Using the Internet as a Resource in Writing and Updating Dictionary Entries

Luanne von Schneidemesser
Dictionary of American Regional English, University of Wisconsin-Madison

In the midst of writing volume IV of the *Dictionary of American Regional English (DARE)* we realized that we had wonderful resources available that had not been available to us earlier, namely the Internet and its many historical and other websites. We started checking a few sites in the course of writing some of our entries, thereby improving some of them considerably; we pushed back the earliest known date of usage by one hundred years or more in some cases, and showed that other terms were not obsolete but still current. By the time we reached volume V, which will contain *sl* through *z*, we were checking some sites routinely. When NewspaperARCHIVE.com with its holdings of newspapers dating back into the eighteenth century became available to us last year, it was like a whole new ballgame.

This paper will discuss three aspects of Internet use on which *DARE* has embarked in the last year, namely, first, a small pilot project to determine how much of an effect searching the web would have if we were to update the text of our published volumes, as reported in a paper given at LAVIS, a conference on Language Variation in the South, in April 2004; secondly, an updating of entries already written for volume V, on which we are presently working and which we hope to have published in 2009; and thirdly, a follow-up pilot project to the first one to determine the feasibility of colleagues around the country working with their students to help us update our already published entries to prepare for an on-line version of *DARE*, something that we *DARE* staff members simply don't have the time to do.

Allow me to start with a brief introduction to *DARE,* as presented to University of Wisconsin-Madison students who worked on the third of these projects in September 2004:

DARE INTERNET PROJECT

The *Dictionary of American Regional English* (*DARE*) is a reference tool unlike any other. Its first four volumes were published between 1985 and 2002. Volume V is projected to be completed in 2009. *DARE*'s aim is not to prescribe how Americans should speak, or even to describe the language we use as cultivated speakers and writers, but to document the varieties of English that are *not* found throughout the country—those words, pronunciations, and phrases that vary from one region to another, that are learned at home rather than at school, or that are part of our oral rather than our written culture. Although American English is remarkably homogeneous given the tremendous size of the country, there are still many thousands of differences that characterize the various dialect regions of the United States. It is these differences that *DARE* records.

The *Dictionary* is based both on face-to-face interviews carried out in all fifty states between 1965 and 1970, and on a comprehensive collection of written materials (diaries, letters, novels, histories, biographies, newspapers, and government documents) that cover our history from the colonial period to the present. These materials are cited in individual entries to illustrate how words have been used from the seventeenth century through the end of the twentieth. We include the earliest and latest citations we find illustrating each word or phrase in each sense as well as a selection of citations representing variant forms, regions, and social usage. The entries also include pronunciations (if they vary regionally or differ from what would be expected), variant forms, etymologies (if *DARE* can add to what is already known about a word's history), and statements about regional and social distributions of words and forms. A feature unique to *DARE* is its inclusion in the text of the *Dictionary* of numerous

maps that show where words were found in the 1,002 communities investigated during the fieldwork.

The present Internet project is an undertaking to update our entries by adding earlier and later citations which are now available through searching on many of the websites which have recently appeared containing historical documents, newspapers, journals, and other materials, as well as websites offering evidence of more recent usage of the terms. During a pilot project involving twenty-eight terms, we found that over half of the terms were antedated, over 70% postdated. While this may not be true to this extent for any given run of entries, we believe that the history of many of the words cited will be extended through this project. These resources were not available to us twenty years ago, or for the most part even five years ago.

Our project assistant has written up instructions for carrying out the project. You will be assigned entries from volume I to search for in certain websites. If you have questions please e-mail. Your findings will be incorporated by an editor into the online version of *DARE*, which is scheduled to come out after publication of volume V, the last volume of dictionary text.

Thanks for your help.

The first pilot project mentioned above was one reported on at LAVIS, at Tuscaloosa, Alabama, in April 2004. When the call for papers came for the LAVIS conference, Joan Hall and I decided to determine just how much of a boon Internet searching would be, something we had been wondering about for a few years. We chose twenty-eight Southern terms from volume I, which was published in 1985, looked at a great many websites, worked out procedures, hired two students, gave them a list of headwords and a list of websites to check, and kept track of results.

We wanted to determine both how effective the various websites would be in supplementing our traditionally-collected materials, and if the nearly thirty-five years since the end of *DARE* fieldwork significantly affected the regionality of these selected *DARE* entries. We could then use the knowledge gained on our continuing work on volume V and on updating our text for an electronic version of *DARE* which will follow the print version.

Adam's housecat n Also ~ *cat,* ~ *house* chiefly S Atl, Gulf
States See Map
=Adam's off-ox 1.
   1908 *DN* 3.285 eAL, wGA, *Adam's (house-)cat* . . . "He wouldn't know me from Adam's house-cat." 1965–70 *DARE* (Qu. II26, . . *"I wouldn't know him from _____ "*) 83 Infs, chiefly S Atl, Gulf States, Adam's housecat; Infs LA25, OH90, VA69, 71, Adam's cat; AL10, Adam's house; FL48, A housecat, [corr to] Adam's housecat. [Of all Infs responding to the question, 26% had less than hs educ; of those giving these responses, 56% had less than hs educ]
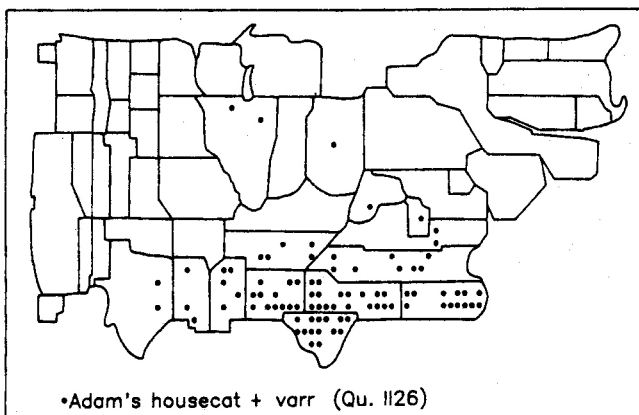
•Adam's housecat + varr (Qu. II26)

Figure 1. Entry with map for *Adam's housecat.*

The terms we chose would be typical of what would face us later, we thought. These included ones easy to search such as *Adam's housecat* (see Figure 1, *Adam's housecat*[1]), and *coffin tack* "a cigarette or cigar," through the range to difficult ones (with many other meanings) such as *back* "to address a letter or envelope" and *change* "to castrate (an animal)." The students each worked individually with me for the first two or three sessions, on four entries and three websites. Throughout this process we added to the instructions, clarified and refined the task. After that, we assigned each of them entries and both received a set of about a dozen URL's for websites they were to search. They could work whenever and wherever they wanted but stayed in contact with me via face-to-face visits, e-mails, and phone calls. After six weeks we interviewed them individually about the project. They agreed on four of the five best websites for our purposes: both Making of America sites (Cornell and Michigan); NewspaperARCHIVE.com in spite of its horrendous searching and its unreliability; and LexisNexis. One said Google was best for postdating. One also thought highly of Wright's American fiction within its twenty-five year range (1851–75), and of ProQuest.[2] We discovered, however, that beyond just choosing which sites to use based on their contents, the search platforms are important as well as the search strategies of the person doing the searching. Relevant questions include: Is there a good search engine? Are there wildcards for various forms and spellings of the terms being searched? Is there a preview of the quotation so meaning can be determined? Are dates of the quotes given in the list of findings? Are the sites reliable or do they break down? Is the Optical Character Recognition on which the searching is based reliable or are there lots of false finds? Can the searches be refined with AND or NOT?

What was found by our students was more or less what we expected: the terms we thought would be easy to find were found, and the hard ones took incredible amounts of time, sometimes without results. *Change,* with its many, many different meanings, could not be found in the meaning we were looking for; but *coffin tack* was found and antedated by five years, postdated by thirty-three, for example. *Clearseed,* "a peach in which the flesh does not adhere to the seed," was a bit of a surprise: we had thought that would be an easy one to find but neither student found anything. The end result of the project: earlier citations were found for 54% of the terms searched, mostly in the range of ten to sixty years, with the greatest extending the history of the word by 114 years (*airish*). Seventy-one percent of the entries now have later citations. Fantastic results, we thought.

And the regionality? We found that it is more difficult to assign regions to citations from websites than in our traditional materials. But for those citations we could locate, most of them proved to be from the South or South Midland, our original area for this project. The tighter knit the region of our original entry, the more this held true. For those original entries not so compact in region, our findings showed somewhat more evidence outside the South. *Airish*, "inclined to put on airs," for example, demonstrated a wider area of use than our original entry showed, leading us to think it might even be a standard term. In a few cases it was not regionality but definition which must be changed, as with *on the carpet,* which we defined as "eager to marry." It is more than that; eagerness indeed, but not just to marry.

The time needed per update (thirty-five of them) worked out to 4.75 hours each. The students together worked a total of 156 hours; that included of course training, becoming familiar with the many different websites, and documenting what they were doing. All in all, a large time-commitment.[3]

---

[1] This entry shows several features of *DARE's* entries. The definition is given as a reference to the entry *Adam's off-ox*, sense 1 ("A person or thing one does not know and cannot identify"). All variant forms or spellings are included, here *Adam's cat* and *Adam's house. DARE* collected data in 1002 interviews it carried out between 1965 and 1970, involving 2,777 informants. Statistics of the informants are included in entries when relevant, as here amount of formal education. Other factors are age, race, sex, and size of community (five categories, rural to urban). Maps showing where *DARE* informants used the word or phrase are included with entries, demonstrating regionality.

[2] Our list is specific to the *DARE* project. Joanne Despres, of Merriam-Webster, came to quite a different list than we did for updating entries in *Merriam-Webster's Collegiate Dictionary*, 11th edition, for example. See Despres 2003.

[3] For more details of this project, see Hall and von Schneidemesser (forthcoming).

But the LAVIS results were so encouraging that we realized we should update those volume V entries produced before NewspaperARCHIVE.com and a few other sites were available to us or before they were routinely checked by editors. So for this second undertaking we kept one of the two students, Lindsay Skotterud, on staff, modified her duties, changed the list of websites to search (added a few to try out, deleted less productive ones, still at about a dozen), and started her working on the already-written volume V galleys, of which we had about forty as of March 2004[4]. Our intent was to examine the findings after five galleys had been completed and at that point to make any changes we felt would make the process more efficient (cutting the number of websites, in the main) without losing too much in outcome, weighing time against resources, or money, if you will. Lindsay basically started on volume V in the fall of 2004. As of March 2005, she had completed searching on only three galleys. She did not search either Making of America site, or American Memory Historical Collections, since the editors looked at those when doing the original editing.

These three runs of entries with their senses taken together, contained 185 entries and senses. Lindsay searched for postdates in only 127 cases, since if we had a citation from 1980 or later in the original entry she was told not to search for a postdate. The preliminary findings show that 108 updates were found, up to 95 years earlier and 105 years later, an average 27 years earlier for antedates, 33 years later for postdates. Having spent around 270 hours on these searches, she averaged 2.5 hours for each update. So, in comparing these results to the LAVIS pilot project, the results are not as spectacular, but 25% of the entries or senses have been antedated, 50% postdated, or taken together, 35% have been updated (counting only those searched). And the time-commitment was cut. Quite a respectable result.

Which sites were the most helpful to *DARE*? Sixty-seven percent of the antedates came from NewspaperARCHIVE.com; 89% of the postdates came from Google. These are the two sites with which Lindsay started her searching on each run of entries, since she knew them to be most productive. Other sites lagged far behind: three updates were found on Wright's American Fiction with its limited dates (1851–1875), LexisNexis, and ProQuest Historical New York Times; all other sites resulted in fewer than three.

What have we learned? That it is definitely worthwhile to do the searching. That different runs of entries are very different in their results. The galleys we looked at contained the entries *slab* to *slashways, slat* to *slew-jawed,* and *slice* to *slip.* In the *slab* galley we found 45% updates, in *slat* only 26%. *Slat* contains several words with many standard meanings: *slat, slave, slay, sled, sledding, sleep, sleepy, sleigh,* and *sleight. Slab* has some but fewer: *slab, slack, slam, slap, slash.* One can spend large amounts of time sorting through such results on sites without good search functions or previews. In the long run this isn't important, but for those people doing the searching, they do need encouragement that their searching will result in valid findings. Lindsay wrote:

> It can be quite frustrating when it seems like I haven't had much luck with a certain galley or entry, but that's just the way it goes sometimes. This has probably been the toughest aspect of my job so far—to know when to draw the line and move on. When you come to a word that yields hundreds or thousands of results and you just can't narrow it down further, there comes a point when it's necessary to call it quits. . . . Some galleys will yield many results, and some won't.

The new citations have led to a new sense of *sleeper,* "to apply a temporary brand (to an animal)." A couple of entries have been deleted, for example, *slat and wire fence,* which turned out to be fairly standard, not just especially in Kentucky, as our original material showed. A headword was changed, from *slab harrow* to *slab drag,* the latter being more common according to this additional research. We changed the region in a couple of cases. *Slap* sense *1,* an adverb meaning "directly, abruptly," as in 'I clocked him right slap in the face,' was changed from 'chiefly South, South Midland,' to a broader region, with quotations added from Massachusetts, Pennsylvania, etc. And *slab drag* was assigned a region, of 'scattered but especially Maryland;' a hundred-year gap in the entry was also filled.

---

[4] A galley for us is around 80,000–90,000 characters, roughly between 55 and 85 entries and senses.

These findings do add an enormous amount of evidence to our *Dictionary* text.

The pilot of the third project was finished in January 2004. At the LAVIS conference, I jokingly suggested at the end of our talk in response to a question that due to the amount of time the Internet searching requires, we could just assign certain pages to each of the members of the audience to update for us. Connie Eble and Catherine Davies immediately picked up on this idea and said they would be interested in having students in their classes work on this for us. Others joined in with the offer. We thought this was a wonderful idea and plan to follow up on it. The two students who did the initial pilot project had worked directly with me in three or four sessions, for training, to clear up questions, so we knew they were doing what we wanted. After that they were basically on their own. They did well. But how would it work if we simply sent out instructions and assigned entry words from our first four volumes to whole classes of students? We decided to try it out first on a smaller scale at the University of Wisconsin-Madison, where we would be available for help if needed.

Prof. Anja Wanner was open to the idea; she happened to be teaching a class on 'The Word' and had asked me to discuss in her class not just what *DARE* was about, but to elaborate on how *DARE* editors decide what to enter in the dictionary, and in what form. Perfect. An added assignment for the nine graduate students in the class would be the task of searching selected websites to antedate and postdate entries from volume I. We updated our instructions, developed by our two students during the course of their work for LAVIS and now expanded with this different purpose in mind (using student-speak). We chose a dozen websites to be searched. Was that the right number? Enough? Too many? We described each of the dozen sites briefly in a handout with a few tips on using them. We wrote a short introduction to *DARE*, the one quoted earlier in this paper. We figured that since the students were studying dictionaries among other topics in their class they would have basic knowledge of dictionaries. The first few entries in volume I were left off (consisting of the letter *a* in different meanings, not an easy thing to search), and a grid of entries was set up, starting with *aback* ("behind" and "ago", as adv.). How many headwords and senses should each student search for? Working with Prof. Wanner, we decided the project should be about fifteen to eighteen hours and include writing a short paper of two to three pages about the experience as well as printing out pages and title pages with the citations found, as had been done for LAVIS. That seemed like it would work.

We circulated the documents to the students via e-mail, then Lindsay and I (against the principles of how this project would be carried out at remote campuses in the future) held an introductory session in a computer lab, explaining more about *DARE* and the structure of entries, and demonstrated two websites briefly, adding to the tips written up, thinking that this should be a breeze for graduate students, a generation which routinely spends a lot of time with computers. We received a lot of blank looks when we asked the students to replicate what we were doing on the computers in front of them in the lab. This should have been our first clue. But we went ahead and assigned each student five to eight headwords/senses to search for, and told them we would be available via e-mail or appointment if they ran into problems.

We heard from no one. One student worked as a student hourly for *DARE*, so she kept me apprised: people were frustrated; they weren't used to searching the web, or at most used only Google; the work assigned was way too much, the hours and the number of websites too many. Prof. Wanner had a meeting with them; our *DARE* student tutored them on website usage; and the requirements of the assignment were reduced. It should perhaps be mentioned that three of the students were not native speakers of English. Once they got into the project (they had the same problems as the native speakers with using the websites) I don't believe this was a great handicap, but it was a handicap. For instance, our entry *Abraham* is very short: the definition is '= **Adam 1**,' actually referring the reader to the first sense of the entry *Adam*. The entry *Abraham* contains only one quotation, from the *DARE* fieldwork: '**1968** *DARE* (Qu. II26, . . *I wouldn't know him from* _____.) Inf **IN**35, Abraham.' The meaning of this was not clear to the non-native speaker working on the entry.

The reduction in the assignment included the following: the amount of time required had been lowered to about ten hours, and the number of websites to be searched had been changed to focus on one, with the concentration on a few (one or two) of the headwords assigned. OK. This was a learning experience for all of us. Actually used were between two and twelve sites, checking *most* of the words assigned. The downside of the results was that we had made assumptions which were totally untrue,

namely the familiarity of the students with doing web searching, the amount of time the assignment would take, and the precision with which the students would follow instructions, allowing us to receive some findings which were quite unhelpful. And they wanted more feedback as they worked (although they had not contacted us with questions or comments). The upside was that we got about twenty updates, either antedates or postdates. That's not a lot, but we'll get better at explaining what we want, what we are looking for. And we learned a lot about how we have to restructure our undertaking, especially considering we won't be there to explain in person to colleagues' students at, say, the University of North Carolina-Chapel Hill or the University of Alabama. More specifically, what we learned from the students' papers and results, what we will have to address or emphasize more in future instructions to students is as follows:

Some of the students felt that they should obtain results for every headword/sense, although we had said this would not be the case. They became frustrated with no results, as well as with too many results, the latter which meant the students had to spend a lot of time going through them. And many website searches do not provide a preview of the findings so they could check on meaning, region, date, etc., before they loaded the .pdf of the page, a process which can be quite tedious. Narrowing down a plethora of results can be very time-consuming, as we knew and they discovered.

Some of the fine points of definitions and our entry structure were not clear; we received citations for *abeam* and *abscond* in their standard meanings, not in a transferred meaning, e.g., 'When you've brought the meetin' house abeam you'll see the Town Hall round the bend,' or reflexive use, 'I immediately absconded myself' in the sense "hide, conceal;" or we received citations with the correct headword but with meanings not in our entries at all.

Instructions were given to the students before they started the project. But it evidently also wasn't clear enough that we were interested only in predating or postdating what was in the entries; we got other, in-between dates. Very frustrating for us because it shows us that there will always be problems with having students do such searches for us, is that several quotations were from non-U.S. sources (London, Jamaica, Cornwall). Those are easily dealt with: we omit them. But several other quotations would have to be researched further to determine if they were indeed from American sources, since title pages were seldom printed out with the page containing the citation. And much more basic, disregarding the lack of title page even, is that in many cases no quotation was printed at all but simply the URL, sometimes running to five lines in length, where the word or phrase evidently occurred. (The students were given the choice of turning the assignment in electronically or in hard copy, I might add, the majority choosing the latter option.) These give us no clue as to usefulness: Is it our definition? Is it American? Where is it used? The date? The source (beyond the website URL)? Can we get more basic, more direct, about their assignment? We obviously need to state several times in several ways that citations should be printed out, from the original; if from a book/journal, the title page should also be printed and attached, with date, page number—all those 'little' points are essential to us. We thought we had it covered; quoted here are examples from our instructions, the first under the heading 'What to do once you've found a relevant usage:'

First, make sure that the author of the document (whether it be a newspaper article, random website, online periodical, or a novel) is from the U.S….! Don't waste your time (not to mention paper and ink) only to find that the author is from another country. After all, this is the *Dictionary of **American** Regional English*.

We continue:

When you know you've got the earliest you're going to find on that database, print out the PDF image of the page that the word is used on. Also, be sure to print out the title page of the book, or masthead (front page) if the citation comes from a newspaper.

And a bit later:

For books, make sure the title page includes book name, author, publishing house, publication location, and publication date.

And another reminder:

> Staple the title page/masthead behind the page containing the citation.

The project did not yield results of the type we had hoped for but there is definite potential. We are still working to incorporate in the best way we can all we have learned from this exercise. It is easy to train one or two students, students we have interviewed and chosen for the job. It is more difficult to draft instructions which will be sufficient to cover everything a group of students will need to know (and will assimilate) who will spend only a short time working on the project. And there are some basics which not all students have grasped in their study of language. Can we teach students the difference between past tense and past participles? Between adverbs and adjectives? Transitive and intransitive verbs? This we can do with one or two students, but we'll need help from the professors of the classes to deal with these basic language questions. Going beyond grammar, can we teach them that some websites may have wildcards to allow all forms to be searched, e.g., *abaloning, abaloned,* and *abalones*; that others will require a search for singular and plural, or present and past, separately? Can we teach them to read the instructions, both ours and those of the websites? We think so. We hope so. We're going back to the drawing board to restructure and redefine, and then we'll ask for the help of colleagues so as much of our text as possible will be updated when it finally goes on-line. And our web searching for volume V text is continuing, with one or two chosen students in-house, to produce as complete a volume as possible, given our restrictions of time and resources. Perfection will never be possible in lexicography, but we're trying our best. Overall these three Internet projects have produced results which are definitely quite encouraging in working towards this end.

References

American Memory, Library of Congress. 4 April 2004 <http://memory.loc.gov/ammem/>.

Despres, Joanne (2003) Dating in *Merriam-Webster's Collegiate Dictionary*, Eleventh Edition. Paper presented at the XIV Biennial Meeting of the Dictionary Society of North America, Durham, NC, May 29–31.

Dictionary of American Regional English (1985–) Cassidy, Frederic G. and Joan Houston Hall eds. Vols. I–. Cambridge, MA: Belknap Press of Harvard University Press.

Google. 4 April 2004 <http://www.google.com>.

Hall, Joan Houston and Luanne von Schneidemesser (forthcoming) The South in *DARE* Revisited. In Picone, Michael D. and Catherine Evans Davies eds. *Language Variation in the South III: Historical and Contemporary Perspectives.* Tuscaloosa, AL: University of Alabama Press. (Volume title and press pending.)

LexisNexis. 4 April 2004 <http://www.lexisnexis.com/>.

Making of America, Cornell University. 4 April 2004 <http://cdl.library.cornell.edu/moa/>.

Making of America, University of Michigan. 4 April 2004 <http://www.hti.umich.edu/m/moagrp/>.

NewspaperARCHIVE.com. 4 April 2004 <http://www.newspaperarchive.com/>.

ProQuest. 4 April 2004 <http://www.proquest.com/>.

Wright American Fiction 1851–1875. Committee on Institutional Cooperation, Indiana University Digital Library Program. 4 April 2004 <http://www.letrs.indiana.edu/web/w/wright2/>.

Selected Proceedings of the 2005 Symposium on New Approaches in English Historical Lexis (HEL-LEX)

edited by
R. W. McConchie, Olga Timofeeva,
Heli Tissari, and Tanja Säily

Cascadilla Proceedings Project   Somerville, MA   2006