www.nytimes.com/2018/02/23

Inside the OED: can the world's biggest dictionary survive the internet?
By Andrew Dickson

In February 2009, a Twitter user called @popelizbet issued an apparently historic challenge to someone called Colin: she asked if he could "mansplain" a concept to her. History has not recorded if he did, indeed, proceed to mansplain. But the lexicographer Bernadette Paton, who excavated this exchange last summer, believed it was the first time anyone had used the word in recorded form. "It's been deleted since, but we caught it," Paton told me, with quiet satisfaction.

In her office at Oxford University Press, Paton was drafting a brand new entry for the Oxford English Dictionary. Also in her in-tray when I visited were the millennial-tinged usage of "snowflake", which she had hunted down to a Christian text from 1983 ("You are a snowflake. There are no two of you alike"), and new shadings of the compound "self-made woman". Around 30,000 such items are on the OED master list; another 7,000 more pile up annually. "Everyone thinks we're very slow, but it's actually rather fast," Paton said. "Though admittedly a colleague did spend a year revising 'go'".

Spending 12 months tracing the history of a two-letter word seems dangerously close to folly. But the purpose of a historical dictionary such as the OED is to give such questions the solemnity they deserve. An Oxford lexicographer might need to snoop on Twitter spats from a decade ago; or they might have to piece together a painstaking biography of one of the oldest verbs in the language (the revised entry for "go" traces 537 separate senses over 1,000 years). "Well, we have to get things right," the dictionary's current chief editor, Michael Proffitt, told me.

At one level, few things are simpler than a dictionary: a list of the words people use or have used, with an explanation of what those words mean, or have meant. At the level that matters, though – the level that lexicographers fret and obsess about – few things could be more complex. Who used those words, where and when? How do you know? Which words do you include, and on what basis? How do you tease apart this sense from that? And what is "English" anyway?

In the case of a dictionary such as the OED – which claims to provide a "definitive" record of every single word in the language from 1000AD to the present day – the question is even larger: can a living language be comprehensively mapped, surveyed and described? Speaking to lexicographers makes one wary of using the word "literally", but a definitive dictionary is, literally, impossible. No sooner have you reached the summit of the mountain than it has expanded another hundred feet. Then you realise it's not even one mountain, but an interlocking series of ranges marching across the Earth. (In the age of "global English", the metaphor seems apt.)

Even so, the quest to capture "the meaning of everything" – as the writer Simon Winchester described it in his book on the history of the OED – has absorbed generations

of lexicographers, from the Victorian worthies who set up a "Committee to collect unregistered words in English" to the OED's first proper editor, the indefatigable James Murray, who spent 36 years shepherding the first edition towards publication (before it killed him). The dream of the perfect dictionary goes back to the Enlightenment notion that by classifying and regulating language one could – just perhaps – distil the essence of human thought. In 1747, in his "Plan" for the English dictionary that he was about to commence, Samuel Johnson declared he would create nothing less than "a dictionary by which the pronunciation of our language may be fixed, and its attainment facilitated; by which its purity may be preserved, its use ascertained, and its duration lengthened". English would not be merely listed in alphabetical order; it would be saved for eternity.

Ninety years after the first edition appeared, the OED – a distant, far bulkier descendant of Johnson's Dictionary – is currently embarked on a third edition, a goliath project that involves overhauling every entry (many of which have not been touched since the late-Victorian era) and adding at least some of those 30,000 missing words, as well as making the dictionary into a fully digital resource. This was originally meant to be completed in 2000, then 2005, then 2010. Since then, OUP has quietly dropped mentions of a date. How far had they got, I asked Proffitt. "About 48%," he replied.

The dictionary retains a quiet pride in the lexical lengths to which it will – indeed, must – go. Some time in the late 1980s, Proffitt's predecessor as chief editor, John Simpson, asked the poet Benjamin Zephaniah about the origins of the noun "skanking". Zephaniah decided that the only way to explain was to come to OED headquarters and do a private, one-on-one performance. Skanking duly went in, defined as "a style of West Indian dancing to reggae music, in which the body bends forward at the waist, and the knees are raised and the hands claw the air in time to the beat".

The tale touches something profound: in capturing a word, a sliver of lived experience can be observed and defined. If only you were able to catch all the words, perhaps you could define existence.

The first English dictionary-makers had no fantasies about capturing an entire culture. In contrast to languages such as Chinese and ancient Greek, where systematic, dictionary-like works have existed for millennia, the earliest English lexicons didn't begin to be assembled until the 16th century. They were piecemeal affairs, as befitted the language's mongrel inheritance – a jumbled stew of old Anglo-Germanic, Norse, Latin and Greek, and Norman French.

The language was perplexing enough, but in the mid-1500s it was getting ever more confusing, as political upheavals and colonial trade brought fresh waves of immigration, and with it a babel of recently "Englished" vocabulary: words such as "alcohol" (Arabic via Latin, c1543) and "abandonment" (French, c1593). Scientific and medical developments added to the chaos. In 1582, the schoolmaster Richard Mulcaster issued a frantic plea for someone to "gather all the wordes which we use in our English tung … into one dictionarie". Such a book would stabilise spelling, a source of violent disagreement. Also, there would finally be rules for "proper use".

In 1604, a clergyman named Robert Cawdrey attempted a stopgap solution: a slender book entitled A Table Alphabeticall. Aimed at "Ladies, gentlewomen and other unskillful persons", it listed approximately 2,500 "hard usuall words", less than 5% of the lexis in use at the time. Definitions were vague – "diet" is described as "manner of foode" – and there were no illustrative quotations, still less any attempt at etymology. A Table Alphabeticall was so far from being completist that there weren't even entries for the letter W.

Lexicographers kept trying to do better – and mostly kept failing. A new "word book" edited by John Bullokar appeared in 1616 (5,000 words); another by Henry Cockeram in 1623 (8,000 words and the first to call itself a "dictionary"); yet another by Thomas Blount in 1656 (11,000 words). But no one could seem to capture "all the wordes" in English, still less agree on what those words meant. The language was expanding more rapidly than ever. Where would you even start?

Comprehensive dictionaries had already been produced in French, Italian and Spanish; Britain's failure to get its house in order was becoming an international embarrassment. In 1664, the Royal Society formed a 22-person committee for "improving the English language", only to disband after a few meetings. In 1712, Jonathan Swift published a pamphlet on the subject, pouring scorn on sloppy usage and insisting that "some Method should be thought on for ascertaining and fixing our Language for ever" – arguing that English should not merely be exhaustively surveyed, but that its users should be forced to obey some rules. This task defeated everyone, too. It wasn't until 1746, when a consortium of publishers managed to convince Samuel Johnson to take on this "great and arduous post", that it seemed remotely likely to be completed.

Johnson's Dictionary, eventually finished in 1755, was a heroic achievement. He corralled 43,500-odd words – perhaps 80% of the language in use at the time. But in some eyes, not least the editor's, the book was also a heroic failure. In contrast to the jaunty Enlightenment optimism of his 1747 Plan, with its talk of "fixing" and "preservation", the preface to the published Dictionary is a work of chastened realism. Johnson explains that the idea of taming a fast-evolving creature such as the English language is not only impossible, but risible:

"We laugh at the elixir that promises to prolong life to a thousand years; and with equal justice may the lexicographer be derided, who being able to produce no example of a nation that has preserved their words and phrases from mutability, shall imagine that his dictionary can embalm his language, and secure it from corruption and decay."

Much as lexicographers might fantasise about capturing and fixing meaning – as Johnson had once fantasised – a living language will always outrun them.

Still, the dream lingered. What if one could get to 100% – lassoing the whole of English, from the beginning of written time to the present day? Numerous revisions or rivals to Johnson were proposed, though few were actually created. After a Connecticut

schoolteacher named Noah Webster published his American Dictionary of the English Language in 1828 (70,000 entries), British pride was once again at stake.

In November 1857, the members of the London Philological Society convened to hear a paper by Richard Chenevix Trench, the dean of Westminster, entitled "On Some Deficiencies in our English Dictionaries". It was a bombshell: Trench argued that British word banks were so unreliable that the slate needed to be wiped clean. In their place, he outlined "a true idea of a Dictionary". This Platonic resource should be compiled on scholarly historical lines, mining deep into the caverns of the language for ancient etymology. It should describe rather than prescribe, casting an impartial eye on everything from Anglo-Saxon monosyllables to the latest technical jargon (though Trench drew the line at regional dialect). Most of all, it should be comprehensive, honouring what Trench called – glancing jealously at Germany, where the brothers Grimm had recently started work on a Deutsches Wörterbuch – "our native tongue".

The quest to capture the language in its entirety may have been centuries old, but, like a great railway line or bridge, this new dictionary would be thoroughly Victorian: scientific, audacious, epic and hugely expensive. Building it was a patriotic duty, Trench insisted: "A dictionary is a historical monument, the history of a nation".

For the first two decades, the New English Dictionary, as it was called, looked as if it would go the way of so many previous projects. The first editor died a year in, leaving chaos in his wake. The second had more energy for young women, socialism, folksong and cycling. Only after it was taken over by Oxford University Press, who in 1879 were persuaded to appoint a little-known Scottish schoolteacher and philologist called James Murray as chief editor, did things begin to move.

Murray's masterstroke was to put out an "appeal" in newspapers and library books for volunteer readers to search for quotations, which would illustrate the ways words changed over time – a "corpus" of data that would make the dictionary as accurate as possible. More than 2,000 enthusiasts from across the world and all walks of life assembled some 5m quotations to feed Murray's team of lexicographers as they churned through the alphabet, defining words as they went. Even when it became evident that it would all take far, far longer than scheduled – after five years they were still halfway through the letter A – Murray kept the dictionary going. "It would have been impossible without him," says the lexicographer and OED historian Peter Gilliver.

The first part was published in 1884, A to Ant, and instalments emerged at regular intervals for the next 40-odd years. Although Murray died in 1915 – somewhere between "Turndun" and "Tzirid" – the machine churned on. In 1928, the finished dictionary was eventually published: some 414,800 headwords and phrases in 10 volumes, each with a definition, etymology and 1.8m quotations tracking usage over time.

It was one of the largest books ever made, in any language: had you laid the metal type used end to end, it would have stretched from London to Manchester. Sixty years late it

may have been, but the publisher made the most of the achievement, trumpeting that "the Oxford Dictionary is the supreme authority, and without a rival".

Yet if you knew where to look, its flaws were only too obvious. By the time it was published in 1928, this Victorian leviathan was already hopelessly out of date. The A-C entries were compiled nearly 50 years earlier; others relied on scholarship that had long been surpassed, especially in technology and science. In-house, it was admitted that the second half of the alphabet (M-Z) was stronger than the first (A-L); the letter E was regarded as especially weak. Among other eccentricities, Murray had taken against "marzipan", preferring to spell it "marchpane", and decreed that the adjective "African" should not be included, on the basis that it was not really a word. "American", however, was, for reasons that reveal much about the dictionary's lofty Anglocentric worldview.

The only solution was to patch it up. The first Supplement to the OED came out in 1933, compiling new words that editors had noted in the interim, as well as original omissions. Supplements to that Supplement were begun in 1957, eventually appearing in four instalments between 1972 and 1986 – some 69,300 extra items in all. Yet it was a losing battle, or a specialised form of Zeno's paradox: the closer that OED lexicographers got to the finish line, the more distant that finish line seemed to be.

At the same time, the ground beneath their feet was beginning to give way. By the late 1960s, a computer-led approach known as "corpus linguistics" was forcing lexicographers to re-examine their deepest assumptions about the way language operates. Instead of making dictionaries the old-fashioned way – working from pre-existing lists of words/definitions, and searching for evidence that a word means what you think it does – corpus linguistics turns the process on its head: you use digital technology to hoover up language as real people write and speak it, and make dictionaries from that. The first modern corpus, the Brown Corpus of Standard American English, was compiled in 1964 and included 1m words, sampled from 500 texts including romance novels, religious tracts and books of "popular lore" – contemporary, everyday sources that dictionary-makers had barely consulted, and which it had never been possible to examine en masse. The general-language corpora that provide raw material for today's dictionaries contain tens of billions of words, a database beyond the wildest imaginings of lexicographers even a generation ago.

There are no limits to the corpora that can be constructed: at a corpus linguistics conference in Birmingham last year, I watched researchers eavesdrop on college-age Twitter users (emojis have long since made "laughter forms" such as LOL and ROFL redundant, apparently) and comb through English judges' sentencing remarks for evidence of gender bias (all too present).

For lexicographers, what's really thrilling about corpus linguistics is the way it lets you spy on language in the wild. Collating the phrases in which a word occurs enables you to unravel different shades of meaning. Observing how a word is "misused" hints that its centre of gravity might be shifting. Comparing representative corpora lets you see, for example, how often Trump supporters deploy a noun such as "liberty", and how

differently the word is used in the Black Lives Matter movement. “It’s completely changed what we do,” the lexicographer Michael Rundell told me. “It’s very bottom-up. You have to rethink almost everything.”

But while other dictionary publishers leapt on corpus linguistics, OED editors stuck to what they knew, resisting computerisation and relying on quotation slips and researchers in university libraries. In the 1970s and 80s there was little thought of overhauling this grandest of historical dictionaries, let alone keeping it up to date: it was as much as anyone could do to plug the original holes. When the OED’s second edition was published in March 1989 – 20 volumes, containing 291,500 entries and 2.4m quotations – there were complaints that this wasn’t really a new edition at all, just a nicely typeset amalgam of the old ones. The entry for “computer” defined it as “a calculating-machine; esp an automatic electronic device for performing mathematical or logical operations”. It was illustrated by a quotation from a 1897 journal.

By astonishing coincidence, another earthquake, far bigger, struck the very same month that OED2 appeared in print: a proposal by an English computer scientist named Tim Berners-Lee for “a large hypertext database with typed links”. The world wide web, as it came to be called (OED dates the phrase to 1990), offered a shining path to the lexicographical future. Databases could be shared, and connected to one another; whole libraries of books could be scanned and their contents made searchable. The sum of human text was starting to become available to anyone with a computer and a modem.

The possibilities were dizzying. In a 1989 article in the New Yorker, an OUP executive said, with a shiver of excitement, that if the dictionary could incorporate corpus linguistics resources properly, something special could be achieved: “a Platonic concept – the ideal database”. It was the same ideal laid out by Richard Chevenix Trench 132 years before: the English language over a thousand or more years, every single word of it, brought to light.

The fact that so much text is now available online has been the most cataclysmic change. Words that would previously have been spoken are now typed on social media. Lexicographers of slang have long dreamed of being able to track variant forms “down to the level, say, of an individual London tower block”, says the slang expert and OED consultant Jonathon Green; now, via Facebook or Instagram, this might actually be possible. Lexicographers can be present almost at the moment of word-birth: where previously a coinage such as “mansplain” would have had to find its way into a durable printed record, which a researcher could use as evidence of its existence, it is now available near-instantly to anyone.

Anyone, and anywhere – when the OED was first dreamed up in the 1850s, English was a language of the British Isles, parts of North America, and a scattering of colonies. These days, nearly a quarter of the world’s population, 1.5bn people, speak some English, mostly as a second language – except, of course, that it isn’t one language. There are myriad regional variants, from the patois spoken in the West Indies and Pidgin forms of West Africa to a brood of compound offspring – Wenglish (Welsh English), Indlish or

Hinglish (Indian/Hindi English), and the "Chinglish" of Hong Kong and Macau. All of these Englishes are more visible now than ever, each cross-fertilising others at greater and greater speed.

"The circle of the English language has a well-defined centre but no discernible circumference," James Murray once wrote, but modern lexicographers beg to differ. Instead of one centre, there are many intersecting subgroups, each using a variety of Englishes, inflected by geographical background or heritage, values, other languages, and an almost incalculable number of variables. And the circumference is expanding faster than ever. If OED lexicographers are right that around 7,000 new English words surface annually – a mixture of brand-new coinages and words the dictionary has missed – then in the time you've been reading this, perhaps two more words have come into being.

Most people, of course, now never go near a dictionary, but simply type phrases into Wikipedia (used more often as a dictionary than an encyclopedia, research suggests) or rely on Google, which – through a deal with Oxford Dictionaries – offers thumbnail definitions, audio recordings of pronunciations, etymology, a graph of usage over time and translation facilities. If you want to know what a word means, you can just yell something at Siri or Alexa.

Dictionaries have been far too slow to adjust, argues Jane Solomon of Dictionary.com. "Information-retrieval is changing so fast," she said. "Why don't dictionaries respond intelligently to the semantic or user context, like figuring out that you're searching for food words, and give you related vocabulary or recipes?" And not just words: "I'd love to include emojis; people are so creative with them. They've become a whole separate language. People sometimes need explanation; if you send your daughter the eggplant emoji, she might think that's weird."

Some have dared to dream even bigger than polysemous aubergines. One is a computer professor at the Sapienza University of Rome called Roberto Navigli, who in 2013 soft-launched a site called Babelnet, which aims to be the dictionary to beat all dictionaries – in part by not really being a dictionary at all. Described as a "semantic network" that pulls together 15 existing resources including Wikipedia, Wiktionary and Microsoft Terminology, it aims to create a comprehensive, hierarchical root map of not just English but of 271 languages simultaneously, making it the largest lexicon/encyclopedia/thesaurus/reference work on the web. Navigli told me that his real aim was to use "semantic technology" to enable the holy grail for software engineers everywhere: autonomous machine-reading of text. "This is the dream, right?" he said. "The machine that can read text and understand everything we say."

Machines already understand a lot, of course. Some have talked of "culturomics", a form of computational lexicology that uses corpus tools to analyse and forecast trends in human behaviour. A 31-month study of Twitter tried to measure the shifting sentiments of the British population about austerity, and there is even a claim – somewhat disputed – that a "passively crowd-sourced" study of global media could have foretold the Arab spring. At least on a large scale, computers, and the information giants who own and

lease the data, may be able to comprehend language better than we comprehend it ourselves.

For lexicographers and Google alike, one linguistic frontier remains stubbornly inaccessible. Whereas it's now easy to assemble written-text corpora and open a window on how language functions in a particular environment, doing so for spoken language has always been far harder. The reason is obvious: recording speech, then transcribing it and creating a usable database, is both time-consuming and hugely expensive. Speech corpora do exist, but are notoriously small and unrepresentative (it's easy to work with court transcripts; far harder to eavesdrop on what lawyers say down the pub).

For lexicographers, speech is the most precious resource of all, and the most elusive. If you could capture large samples of it – people speaking in every context imaginable, from playgrounds to office canteens to supermarkets – you could monitor even more accurately how we use language, day to day. "If we cracked the technology for transcribing normal conversations," Michael Rundell said, "it really would be a game-changer."

For OED's editors, this world is both exhilarating and, one senses, mildly overwhelming. The digital era has enabled Oxford lexicographers to run dragnets deeper and deeper through the language, but it has also threatened to capsize the operation. When you're making a historical dictionary and are required to check each and every resource, then recheck those resources when, say, a corpus of handwritten 17th-century letters comes on stream, the problem of keeping the dictionary up to date expands to even more nightmarish proportions. Adding to that dictionary to accommodate new words – themselves visible in greater numbers than ever before, mutating ever-faster – increases the nightmare exponentially. "In the early years of digital, we were a little out of control," Peter Gilliver told me. "It's never-ending," one OED lexicographer agreed. "You can feel like you're falling into the wormhole."

Adding to the challenge is a story that has become wearily familiar: while more people are consulting dictionary-like resources than ever, almost no one wants to shell out. Sales of hard-copy dictionaries have collapsed, far more calamitously than in other sectors. (OUP refused to give me figures, citing "commercial sensitivities". "I don't think you'll get any publisher to fess up about this," Michael Rundell told me.) While reference publishers amalgamate or go to the wall, information giants such as Google and Apple get fat by using our own search terms to sell us stuff. If you can get a definition by holding your thumb over a word on your smartphone, why bother picking up a book?

"Go to a dictionary conference these days and you see scared-looking people," Rundell said. Although he trained as a lexicographer, he now mainly works as a consultant, advising publishers on how to use corpus-based resources. "It used to be a career," he went on. "But there just aren't the jobs there were 30 years ago." He pointed to his shelves, which were strikingly bare. "But then I'm not sentimental about print; I gave most of my dictionaries away."

Even if the infrastructure around lexicography has fallen away or been remade entirely, some things stay pleasingly consistent. Every lexicographer I spoke to made clear their distaste for "word-lovers", who in the dictionary world are regarded as the type of person liable to scrawl "fewer" on to supermarket signs reading "10 items or less", or recite "antidisestablishmentarianism" to anyone who will listen. The normally genial John Simpson writes crisply that "I take the hardline view that language is not there to be 'enjoyed'"; instead, it is there to be used.

But love is, most grudgingly admit, what draws people to spend their lives sifting and analysing language. It takes a particular sort of human to be a "word detective": something between a linguistics academic, an archival historian, a journalist and an old-fashioned gumshoe. Though hardly without its tensions – corpus linguists versus old-school dictionary-makers, stats nerds versus scholarly etymologists – lexicography seems to be one specialist profession with a lingering sense of common purpose: us against that ever-expanding, multi-headed hydra, the English language. "It is pretty obsessive-compulsive," Jane Solomon said.

The idea of making a perfect linguistic resource was one most lexicographers knew was folly, she continued. "I've learned too much about past dictionaries to have that as a personal goal." But then, part of the thrill of being a lexicographer is knowing that the work will never be done. English is always metamorphosing, mutating, evolving; its restless dynamism is what makes it so absorbing. "It's always on the move," said Solomon. "You have to love that."

There are other joys, too: the thrill of catching a new sense, or crafting a definition that feels, if not perfect, at least right. "It sounds cheesy, but it can be like poetry," Michael Rundell reflected. "Making a dictionary is as much an art as a craft."

Despite his pessimism about the industry, he talked with real excitement about a project he was about to join, working with experts from the Goldfield Aboriginal Language Centre on indigenous Australian languages, scantily covered by lexicographers. "Dictionaries can make a genuine difference," he said. "They give power to languages that might have had very little power in the past; they can help preserve and share it. I really believe that."

Throughout it all, OED churns on, attempting to be ever so slightly more complete today than it was yesterday or the day before. The dictionary team now prefer to refer to it as a "moving document". Words are only added; they are never deleted. When I suggested to Michael Proffitt that it resembled a proud but leaky Victorian warship whose crew were trying to keep out the leaks and simultaneously keep it on course, he looked phlegmatic. "I used to say it was like painting the Forth bridge, never-ending. But then they stopped – a new kind of paint, I think." He paused. "Now it's just us."

These days OED issues online updates four times a year; though it has not officially abandoned the idea of another print edition, that idea is fading. Seven months after I first asked how far they had got into OED3, I enquired again; the needle had crept up to

48.7%. “We are going to get it done,” Proffitt insisted, though as I departed Oxford, I thought James Murray might have raised a thin smile at that. If the update does indeed take until 2037, it will rival the 49 years it took the original OED to be created, whereupon it will presumably need overhauling all over again.

A few days ago, I emailed to see if “mansplain” had finally reached the OED. It had, but there was a snag – further research had pushed the word back a crucial six months, from February 2009 to August 2008. Then, no sooner had Paton’s entry gone live in January than someone emailed to point out that even this was inaccurate: they had spotted “mansplain” on a May 2008 blog post, just a month after the writer Rebecca Solnit had published her influential essay Men Explain Things to Me. The updated definition, Proffitt assured me, will be available as soon as possible.