The Coming Boom in English Lexicography –
Reconsidered (Part Two)

Charles M. Levine

Recap. In July 2001, I optimistically crawled out on a limb in these pages talking about an imminent boom in English lexicography.[1] Then in the July 2002 issue, my good friend and colleague Joseph Esposito countered that traditional, legacy dictionary publishers, like OUP and Merriam-Webster, "will muddle along, with growth becoming harder to come by except at the expense of their smaller and declining rivals; eventually they will stop publishing for broad markets altogether and the remaining activity will be to focus on the scraps Microsoft leaves on the floor."[2]

As a graduate in the history and philosophy of science, I have always been amazed at the uncanny ability of Esposito, a literature graduate, to anticipate important trends in technology. Possibly I get too preoccupied chasing down such details as how much, when? For example, how much will online dictionary searches replace print look-ups in the next three, five, and ten years? Twenty percent or 50 percent within ten years? But, as Esposito wrote (in a more general point about the future of traditional dictionary publishing), "Who knows?…In the absence of growth, the old [dictionary] business will be strained for capital, which will beget smaller investments, which will in turn hasten the decline."

Recently in the States. Certainly during the past two years, American lexicography has shown many of the strains Esposito wrote about. One could say these wounds were self-inflicted through corporate ownership dramas – but no matter, the

Charles M. Levine was recently the vice-president and publisher of Random House Reference Division, and has held a number of senior executive positions in digital and print content management in Asia and the USA. He is currently consulting, agenting, and packaging for a number of publishing clients.
charlev@att.net

cause of lexicography in America was not advanced, and possibly permanently hurt.

For example, Random House Webster dismantled its lexicography staff, as its relatively new Bertelsmann-owned parent grew tired of dictionary making. *Webster's New World*, sold to Hungry Minds (formerly, IDG Books), was off-loaded to John Wiley. *American Heritage*, as part of Houghton Mifflin, went to Vivendi, which then sold it to a private American-led investment group (headed by Bain Capital and Thomas H. Lee). The new *Microsoft Encarta* print dictionaries (created by Bloomsbury in the UK and distributed by St. Martin's in the US), after creating a splash of publicity, retreated in sales at retail. And the recent US$55 *New Oxford American Dictionary*—by all appearances an excellent product—assumed a dignified, but quiet, presence on bookshelves.

Only Merriam-Webster seems to have gained market share in America at the expense of its smaller competitors, who were hobbled by corporate problems. Currently shipping the new eleventh edition of its flagship *Collegiate Dictionary* (bundled for the first time with an electronic version and a free introductory online subscription – all for US$25.95), M-W recently boasted a 17 percent increase in dictionary sales, while their website bustled with more than 150,000 daily visits. The private investor (Jacqui Safra) who now owns Merriam, however, in a possible sign of impatience, recently brought in an outside CEO (G. Macomber) to find ways to grow the business more quickly. We will have to stay tuned for what develops at America's leading dictionary publisher.

In the aggregate, though, the American print dictionary market seems to stay at about the same size, year after year, even as online look-ups are apparently booming.[3] Is the resilience of print dictionaries in America a good sign? Is overall dictionary use (counting both print and digital look-ups) increasing? Possibly, yes to both questions. If I unscientifically use myself as an example, I now routinely consult two online dictionaries in addition to printed standbys – the OED online (accessible for free as a member of the Quality Paperback Bookclub[4]) and the faithful friend I once published, the *Random House Webster's Unabridged Dictionary* on CD-ROM. If a word still perplexes me, I search Google[5] or visit well-developed reference sites like Webopedia[6] (for technical terms). Because of the availability of multiple sources in both print and online, I am now much more dictionary literate – as are, I extrapolate, other serious word users.

Another noteworthy development in the States is the long-overdue progress toward creating the first comprehensive American National Corpus (ANC) – the first 10 million words of which are now being made available. My friend and colleague Wendalyn Nichols says more about this elsewhere in this newsletter[7].

The Bigger Picture. Looking beyond American shores and around the world, lexicography seems very much alive and well, if not booming to my optimistic drumbeat. The continued use and exploration of corpora and the vigorous linguistic research into world Englishes are two important signs of continued vigor. While at Random House Webster, I helped initiate—with the assistance of Nichols and others—an all-too-brief foray into creating entirely American-bred ESL/EFL dictionaries, partnering with publishers like FLTRP (Foreign Language Teaching and Research Press[8]) in Beijing, under the innovative leadership of Li Pengyi[9]. (Houghton Mifflin created the *American Heritage English as a Second Language Dictionary* in 1997 and revised it in 2002, primarily I believe to reach American schools and colleges – without strong marketing internationally.[10]) Except for these efforts the major American dictionary companies still appear blissfully lackadaisical about the potential of the global ELT market – which is probably the single largest area for growth in the English-dictionary business.

Maybe it would be more accurate to say that the global reach and penetration of English—especially as reflected on the WWW—will keep linguists and lexicographers busier for some time to come analyzing and mediating exchanges in which English is the lingua franca, and helping build the next iteration of the WWW, called the Semantic Web.

In the absence of an American corpus, for several years I have relied on the Web as a surrogate. For example, in a pop-reference book on Yiddish I co-created, playfully entitled the *Meshuggenary*[11], the Web was the best source of finding current uses in English of Yiddish-origin words and phrases (see the sidebar).

I am grateful to Michael Rundell, editor-in-chief of the *Macmillan English Dictionary for Advanced Learners*[12], for alerting me to the forthcoming special issue of *Computational Linguistics*[13], edited by Adam Kilgarriff and Greg Grefenstette, which will be devoted to the question of the Web as a corpus. Rundell notes that there are "many computational linguists who are beginning to see the Web as

the only corpus worth looking at (well, maybe I exaggerate somewhat), and as the solution to the long-running problem of 'data-sparseness'."

Of course, as Kilgarriff and Grefenstette point out, search engines like Google still have significant limitations as lexicographic tools—for example, in giving too much weight to words in the titles and headings of Web pages, and in missing the vast volume of material now hidden from the eyes of search engines behind "fee walls," in archives that charge to retrieve documents—but nonetheless it is clear that the Web presents lexicographers with a whole new set of opportunities to research current language use and should be considered a valid linguistic corpus.

Rundell also pointed out the heating up of discussions in the UK and Europe about English as a lingua franca – a development that looms so large throughout the world that it can still actually seem invisible to many native (English-centric) speakers. I noted in my first installment that an estimated 80 percent of Web pages are written in English (though that percentage may actually decrease somewhat over time; for example, Internet Explorer now supports the use of dozens of special scripts of the world's languages).

Somewhat hidden to those outside of Europe is the growing role that English plays (and a controversial one at that) in the affairs of the European Union. In 1970, about 60 percent of all documents coming out of Brussels were written in French, few if any in English. By 1997, English (45 percent) had surpassed French (40 percent) as the most frequently used official language.[14] (Is it less CNN, MTV, and MacDonald's that so haunts the French than ELF, English as a Lingua Franca? As language buffs, we can be sympathetic about the potential decline of any vibrant language.) The Semantic Web should further accelerate the use, promulgation, and importance of ELF. And by providing even richer data for research, the Semantic Web should also accelerate the business of global linguistics and lexicography.

The key step in building the Semantic Web will be the addition of metadata URIs—Universal Resource Identifiers—that "define or specify an entity, not necessarily by naming its location on the Web."[15] Put simplistically, the Semantic Web will establish protocols to identify types of content on each Web page – in ways to make the content elements computer readable and useable. As Tim Berners-Lee, who laid the groundwork for the WWW, writes:

The Semantic Web, in naming every concept simply by a URI, lets anyone express new concepts that they invent with minimal effort. Its unifying logical language will enable these concepts to be progressively linked into a universal Web. This structure will open up the knowledge and workings of humankind to meaningful analysis by software agents, providing a new class of tools by which we can live, work, and learn together.[16]

I am obviously skipping lightly over a number of important new works-in-progress that will profoundly affect linguistic and lexicographic research in the coming years. (Searching on Google for the phrase "Semantic Web" plus "lexicographer or lexicography," restricted to English pages cached in the past year, yielded about 850 results, with a number of fascinating leads to recent papers and conferences. Adding "linguist or linguistics" to the search criteria increases the results to 4000.[17]) Work is just gearing up, and while I bemoan not having more hard data and numbers, my instincts tell me that lexicography and linguistics are on the verge of a revolution as a result – though, sadly for me, much of this new linguistic and lexicographic innovation may take place outside of America, even ironically as American-English is the driving force behind the increasing global use of English.

Postlude
Some interwoven comments follow, from those whose help in writing this article has been most welcome and appreciated.

Rundell: There's definitely a big growth in corpus development worldwide (especially but by no means exclusively for use in dictionary making) – sometimes, it seems, almost anywhere but the US. The big Japanese publishers like Kenkyusha and Shogakukan are all partners in the ANC consortium, but also busy with corpus development of their own. There is, for example, a 100-million-word Corpus of Professional English under development in Japan.[18]

Levine: All this corpus work is immensely exciting, and it is going to be interesting to see how it will influence the look and feel of future native-speaker dictionaries of English as well.

Instant Yinglish – Google's Top Hits
(adapted from the *Meshuggenary*)

Searching on google.com gives a good indication of which Yinglish words are most frequently used today. **Glitch** and **kosher** top the charts, way ahead of all the rest; while even such well-known Yinglish words as **nosh**, **knish**, **schnoz**, **schmuck** and **gonnif** seem to have fallen behind in popularity, if one accepts Google's results. With the following list, you'll be instantly up and running in Yinglish.

glitch Slip-up; bug in the system. [232,000 hits]

kosher Legit, on the up-and-up; ritually clean. [222,000]

bagel The doughnut-shaped bread of champions. [145,000]

maven Expert; pundit; smart aleck. [70,800]

yid Jew, pronounced <yeed>. (But use with care: in U.S. slang, pronounced with a short i (as in bid), it is very disparaging.) [62,800]

klezmer Lively, heart-tugging Yiddish folk music. [46,800]

mensch Decent, trustworthy person. [42,600]

tush Backside; rear end. [39,500]

schlock Cheap or shoddy goods; junk. [39,300]

klutz Clumsy, inept person; blockhead. [39,000]

schmooze To chat or gossip; by extension, to network. [38,100]

chutzpah Impudence; moxie; cojones. [32,700]

The above results were derived from searching about 80 Yiddish-origin words that are now accepted in standard American English and would appear in up-to-date larger or unabridged dictionaries. The search was restricted to English Web pages, searching on word clusters such as [glitch glitsh glitchy], to take into account alternative spellings and closely related uses.

Kernerman: The implications may be more far-reaching and actually concern other languages too, not just English. For example, expect a dramatically growing demand for bilingual dictionaries of "unorthodox language-pairs", including so-called non-major languages, be it Icelandic/Korean, Latvian/Slovene, etc.

Levine: One of the key developments that I hope to see is the sharp increase in the quality of these dictionaries covering odd bilingual couples.

Esposito: I see a rhetorical error in the paper – rhetorical, not substantive. You are using my paper as a pushing-off point, which is fine: I have been a straw man before. But the contrast is imprecise: you are referring to lexicography, I to dictionary-makers. Lexicography is bound to grow. The current crop of dictionary companies can't grow. Apples and oranges.

Levine: There actually may be two lacunae in what I write: (1) The boom may be more in computational linguistics than in lexicography as such, especially when it comes to the Semantic Web. Separately, I learned, for example, that many students who major in linguistics go on to careers in software. (2) I am implying if not stating that because of ELF, you could grow the dictionary-making business; but I fudge about addressing the key question – whether you could make it a "growth" business. It's somewhat like talking about the U.S. economy. If it grows at only 1-2% a year, it would still be considered an investment crisis. You are saying you doubt dictionary making as a business can grow at all and will decline.

Esposito: I think you can grow any business that adds value beyond the default value of a bundled Microsoft dictionary, and learners' dictionaries add value.

Levine: But, I do not fully address the

O brave new worldictionaries

Ilan J. Kernerman

Miranda   O, wonder !
How many goodly creatures are there here !
How beauteous mankind is ! O brave new world,
That has such people in 't !
Prospero   'T is new to thee.
(W. Shakespeare, *The Tempest*, Act V. Scene I.)

When I began working in dictionaries in the early 1990s, our prime concern was the advent of English as a lingua franca—ELF (also here by Levine)—and its forthcoming consequences, for everyone to now learn and use worldwide, in equilibrium with their native language. This trend has indeed evolved, with critical impacts and interactions of various sorts.

Meantime, while we were snoozing, within that arguable globalization process—disseminating communication and information, grinding all into ubiquitous uniformity and mediocrity—other languages have also been awakening to each other and to themselves. Although ELF is champion—simultaneously and complementarily—it has become necessary and easier to create dictionaries for unorthodox language pairs, as well as to reach and explore—and sometimes even safeguard and enhance—any language still spoken.

With growing direct contact between languages not involving English, there are more trilingual and multilingual persons who want bilingual dictionaries without English, or dictionaries with two or more languages and ELF as an underlying bridge. Their quality might be painstaking to start with, but improvement usually follows suit. One way or another, soon you will be able to get any kind of dictionary, and via modern magic—such as Wi-Fi, broadband, cellphone—virtually anyhow, anywhere, anytime.

Some forecasts—such as teasingly or skeptically by Esposito—warn against the ill effects of such bliss on the traditional business of dictionary making. Yet, life forever intermingles so-called good with bad, bad with good, counter-running contradictions in cohabitation, bringing all together and centralizing, whilst breaking farther apart into atoms, quarks and who knows what next. So, big feed on small, the small disappear, yet big transform too. Giants come and go, while little men and women bear on. Change is—as Prospero might say—"such stuff as dreams are made of". You can tickle the Establishment, but the Establishment never goes away; its characters may be replaced but the roles remain.

Dictionaries pertain to civilized society, with an aim to confine the lawless jungle into order and fairness. How sad when they succumb to this very same jungle law, such as the recent change in Random House, which meant a mortal blow and terrible loss. On the other hand, I cannot lament the fate awaiting—according to Esposito—whichever legacy dictionary publisher whose so-called quality is undermined by a big name. The defamed Microsoft dictionaries could—eventually, if not yet—not be worse than theirs.

Sadly, established brand names are often something to beware of (so no surprise if sacred-cow slaughter becomes a global hobby). When their originality has evaporated ages ago, big names get fat and preoccupied with self-preservation and enforcing monopoly, then impede the advance of new spirits whom they copy. That is not a trait of dictionaries, but of humankind.

key point (because it is so hard to pin down), which is whether you could grow a dictionary business enough to attract serious investment money. The sad truth of the matter is that it may be difficult that one could. A little growth is not enough, and there hangs a tale (as you would say) – a tale of most of the corporate problems we have recently seen in American dictionary companies, including Merriam-Webster. Even if you could grow a dictionary business, under ideal circumstances, the growth may not be interesting or attractive enough for investors, American-style ones at least.

Other publishing entrepreneurs with a more worldly view, might be willing to accept a little growth. What puzzled me most, for example, about the dismantling of Random House Webster was that the dictionaries could be of immense benefit to the global branding of the Random House name, which fits in with the Bertelsmann global view, although not necessarily with the one emanating from the Broadway headquarters. For example, I am told that the Random House name is well known in both Japan and China, largely because of the local translations of the *Random House* dictionaries. But the latter benefits were apparently not strong enough to overcome the issue of growth for Random House/Bertelsmann.

I do lament this development, not only out of self-interest, but selfless interest when it comes to many good American lexicographers whose careers are being turned upside down by the stranglehold that dummies seem to have on dictionary publishing in the States today. But, all this is a natural process, in which the mighty fall by the wayside—out of hubris, complacency, or too much past success— and room is created for new scrappy small guys who are willing to take risks and innovate.

Links, notes, references

1 http://kdictionaries.com/newsletter/kdn9-1.html
2 http://kdictionaries.com/newsletter/kdn10-1.html
3 I roughly estimate that the combined size of the American monolingual dictionary business—measured in total annual sales dollars and including dictionaries sold directly to schools, exported abroad, as well as electronic versions—falls somewhere in the range of US$80-100 million annually. This number does not appear to have grown much during the past decade. One could point out that sales of electronic dictionaries, like the *Microsoft Encarta World English Dictionary*, do not seem to have greatly added to the total sales dollars, because of bundling with other products and/or the purchase of an electronic product in substitution for an equivalent one in print. This is all, of course, educated guessing.
4 http://qpb.com
5 http://google.com
6 http://pcwebopedia.com
7 http://kdictionaries.com/newsletter/kdn11-02.html
8 www.fltrp.com.cn
9 *The Random House Webster's Easy English Dictionary*, New York: 2001, paperback US$12.95, was certainly the first entirely new American English dictionary written specifically for middle school learners, a rapidly growing and underserved market around the world. In China, for example, soon every student, from the earliest grades, must master three core subjects: Mandarin, mathematics, and English.
10 One should also mention the *Newbury House Dictionary of American English*, edited by Philip M. Rideout, and published by Heinle in a revised edition in 1999. There is also a basic edition. Although Heinle is not a major dictionary publisher, these two editions are, according to Nichols, giving Longman a run for its money in the intermediate ESL market. McGraw-Hill, like Houghton Mifflin, has a small ESL group as well that tries to reach the domestic U.S. market.
11 *Meshuggenary: Celebrating the World of Yiddish*, by Payson R. Stevens, Charles M. Levine, and Sol Steinmetz. Simon & Schuster, New York: 2002.
12 Macmillan, London: 2002.
13 *Volume 29, Issue 3, September 2003*, MIT Press. www.mitpress.mit.edu/coli
14 See "Debate: The European Lessons," in *The Guardian Weekly*, April 18, 2001, posted at www.onestopenglish.com/culture/global/debate.htm.
15 Tim Berners-Lee, James Hendler and Ora Lssila, "The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of possibilities," *Scientific American*, April 2002.
16 Op. cit.
17 Also see the recent paper delivered by David Jost of Houghton-Mifflin and Win Carus of the Dictaphone Corporation, at the DSNA (Dictionary Society of North America) conference held this past May 2003, entitled "Is the Semantic Web Possible?" (www.duke.edu/web/linguistics/dsnaabstracts.htm-jost).
18 See www.perc21.org/cpe_project/cpe_project.html.