interrogating the largest corpora, and extracting information in multiple ways (not only from concordances) at very high speeds. A number of off-the-shelf software packages for creating, editing and publishing dictionary content are now widely available, including IDM's DPS, Tshwanelex, and Lexonomy. And the final product is more likely to be published online than in book form, opening up opportunities for dictionaries to provide new types of information. With corpus data and software tools being cloud-based, the Lexicom workshop is no longer confined (as it was for its first ten years or so) to university computer labs. Nowadays, Lexicom is a 'BYOD' event ('bring your own device') and can be held anywhere with a good Internet connection.

3. Who comes to Lexicom?

The first Lexicom workshop attracted 50 participants. (This was actually too large a group to allow for much one-on-one interaction, and in recent years we have put a cap of 25 on the number of attendees.) Among this initial cohort were several people who were (or have since become) well-known figures in the lexicographic community, including Anna Braasch, Krista Varantola, Gilles-Maurice de Schryver, Ilan Kernerman, Bolette Pedersen, and Sussi Olsen. The high number of participants in 2001 hinted at a large pent-up demand, and in fact we were concerned that we may have exhausted the pool of potential attendees in the very first year, and that Lexicom would not be repeated.

[2] This isn't strictly true: an early dictionary writing system called GestorLEX had been used since the mid-90s by some publishers (including Longman). Unfortunately, it only worked on IBM's OS/2 operating system, which did not survive long.

But the second iteration in 2002 attracted 39 attendees, and Lexicom has run annually since then, with the single exception of 2020, when the global pandemic put paid to face-to-face gatherings of all kinds. At the time of writing, there are almost 600 'graduates' from the annual Lexicom workshops in Europe, and well over 100 more from 'spin-off' versions – either regular workshops in non-European locations, or Lexicom-type courses customised for specific institutions. Participants have come from 71 countries, including Bhutan, Aruba, and the Maldives, but predominantly from Europe (with, incidentally, a grand total of one person from France). Typically, attendees will be working or aspiring lexicographers, lexicographic project managers, computer scientists and computational linguists, translators, experts in language teaching, and research students in relevant disciplines. The maps show (1) the European locations where Lexicom has been held; (2) the locations where spin-off editions of Lexicom have been held; and (3) the countries of origin of Lexicom participants.

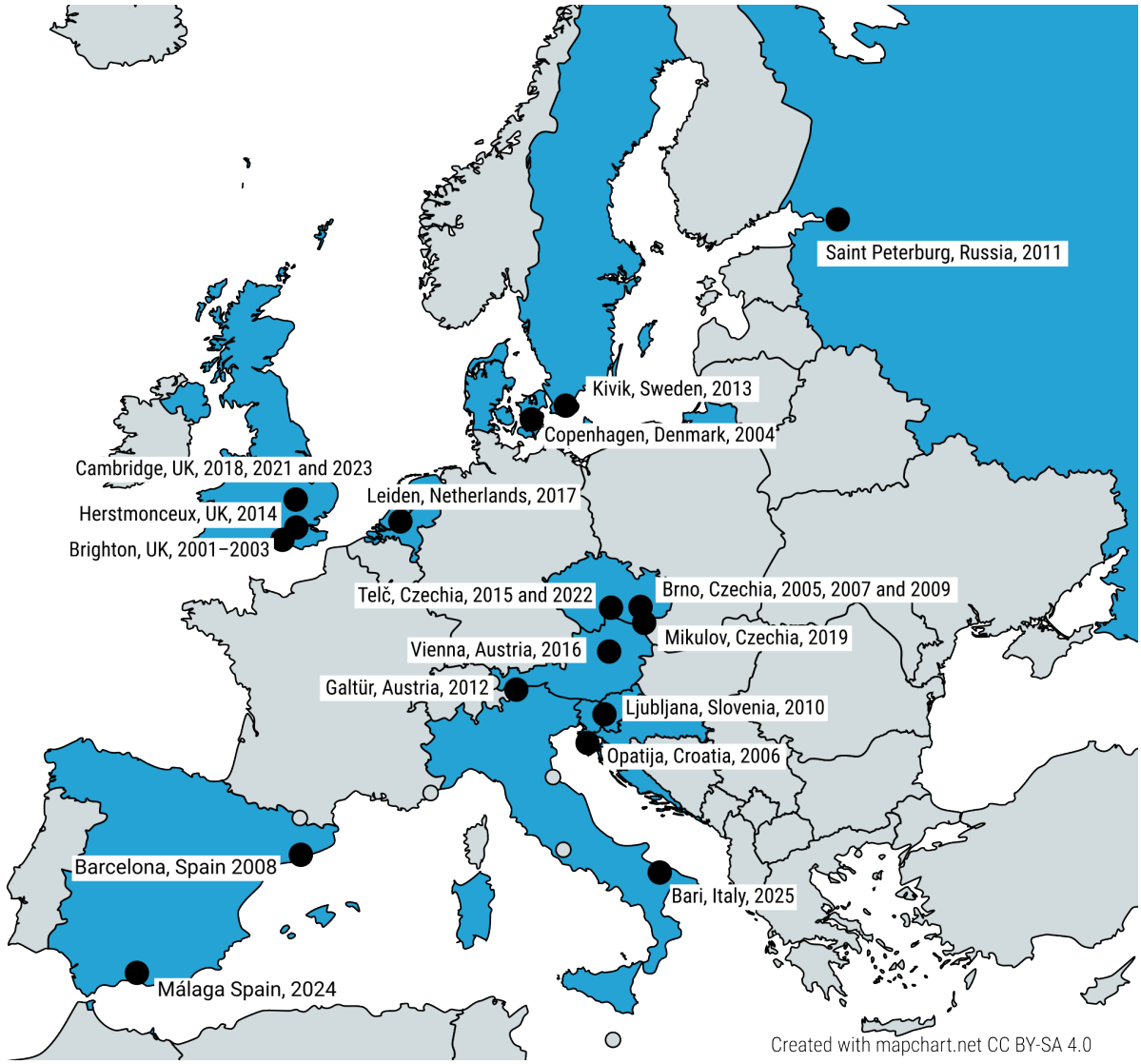

Figure 1: The European locations where Lexicom has been held

One of the most challenging, but also most rewarding, workshops was held in the interior of Western Australia, in the gold-mining town of Kalgoorlie in 2018. We had been invited

by the Goldfields Aboriginal Language Centre to run a workshop for a dozen or so linguists working with some of Australia's numerous indigenous languages. Not surprisingly, the biggest issue was the sparsity of data: the languages of the First Australians are predominantly oral. But we devised or adapted strategies for collecting language data (and transcribing archived recordings), and we were able to build small corpora for several languages. Of course, we were only scratching the surface: there are hundreds of Aboriginal languages, and many are on the brink of extinction. But it was a moving experience to watch the students using corpus tools, usually for the first time, to investigate the languages they were working with. It was one of those cases where we as teachers probably learned more from the participants than they learned from us.

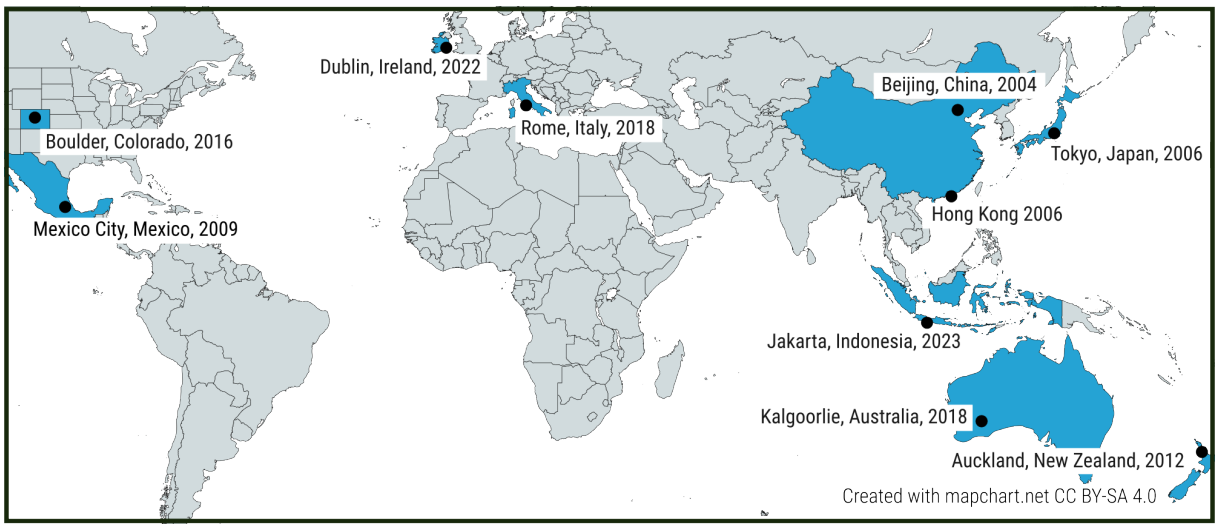

Figure 2: The locations where spin-off editions of Lexicom have been held

The kind of work the Kalgoorlie participants were engaged in highlights a broader trend. In the first decade or so of Lexicom, we had numerous participants from British or European dictionary publishers (public or private) working on standard monolingual or bilingual dictionaries for 'large', well-resourced languages. Recent years have seen a greater focus on the development of more specialised lexical resources: for people with disabilities (e.g. sign language dictionaries); for specialised domains (dictionaries of tourism, musical terms, etc.); for endangered or poorly-documented languages; or for specialised areas of the lexicon such as collocation or phraseology. Over the years, these changes in the profile of Lexicom attendees have tended to reflect shifts in the broader lexicographic landscape.

4. Changes in lexicographic technology since 2001

Aside from advances in the general technological environment (almost universal high-speed internet access, exponentially faster processing speeds, inexpensive data storage), all of the stages involved in creating a dictionary have been transformed through changes in the technologies we depend on. As noted above, very large corpora are now routinely available for most languages, due substantially to the development of NLP tools for automatically harvesting, cleaning, and annotating raw language data from the Web.

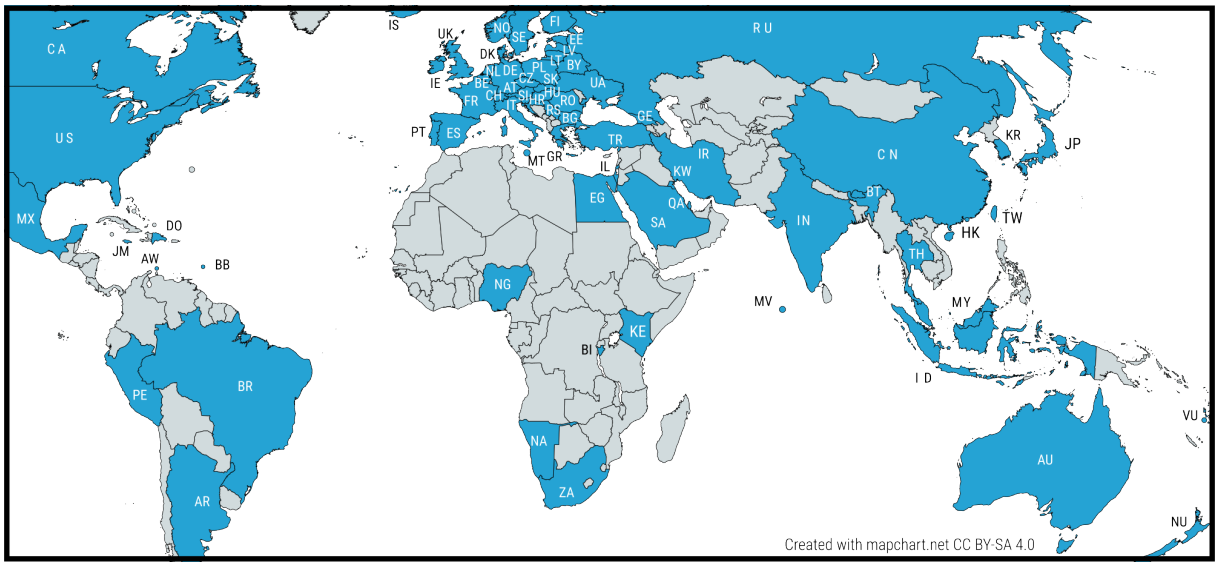

Figure 3: The countries of origin of all Lexicom participants.

Meanwhile the tools used by lexicographers for extracting relevant information from language data have seen dramatic improvements. At the turn of the millennium, lexicographers with access to corpora (by no means the majority at that time) would typically scan KWIC concordances in order to identify word senses and their defining features; gather information about a word's syntactic and collocational preferences; and find candidate example sentences. While the concordance remains a key part of the lexicographer's toolkit, it is no longer practical to 'manually' analyse tens of thousands of concordance lines when working with a large corpus. Consequently, lexical profiling software tools (such as Word Sketches) are a more typical starting point in the corpus-analysis process.

Sketch Engine, which combines concordancing, Word Sketches, and numerous other corpus-building and corpus-querying tools, is effectively a product of the Lexicom workshops. At the second Lexicom in 2002, a researcher from Masaryk University in Brno, Pavel Rychlý, gave a demonstration of his corpus management software, Manatee/Bonito (Rychlý, 2007). The standout feature of this system was its ability to generate concordances at speeds which no-one watching the demo had ever seen before. In the same workshop, course leader Adam Kilgarriff previewed an embryonic version of the Word Sketch, a tool for automatically identifying a word's key collocates and grouping them according to the grammatical relations in which they occur. This was the outcome of a research project with David Tugwell at the University of Brighton (Kilgarriff & Tugwell, 2001). Following the workshop, Kilgarriff and Rychlý embarked on a collaboration and set up a company, Lexical Computing, which to this day continues to develop these resources.

A significant technological trend during the Lexicom years – and one reflected in the workshop's changing syllabus – is the gradual move towards greater automation of the processes involved in creating a dictionary: from corpus building and headword-list development; through lexicographic tasks such as word sense induction, identifying suitable dictionary examples, and finding translation equivalents; to making the final dictionary available to users, and then maintaining it in its post-publication form. (For an overview, see Rundell & Kilgarriff (2011); Rundell (2023)) All these initiatives have come together in more recent years in the rollout of 'post-editing lexicography', an ambitious suite of

technologies which leverage advanced NLP techniques to automatically generate a draft corpus-driven dictionary (Jakubíček et al., 2018) (Rundell et al., 2020). This rough draft is then post-edited by human lexicographers, following a model which will be familiar to people working in the translation industry.

Another innovation has been the greater involvement of dictionary users in some stages of the lexicographic process (e.g. Klosa-Kückelhaus & Tiberius (2025), pp. 5–6), and a session on the different varieties of crowdsourcing was introduced to the Lexicom programme in 2018.

All of these technologies have developed incrementally over many years, gradually becoming more efficient and more reliable. The process has been evolutionary. A far more disruptive event occurred in late 2022, with the arrival of ChatGPT, the first of several Large Language Models (LLMs) to be released around that time.

For some, the implications are existential, heralding 'the end of lexicography' – which can be taken to mean either that lexicographers are redundant (if machines can produce dictionaries without them), or that dictionaries are redundant (if users' reference needs can be adequately met without dictionaries). According to this view, while post-editing lexicography represents a collaboration between humans and machines, LLMs replace humans altogether – if not now, then in due course.

Others, including the current Lexicom course trainers writing this paper, are more sceptical. We believe that while the way lexicographic content is presented to its end-users has changed significantly (and likely will continue changing), and alongside of that the notion of a dictionary as perceived by general public, all the various needs for getting information about how language works remain and it is hard to see a reason why they should vanish. Moreover, a number of studies have shown that LLMs perform impressively in some lexicographic tasks (its definition-writing skills, for example, are often as good as those of most humans), but are embarrassingly bad at others – notably word sense disambiguation – and seem to lack anything that could truly be described as intelligence. (See Jakubíček & Rundell (2023); de Schryver (2023) for a summary of recent work and opinions.) It is too early to form a definitive view, but what we do know is that AI tools can't be left out of the debate. Since the 2023 workshop, we have incorporated this topic into the Lexicom syllabus, not only in lectures but also in practical work, where participants can create dictionary content using LLMs and then evaluate it against 'traditional' dictionaries.

5. Changes in Lexicom's linguistic and lexicographic content

The NLP-derived technologies that underpin lexicography have developed rapidly during the lifetime of Lexicom, bringing transformative changes both in lexicographic practice and in the depth and breadth of coverage that dictionaries can offer. At the theoretical level, the pace of change has been less frenetic. Lectures on topics such as word senses and definitions have evolved quite slowly, and some of the material from the first Lexicom is still being used, with only minor modification, in current iterations of the workshop. (Appendix **??** shows the Programme for the first Lexicom workshop in 2001. Appendix **??** shows the Programme from the 2025 workshop.) We see it as a proof that most of the theory selected 25 years ago was selected appropriately as it stood the hardest test for any theory: the test of time.

Many of the lexicographic (as opposed to technical) elements of the course drew on lectures devised originally for the SALEX/Afrilex workshops held in South Africa in 1997 and 1998. These courses were led by Sue Atkins, Edmund Weiner of the OED, and Michael Rundell. They were set up by a group which included Penny Silva and Danie Prinsloo, following changes in South Africa's constitution post-apartheid. Nine of the country's African languages were added to English and Afrikaans as official national languages, and they each needed new dictionaries created using contemporary methods. Thus the goal of SALEX was 'the development of a cadre of lexicographers with the practical and intellectual skills to tackle every aspect of running a dictionary project' ((Silva, 1998), p. 283). This led to two two-week intensive courses, incorporating a mixture of lectures and practical tasks. The SALEX courses provided a basic template for Lexicom, and indeed for the guidebook written by Atkins & Rundell (2008).

The non-technical sections of Lexicom include some discussion of 'lexicographic theory' (Wiegand, Bergenholtz et al.), but with a degree of scepticism as to its practical usefulness. Linguistic theories, on the other hand, are covered in some depth where they have a clear relevance to what we do as lexicographers. These include references to Frame Semantics (Fillmore), Lexical Functions (Mel'čuk), the Generative Lexicon (Pustejovsky), Regular Polysemy (Apresajan), Prototype Theory (starting with the work of Eleanor Rosch), and the theoretical ideas of John Sinclair and Patrick Hanks. In every case, the emphasis is on demonstrating the links between the theory and its practical application to real-world lexicographic tasks.

6. Conclusions

It was always a goal of Lexicom to give lexicographers a better understanding of the technologies they depend on and of the potential of NLP research to improve lexicographic tools. At the same time, we want to enable those from a computational background to get a clearer view of the needs of lexicographers and how these might be better met. Over the years, the technical and NLP content of Lexicom has taken a more prominent role, and where two of the original three course leaders were lexicographers and one a computationalist, that position is now reversed. This reflects a change in the balance of the workload involved in creating a dictionary, and our ever-greater reliance on software tools. But far from deskilling lexicographers, emerging technologies have transformed the role of human editors. Instead of 'manually' analysing language data, their focus is now on evaluating and curating machine-generated content. In line with this shift, a typical model we apply in Lexicom for covering a subject such as word senses would be: first, a lecture on the linguistic and lexicographic aspects; then another on how NLP technologies can be applied to word sense induction (as well as an introduction to related approaches such as word embeddings); then a practical session where participants analyse corpus data to determine the 'dictionary senses' of specific polysemous words; and finally a look at the performance of LLMs on the same tasks.

Lexicom has always sought to introduce its students to the state of the art in lexicographic debate, practice, and resources. At the same time, we recognise that reality may look very different for some people in the field, and that (for various reasons) the methods and technologies they use may be far from cutting edge. Even now, it is not unheard of for a participant to tell us about a project they are working on where the evidence base is hand-gathered citations and the dictionary is being created in Microsoft Word.

The Lexicom workshops have mirrored the many changes in the lexicographic environment over the last 25 years. And in the widely divergent experiences of its participants, they also neatly illustrate William Gibson's well-known view that 'The future is already here, it's just not very evenly distributed'. We hope the Lexicom course will continue contributing to the education of next-generation lexicographers: the corpus-powered, AI-powered and well-educated ones.

Acknowledgments

We cordially thank our colleague Michal Cukr for collecting all course statistics and rendering the visualizations. We would also like to thank the anonymous reviewers, whose comments have helped us to improve the paper in a number of ways. Any remaining errors are ours alone.

Software

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

de Schryver, G.M. (2023). Generative AI and Lexicography: the Current State of the Art using ChatGPT. *International Journal of Lexicography*, 36(4), pp. 355 –387.

Jakubíček, M., Měchura, M., Kovář, V. & Rychlý, P. (2018). Practical Post-Editing Lexicography with Lexonomy and Sketch Engine. In *Proceedings of the XVIII EURALEX Congress*. Ljubljana, Slovenia: Ljubljana University Press, Faculty of Arts, pp. 65–67. URL http://anthology.aclweb.org/W16-2114.

Jakubíček, M. & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? *Electronic lexicography in the 21st century. Proceedings of the eLex 2023 conference*, pp. 518–533.

Kilgarriff, A. & Tugwell, D. (2001). WASP-Bench: an MT Lexicographers' Workstation Supporting State-of-the-art Lexical Disambiguation. In *Proceedings of Machine Translation Summit VIII*. Santiago de Compostela, Spain, pp. 187–190. URL https://kilgarriff.co.uk/publications.htm.

Klosa-Kückelhaus, A. & Tiberius, C. (2025). The Lexicographic Process Revisited. *International Journal of Lexicography*, 38(1), pp. 1–12.

Rundell, M. (2001). Teaching lexicography, or training lexicographers. *Kernerman Dictionary News*, 9, pp. 6–7. URL https://lexicala.com/wp-content/uploads/kdn9_2001_Teaching_lexicography_or_training_lexicographers_MR.pdf.

Rundell, M. (2023). Automating the creation of dictionaries: are we nearly there? In *Asialex 2023 Proceedings*. Seoul, South Korea, pp. 9–17.

Rundell, M., Jakubíček, M. & Kovář, V. (2020). Technology and English Dictionaries. In S. Ogilvie (ed.) *The Cambridge Companion to English Dictionaries*. Cambridge University Press, pp. 18–30.

Rundell, M. & Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end? In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (eds.) *A Taste for Corpora. A Tribute to Professor Sylviane Granger*. Benjamins, pp. 257–281.

Rychlý, P. (2007). Manatee/Bonito – A Modular Corpus Manager. In *First Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2007*. Brno: Masaryk University, pp. 65–70. URL https://nlp.fi.muni.cz/raslan/2007/.

Scott, M. (1999). *Wordsmith Tools Version 3*. Oxford: Oxford University Press.

Silva, P. (1998). Report on the SALEX '97 Lexicographical Training Course, 15–27 September 1997. *Lexikos*, 8, pp. 282–288.