English Dictionary Making in America Today

Wendalyn Nichols

Wendalyn Nichols is the former Editorial Director of Random House Dictionaries, and prior to that the Editorial Manager of Longman Dictionaries. She is now a freelance consultant and lexicographer, and is on the advisory board of the American National Corpus, the board of the Dictionary Society of North America, and the advisory board of Montclair State University's graduate linguistics program.
wendalyn@nyc.rr.com

Why the UK still leads the way in the development of monolingual learners' dictionaries

The pioneering work in lexicographic publications for non-native learners of English was done in the UK, and the US has never really caught up. There are many reasons for this; the main one, I think, is the large size of the native speaker US domestic market combined with an unwillingness to cater to the special needs of immigrant populations; the prevailing attitude until the 1960s was the "bootstrap" mentality: "I (or my forebears) pulled myself up by my own bootstraps, and you should too."

The isolationism that prevailed in the US until the Second World War meant that few publishers saw the need to serve international markets, and domestically the US is such a large market for school publishing that the local educational publishers found it more lucrative to concentrate on producing school dictionaries geared toward the specific grade levels in elementary school and high school (called "elhi" for short). In contrast, Britain had a large empire (gradually replaced by the Commonwealth) as a ready-made market of people who needed to learn English (as a foreign language) to get ahead.

Once US publishers woke up to the need for special dictionaries for learners of English as a second language, they concentrated mainly on their already-established customers in the US market, specializing in literacy programs and bilingual (Spanish-English) education. These programs did not stress dictionary skills; at the lower levels students relied heavily on their bilingual dictionaries, and at the higher levels students were encouraged to switch to a standard native speaker dictionary.

Enough teachers admired the British EFL dictionaries that the *Oxford Advanced Learner's Dictionary* sold well in the US, and then Longman established a foothold in the 1970s. The *Longman Dictionary of American English* (LDAE) became the best-selling title once it was published in 1981, even though it wasn't truly American, being patchily Americanized from the *Longman Active Study Dictionary*. American publishers stuck to their elhi dictionaries, and so the British and US publishers happily split the market.

Why US publishers have been slow to create corpus-based dictionaries

The reason to keep up with the latest scholarship—like corpus-based lexicography—is an economic one, and too often reactive: if your books stop selling, then you figure out why. In the UK, the rivalry between Oxford and Longman, and the entry into the market of the COBUILD dictionary, meant that to keep up, everybody had to jump on the corpus bandwagon. US publishers, who were content to let the UK publishers have this slice of the market, did nothing about the new trend. Heinle & Heinle was the first US publisher to attempt an all-American ESL dictionary (the *Newbury House Dictionary of American English*), distinct from the Americanized ones coming from Britain, but it was written by one man rather than a team, and had no corpus input. Random House made the same mistake with its first foray into the monolingual ESL market, *Random House Webster's Dictionary of American English*. Now, it has always surprised me that a high percentage of US teachers prefer the Newbury House dictionary with its made-up example sentences to the second edition of the Longman one that is corpus-based; they like the pedagogical nature of the former. They'd gotten used to the first edition of LDAE, which pre-dates corpora and has example sentences that use a limited vocabulary.

It takes a lot of money to develop proprietary corpus data, and there was no equivalent initiative in America to the British National Corpus (BNC), because the US government has never supported lexicographic scholarship in the way that the UK has, and it's my understanding that the BNC would not have been possible without a huge chunk of money from Whitehall. At that time—the late 1980s and early 1990s—the ESL publishing market was undergoing great upheaval, with mergers, buyouts, acquisitions and divestments happening with such dizzying speed that even those US publishers who were aware of the "corpus revolution" could not convince their management to approve a significant, long-term, capital investment. Houses like Random House that did not have a history of selling into the ESL market didn't have the mergers problem to deal with, but they had the problem of financial models that no longer allowed for long-term amortization.

So, the UK educational publishers who have the greatest penetration into the US ESL market—Longman, Oxford, and to a lesser extent Cambridge—already have dictionaries now, and the US educational publishers remain unable to get approval for the kind of funding it would take to produce a product line that would rival the UK titles. McGraw-Hill ought to have seized the day—they had the cash, the sales penetration, and the size—but they chose instead to strike deals with other publishers to present their products to this market. NTC, the National Textbook Company, produces a large line of dictionaries that are, in my view, second-rate, but which people buy because they're cheap.

There is now the American National Corpus (ANC) Consortium, which got investment from enough publishers to start work that is modeled after the BNC so that comparative studies can eventually be done. The first 10 million words are being released this summer (2003). The initial founder investors have exclusive access during the developmental period; other commercial houses that wish to invest may still join, but at a higher fee than was the case for initial investors. Non-commercial educational institutions and individual researchers also have access from the start. The texts are being gathered under the supervision of Randi Reppen at Northern Arizona University; they are being tagged at Vassar under Nancy Ide; and the resultant corpus will be housed on the servers at the Linguistic Data Consortium at the University of Pennsylvania, which is also administering the licenses. (See next page.)

At this point, I see the UK and Japanese publishers as being more likely to take advantage of the ANC than American publishers, and for the disparity between British and American products to continue. I wish it weren't so; Charles Levine and I had great plans for the application of corpus-based lexicography to the Random House line, but what can you do when the visionaries don't hold the purse strings, and the upper management changes so often that you don't have a track record with them you can point to so that they trust you with large investments? This is the problem in nearly every US dictionary house; the one healthy one, Merriam-Webster, has so far remained unconvinced about introducing corpus-based lexicography. American consumers, meanwhile, will continue to make Merriam-Webster native speaker dictionaries their number-one choice; ESL teachers and students will continue to buy Americanized UK products.

The health—or otherwise—of US dictionary publishers vis à vis UK publishers

The top management of the big publishing groups look at the bottom line: dictionary publishing does not make the margins they like to see, so they are perennially putting pressure on the dictionary units to cut costs.

Merriam-Webster is the only major American dictionary publisher that is not under financial threat or at least dealing with perennial uncertainty: the publishers of the American Heritage line at Houghton Mifflin are still settling down after being sold by Vivendi; Random House closed its division in 2001; between 1997 and 2002, Webster's New World had three different owners. Encarta, the corpus-based UK-US collaborative project that was supposed to mark a new breed of dictionary, was done so quickly and edited so poorly that it was a near-complete failure: you now see copies of it everywhere on bargain book tables and street vendors' stalls next to the cut-price brands, because it had unprecedented numbers of returns of unsold copies from booksellers.

The Random House line, especially the great Unabridged Dictionary, is in danger of the fate of declining without any revision, unless another publisher decides to buy the rights to the Random House dictionaries and revive them. The current managers have even moved all of the citation cards into a storage facility where they cannot be readily accessed by anyone! Corporate changes are definitely a threat to the revision schedules and the very existence of the larger US dictionary publishing units.

Outside the US, American products simply do not have enough sales success to make an impact. The few exceptions, I think, included the works that Random House had the foresight (in the old days) to license for translation in Japan, Korea, and China – the beautiful editions of the Unabridged and College dictionaries that made Random House a respected name in East Asia. The American lexicographic tradition for native speaker products is long and illustrious, but the commercial climate has taken such a toll that the most brilliant lexicography now happens in specialized areas: Jonathan Lighter's *Historical Dictionary of American Slang*; the *Dictionary of American Regional English* project under Joan Houston Hall; and the recently-completed *Middle English Dictionary* at the University of Michigan, are examples.

Britain, in contrast, still maintains a commitment to promoting the English

language that is lacking in the US, so the UK-based publishers are less eager to divest themselves of dictionary units. The only dictionary house in the UK to undergo significant restructuring in recent years is Collins (the company is now HarperCollins), and this may have much to do with the fact that it is now owned by Rupert Murdoch's NewsCorp. Its schools assets in the US were sold to Pearson (Longman's parent company) in the 1990s; the COBUILD project was closed in the late 1990s because the sales of the product were disappointing. Collins still owns COBUILD and keeps updating it, but the lexicographic unit that produced it is no longer in operation. The dictionary program now concentrates more on native speaker and bilingual titles, and is based in Glasgow.

Having said that, it is becoming increasingly difficult for any commercially-owned unit, such as Longman Dictionaries, to get approval for new innovative capital projects – they seem to be in the "let's revise what we've got for now" mode. As for the two university presses: Oxford is also penny-pinching in most areas (it's more focused on its biggest capital project, the third edition of the OED); its Americanization of the *Wordpower* dictionary is not selling well. Cambridge now has a New York office and recently produced an American dictionary to compete with the LDAE, but its sales penetration is also disappointing.

What it will take to be a lexicographer in the future

The quality of a lexicographer will still depend heavily on all the traditional skills, as well as talent. I've trained plenty of people who learned the basic concepts but never became truly good, instinctual lexicographers – and unfortunately there are too many people out there who've had lexicographic training whose work is really quite patchy. Anybody can be taught the basic principles in a university course or an in-house training program on lexicography, but it takes someone with an instinct, an ear for the language—a poet, I would argue—to find just the right genus and differentiae and commit those to paper (or electronic database!) within the restrictions of a particular style guide.

A lexicographer will still need to have something of the teacher in him or her: an ability to convey complexity in a clear, simple, consistent form. A lexicographer will still need an unerring knowledge of grammar and a curiosity about usage and new words that keeps him or her alert to changes in the language – new words, new

uses, shifts in sociolinguistic register. He or she will still need to be able to interpret citations, which have their own role to play in an active reading and marking program alongside corpus data. He or she will still need a keen attention to detail.

The skills required of a lexicographer going forward are also going to include an ability to analyze corpus data quickly and judiciously, identifying and differentiating significant patterns from "rogue" uses of language, and making allowances for any bias the corpus may have. The lexicographer will have to understand data tagging and be able to work in an electronic medium, manipulating entries across databases.

Electronic applications and consumer (non-)awareness

There are some good CD-ROM products on the market from reputable companies, and then there are a lot of bad products with very old data sets being offered for license at bargain-basement rates. You get what you pay for. Electronic handhelds are still limited in their usefulness and helpfulness because of the limitation on memory; I think that wireless handhelds could solve that problem. That's where the future is, so whoever is first at successfully manipulating their data into a compelling, flexible, and useful format for wireless access, and can strike exclusive deals with the main manufacturers, is going to make a lot of money.

The perennial problem is that consumers the world over do not know how to tell a good dictionary from a bad one – it doesn't matter if it's print or electronic. They look at the number of definitions the product claims to have, and buy the one with the largest number. And the manufacturers of these devices often choose the cheapest licensing deal they can get rather than the best content. About the only defense against this is strong consumer awareness campaigns – if a manufacturer were to choose a high-quality licensing partner (or develop its own high-quality English content) and then hit the market with a very strong marketing campaign that focused on the quality of the product, educating the consumer in the process, then it might make a dent in this trend. That's how Longman beat out Oxford in many markets: they were quicker to exploit corpus resources and more innovative in their applications, and were able to demonstrate the difference in a global blitz of teacher-training workshops and conference presentations. Therefore, schools that teach English ought to be teaching the students how to choose a dictionary; you're not going to convince

manufacturers to reform their practices, so you've got to teach the consumer not to buy the inferior products.

The Internet also contributes to the confusion of quantity—or ease of access—with quality. Being mindful of the quality of the source matters, regardless of whether the delivery format is print or electronic. I think it was a mistake to offer online dictionaries for free – the newer works that are still under copyright and are the most up-to-date should have been set up with a subscription model from the beginning. Internet users now feel that they have the right to free information, no matter how much it cost the original publisher to produce it. Some publishers, like Columbia University Press, have been successful with encyclopedic works offered online by subscription, and I think people will start to accept this model, especially now that companies like Napster have been barred from allowing free music downloads of copyrighted material.

Dictionaries and references

American Heritage Dictionary of the English Language, 4e, 2000. Boston, MA: Houghton Mifflin.

ANC: the American National Corpus, www.americannationlcorpus.org

BNC: the British National Corpus, www.hcu.ox.ac.uk/bnc

Cambridge Dictionary of American English, 2000. New York: Cambridge University Press.

Collins COBUILD English Dictionary for Advanced Learners, 1987, 1995, 2001. Glasgow: HarperCollins.

Dictionary of American Regional English, 1985, 1991, 1996, 2002. Cambridge, MA: Harvard University Press.

Encarta World English Dictionary, 1999. New York: St. Martin's Press.

Longman Active Study Dictionary, 1983, 1991, 1999/2000. Harlow: Longman.

Longman Dictionary of American English, 1983, 1997, 2002. New York: Longman.

Longman Dictionary of Contemporary English, 1978, 1987, 1995, 2003. Harlow: Longman.

Merriam-Webster's Collegiate Dictionary, 10e, 2002. Springfield, MA: Merriam-Webster.

Middle English Dictionary, 8e, 2001. Ann Arbor, MI: University of Michigan Press.

Napster: www.napster.com.

Newbury House Dictionary of American English, 1996, 2000. Boston, MA: Heinle & Heinle.

Oxford Advanced Learner's Dictionary of Current English, 1948, 1963, 1974, 1989, 1995, 2000. Oxford: Oxford University Press.

Oxford American Wordpower Dictionary, 1998. New York: Oxford University Press.

Random House Historical Dictionary of American Slang, 1994, 1997. New York: Random House.

Random House Webster's Dictionary of American English, 1997. New York: Random House.

Random House Webster's Unabridged Dictionary, 2001. New York: Random House.

Webster's New World Dictionary and Thesaurus, 2e, 2002. Hoboken, NJ: John Wiley & Sons.

The American National Corpus

The ANC Consortium members include publishers, software companies, and academic members. Consortium members have exclusive access throughout the development period and for five years after the full corpus becomes available. The ANC is freely available for the purposes of academic research and education.

Acuired data includes, so far, about 2 million words of spoken data (the LDC Switchboard corpus and a portion of the CallHome corpus); 1.5 million words of previously unreleased newspaper data from the *New York Times*; a few hundred thousand words of "ephemera" (pamphlets, newletters, etc.); several novels published by Oxford University Press USA; Berlitz Travel Guides from Langensheidt; Verbatim magazine; government documents drawn from the web; about 5 million words from Slate magazine (Microsoft); and about 900,000 words of research papers from the Association for Computational Linguistics.

Commercial Members
• Pearson Education
• Random House Reference
• Langenscheidt Publishing Group
• HarperCollins Publishers
• Cambridge University Press
• LexiQuest
• Microsoft Corporation
• Shogakukan Inc.
• ACL Press Inc.
• Taishukan Publishing Company
• Oxford University Press
• Kenkyusha Ltd.
• IBM Corporation
• Obunsha Publishing Co. Ltd.
• Bloomsbury Publishing Plc
• Benesse Corporation
• Sanseido Co. Ltd.
• Sony Electronics Inc.
• Macmillan Publishers

Academic Members
• Vassar College
• Northern Arizona University
• New York University
• Linguistic Data Consortium, University of Pennsylvenia
• International Computer Science Institute, University of California, Berkeley
• University of Colorado at Boulder