Transactions of the American Philological Association 137 (2007) 491–495

3. Finishing the *TLL* in the Digital Age: Opportunities, Challenges, Risks*

MICHAEL HILLEN
Thesaurus Linguae Latinae

WITH THE NEAR-INSTANTANEOUS RETRIEVAL AND PROCESSING OF INFORMATION, the digital age has created the expectation of increased speed in all branches of scholarship involving the sorting and categorization of verbal or numerical data. A lexicographical project like the *TLL*, one might imagine, could be transformed without diminishing its world-renowned standards. Indeed, our team in Munich is very conscious of the need to speed up, not least to satisfy funding agencies that are becoming intolerant of long-term projects. How can we adapt the content of the articles in the lexicon, now that some information is easily obtainable from digital databanks such as the Packard Humanities Institute Latin Databank, the Bibliotheca Teubneriana Latina, the CETEDOC Library of Christian Latin texts, and the epigraphic databases?

In terms of method it matters little whether such resources are available digitally or in print, although digital resources are quicker to use. This is because of the way in which the Thesaurus's own archive of data is organized. The *TLL* is concerned with the vocabulary of the Latin language from its earliest attestations in the third century B.C.E. down to 600 C.E. At the end of the nineteenth century, after intensive discussion, the founding fathers of the Thesaurus decided to divide the testimonials of this 900-year period in half, and deal with the two halves differently. For the period down to Apuleius in the mid-second century, each occurrence of every word in all the textual sources is recorded, and the inscriptions and coins are recorded in full down to the beginning of the Principate; in practice, this means that for every occurrence of a word there is a handwritten slip displaying it in its surrounding context. But, for the period from Apuleius down to 600 C.E., most of the textual sources have been excerpted: specialists on the respective author, genre, or subject worked through the text, marking noteworthy usages

* Translated by Kathleen Coleman.

that depart from the norm. (The exceptions, recorded in their entirety, are significant texts such as Augustine's *City of God*, and the works of Tertullian and Commodian.) Excerption continued until the end of the nineteen-seventies, and references to newly published indexes and vocabulary studies were also added. New discoveries continue to be incorporated in accordance with the same principles: a text dated before Apuleius—e.g. the *senatus consultum de Pisone* from the year 20 C.E.,which was found in Spain in the nineteen-eighties—is fully recorded in the archive; a later text—e.g. a new sermon by St. Augustine—will be excerpted. Nowadays, in this second category we limit ourselves to looking for words that are seldom attested, so as to increase our data as rationally as possible.

The process of excerpting the second half of our material greatly reduces the amount that we have to deal with, making it possible to do the work at a reasonable pace (or, indeed, at all). Just as important is the overall relevance of the excerpted material. The process of excerption also chimes with an important principle underlying the composition of the individual articles. An article ought to present a concise overview of what is normal, unexceptional, and predictable, whereas what is unusual, exceptional, and abnormal should be treated more fully, if not exhaustively. As a result, lemmata that have fewer testimonia are often dealt with in full, unless all the attestations conform to the same pattern. In short: in assembling the archive of material, which today comprises more than 11 million slips, a huge amount of scholarship, energy, and expertise has already been invested.

It is clear that, for the first half of our material, the digital databanks cannot contribute anything new, other than plugging gaps in authors like Gellius, whose *oeuvre* was not comprehensively entered on slips. For the second half, with the exception of specific texts and genres that are not yet fully available in digital form (e.g. the medical authors), the databanks contain three categories of material: everything that our archive already contains; everything that our excerptors deliberately omitted; and much that the excerptors and the original compilers either did not know or could not have known. Hence, for the second half of our material meaningful expansion is possible; at the same time, however, decades of energy and expertise could be vitiated by expansion that is indiscriminate and unregulated. Material that is accessible at a keystroke cannot be analyzed so fast.

The difference between the inclusiveness of the databanks and our excerpted material is easily demonstrated. For the lemma *peccatum*, disregarding later expansions and addenda the basic archive records 860 instances, CETEDOC more than 30,000. For *poena* the archive records 2,280 instances, CETEDOC more than 8,000. For the preposition *per* more than 15,000 slips had to be dealt with, a task that took two years; the databanks number over 90,000 instances. And, to cite neither a preposition nor an overtly Christian candidate, even *pondus* more than doubles its recorded appearances. It is obvious that such enormous amounts of material are a great hindrance to a manageable undertaking, and would achieve the exact opposite of the increased rate of progress that is so urgently needed. Less dramatic examples are compelling too: for the verb *plaudo* the databanks yield 509 instances (including forms of the verbal noun *plausus*), approximately twice the number of slips in the archive; furthermore, this example illustrates a disadvantage of the databanks in comparison with the archive, in that words that look the same are not distinguished from one another. In a case like this, where the increase in attestations is moderate, it is very tempting to expand the raw material for the *TLL* article; nevertheless, it is all too likely that the extra time needed would far exceed any advances in our knowledge of what the word means and how it behaves. For the noun *plausus*, for example, the excerpted material in the archive includes four instances (one in Ambrose and three in Jerome) with the sense of "words that sound good without meaning anything." A search through CETEDOC might have been expected to reveal further instances of this usage; yet a study of the 160 post-Apuleian instances that it records, including the identical form of the past participle passive *plausus*, yielded no further examples, and the time expended on this enquiry proved instead how well the excerptors had done their job.

A further indication of the esteem in which the excerpted material should be held becomes evident in considering how to deal with the enormous number of attestations for negatives and particles beginning with N. The problem here is not so much a matter of adding new material to what has been excerpted but, rather, the opposite: how to cope with the exhaustive documentation for the first chronological period. For *nam* the archive has 12,000 slips, of which approximately 2,200 comprise excerpted material. *Ne*, which is the same size, has about 1,650 excerpts. For *neque* and *nec* we have 24,000 slips, about 5,500 of them excerpts. And, finally, *non* has 44,000 slips, of which about 9,000 comprise excerpts; in other words, for *non* there are about 35,000 instances down to the time of Apuleius. On average, the exhaustively documented material outnumbers the excerpted material by a factor of 4.5 to 1. Practically speaking, if the project is to pick up speed, it is for the exhaustively documented material that methods must be found to reduce the number of slips that are to be given comprehensive treatment, with emphasis being laid on texts that do not conform to grammatical and stylistic norms.

The main value of digital databanks for the work of the Thesaurus cannot, therefore, be a systematic increase in the raw material. Rather, they are useful in three specific areas: reproduction, checking, and regulated expansion of our sources. To take reproduction first: attestations of the material of the first period can easily be reproduced by electronic means. This is useful where, for example, comprehensive digital indexes are available, as is the case for the slips containing citations for *pretium* or *poena* that are cryptically attributed to the jurists without further specification. It is also helpful that the databanks are based on up-to-date editions; in this respect they are superior to the slips, although the absence of a digital *apparatus criticus* means that one cannot dispense with checking the printed editions. All the same, in this respect the databanks have the potential to increase the speed of our work.

Second, checking: the philological work of the contributors to the project, like that of all philologists, is made much easier now, because in hard cases, such as the word-choice of a particular author, it is easy to do a quick check. The distribution of a word in particular authors or genres can be checked quite quickly as well. For a good *TLL* article, this is information of no small importance. So here, too, there is potential to pick up speed. In terms of content, the *TLL* can also profit, via the third category: expansion. Even the excerptors were not perfect, and it is therefore assumed that, in the case of lemmata with only a few attestations, contributors will consult the databanks; in light of the founding principles of the Thesaurus, it should be clear that there is greater rejoicing over the gain of a single attestation for a word that only has ten others, than over the gain of fifty new attestations for a word that already has 500, even though the ratio of expansion is the same (10%).

The expansion of citations for authors whose works have not been comprehensively entered on slips, the means of checking the completeness of coverage in individual authors: these are tasks quickly accomplished with the aid of digital resources. Qualified expansion and enrichment of the material in the archive, however, do not usually occur until work on an article is under way, since it is then that questions arise that can be answered by recourse to databanks, i.e., questions to do with phenomena that can be defined in formal terms: singular or plural, combination with a specific adjective, prepositional expressions, etc. Of particular benefit is the capacity to search databanks for lemmata with enormous numbers of attestations (such as prepositions and the tidal wave of negatives), because at a late stage questions may arise concerning material that has already been dealt with, and a contributor cannot afford the time to answer them by working through the material all over again. For instance, in the article *per* the question arose, under what circumstances the reflexive pronoun in the phrase *per se*—contrary to the normal rule—no longer refers to the subject of the clause. A search of the Latin databank produced about 100 pages of examples, a formidable but still manageable body of material, whereas a manual search of the enormous archive of slips would have taken a very long time. Previously, research like this would have been unthinkable, because of the time it would have taken. Nowadays, however, we can easily pose these questions, and in terms of the quality of the lexicon this is a tremendous gain. But, because every answer costs time, one must always ask whether the result justifies the effort. This aspect of digital resources does not, in fact, speed up the work; it makes it slower, precisely because it is now possible to deal with issues that could not be dealt with before.

One practical gain is worth noting. Everybody who has used the *TLL* is familiar with the technique of employing brackets and parentheses so that expressions of the same form, or citations with shared content, can be grouped together. The same user probably also recognizes the technique of omission, where *al.* signals that in the relevant author or passage more instances can be found in the Thesaurus archive. Nowadays, if linguistic usage rather than context or content is at stake, the author of the article can keep the bracketed material very brief, knowing that it is comprehensively available in the databanks. Hence, the knowledge that the user of the *TLL* has access to digital resources can relieve contributors of the duty of documenting in full phenomena that accommodate purely formal definitions. With a little work a competent user can easily fill in gaps and omissions that are obvious or to which the author has drawn attention. It is therefore crucial for a potential user to study the Introduction to the *TLL*, in order to realize where investment of effort will reap the richest dividends.

Reducing the number of comparanda listed between brackets and parentheses creates more room to fulfill the central goal of the Thesaurus, formulated by Eduard Wölfflin: to provide the most nuanced and comprehensive picture of a word, its meaning (or meanings), and its history. To this goal belong details about the word's usage, development, transferred use, metaphorical meanings, associations (if any) with specialized jargon, etc. An article in the *TLL* gives guidance and orientation through the material that its author was able to look at. By means of its structure, it offers the user criteria and tools for analyzing other instances of the lemma that were left out or were unknown when the article was composed; hence, a new attestation could render the structure of an article partially obsolete or, at least, in need of supplementation. The view of a word sketched in the article is very often only one of several possible views; the author, through intensive engagement with the material, tries to define its characteristics and lay them out for the reader. An article in the *TLL* is the product of a thorough thought-process, not a matter of recording data. It demands of its users patience and thought on their part too, in order to make fruitful use of this nuanced instrument.