Essays of an Information Scientist, Vol:2, p.415-418, 1974-76 Current Contents, #5, p.5-8, February 2, 1976

"Current Comments"

Citation Analysis, Mechanical Translation of Chemical Nomenclature, and the Macrostructure of Science†

February 2, 1976 Number 5

The microstructure of science is very different from its macrostructure. For example, I can confidently assert that "milestone" papers—those which are subjectively rated as important by a large number of scientists—are, on average, frequently cited. However, I cannot truthfully assert that *every single* "milestone" paper is highly cited. A few may have been almost totally ignored, for a variety of reasons. In fact, some portions of my own work that I regard most highly have been least cited. Thus it is painfully apparent to me that models which are valid and reliable in the macrostructure of science can crumble when the focus is narrowed to the microstructure of science.

How extensive this phenomenon is we have not yet been able to determine. I wonder whether Watson and Crick would agree that their 1953 paper in *Nature*[3] represents the pinnacle of their work? I know that Oliver Lowry[4] correctly asserts that his most important papers are not his most cited. But that does not say his most important are not heavily cited. Keep in mind that I am not saying citation analysis cannot detect the significant though infrequently cited paper. Back in 1964 we produced a "computerized" history of DNA[5] which showed some papers that were infrequently cited but were significant in breaking the genetic code.

Examples of this kind have given me reason to question the assertion by Cole[6] that there is no validity in the Ortega hypothesis.[7] This theory asserts that advances in science depend in part on the contributions of mediocre scientists. While we may all stand on the shoulders of giants, they in turn depend upon many average or less eminent scientists. Whether they depend upon dwarfs is another question.

All this is leading up to a discussion of some work I did which is rarely cited but which gave me fantastic satisfaction. I refer to a paper on mechanical translation of chemical nomenclature.[8] This was the subject of my doctoral dissertation. Since I'm so often asked why, I'd like to tell you how I happened to take a degree in linguistics rather than library science.

I entered the field of documentation, now information science, from chemistry by joining the Johns Hopkins University Indexing Project in 1951. I stayed until its demise in 1953. By the middle of 1954 I had already accumulated a master's degree in library science and sufficient graduate credits to satisfy the minimum requirements for a Ph.D. But it proved impossible for me to find a faculty member at Columbia University who would approve for my dissertation topic the use of machine methods in scientific information. The only sympathetic ear was that of Professor Merrell Flood, but in order to take a degree with him, I would have had to take undergraduate training in industrial engineering. In retrospect, I see more clearly how relevant systems work has been in my career.

I tried to form an interdisciplinary faculty group, but I was not interested in spending ten years trying to satisfy an interfaculty group that would supervise my work. By that time my family had already been convinced I was going to be a student forever. I left Columbia disappointed. But in 1954, through my friend and colleague, Casimir Borkowski, I met Professor Zeliig Harris at the University of Pennsylvania, Department of Linguistics. His work in structural linguistics was already well known to scholars, but in the field of scientific information he was unknown. In 1956, I wrote a paper on the application of

†Reprinted from *J. Chem. Inform. & Comp. Sci.* 15(3):153-55, 1975.

structural linguistics to mechanized indexing[9] and showed it to Harris. Though it was never published, Harris became sufficiently interested in the field of information retrieval to accept some huge grants from NSF over a ten-year period. Most of this work is now continued primarily by Naomi Sager at New York University.[10] Some of you may recall transformational and discourse analysis.

I suppose it was prestige that made me seek a Ph.D. I ultimately worked out a doctoral program with Professor Harris which commenced officially in 1958. We had agreed on the amount of course work and my ultimate dissertation topic. By then I was quite preoccupied with problems of chemical indexing. We were encoding all new steroids for the U.S. patent Office under a contract with the Pharmaceutical Manufacturers Association.

By 1960, the Institute for Scientific Information (ISI) was publishing *Index Chemicus*. The original purpose of this service was to index compounds by molecular formula. So it was natural for me to want to find a way of calculating molecular formulas in the simplest way possible. Until that time *everyone* assumed that it was necessary to draw a structural diagram in order to calculate a molecular formula. Even Ascher Opler,[11] who wrote the pioneering paper in 1956 on "New Speed to Structural Searches", assumed this was the case. That is why he first wanted to represent the compound in a topological matrix which later was called a connectivity table.

My linguistic studies convinced me that the "meaning" of chemical nomenclature had to include enough information for calculating molecular formulas straight away. Otherwise, how could we do this so quickly in our heads for simple compounds? I told Professor Harris my theory and he accepted it as my doctoral thesis, the first in the new field of chemico-linguistics. Thanks to the recognition by Professor Allen Day of Penn's Chemistry Department that it was a nontrivial problem, the topic was agreed upon in the graduate school. However, before I could work on my dissertation, I had to prove my theory worked. If it did not, I would have to choose another topic, no matter how long I spent on the research.

Recognizing that the dictionary work alone might take me several years unless I got help, I proposed that the theory be proven with respect to acyclic compounds. During the next few years I got into the detailed problems of discourse analysis for my target language—chemical nomenclature. The details are not essential to this story. When I was ready for actual computer trials, I got the help of John O'Connor in programming Univac I, which was then in use at Penn. But I found that I could never get time on the computer, so I had to buy time at the Franklin Institute computer center.

The outcome of all this was "an algorithm for translating chemical nomenclature into molecular formulas."[12] When I submitted it to the department it was only ten pages. My substitute adviser was dumbfounded by this. Dissertations in linguistics are written by the pound—not the page. I spent a whole semester filling it out with interesting theoretical statements and formal analyses of chemical morphology, etc. By late 1960 I had made the first successful computer run in calculating a molecular formula directly from a systematic name.[8] I had done this manually hundreds of times earlier in the year.

As it turned out ISI was never able to finance the research necessary to complete this work. NSF was not very kindly disposed to us in those days. We also were up to our ears in the *Genetics Citation Index* project so I had to put chemical nomenclature work on the back burner. We never did input compound names for *Current Abstracts of Chemistry* (CAC); on the contrary, we now input Wiswesser Line Notation (WLN) for each compound and that is what we use to compute the molecular formula. However, the double bond checking routines that we used for so long were included in my algorithm.

About eight years ago I saw the proposal *Chemical Abstracts* made to NSF regarding chemical nomenclature translation research. Naturally I felt envious that they should get this support when it was clearly an operational development they needed more than ISI. That's what made it applied for them and academic for us.

However, I was very glad someone was

doing this and read with mixed feelings the first reports of this research in 1967.[13] A recent paper in the *Journal of Chemical Documentation*[14] shows that this work is finally coming to fruition, and I congratulate the CA group on their accomplishment.

Returning to the main point of my essay. Here is a topic of research which has multi-million dollar economic significance. There are only a few people in the world interested in it, so the number of times this kind of work will be cited is bound to be small. Clearly it is the kind of thing that is less cited than, e.g., papers on WLN, but there is an important connecting thread. Perhaps historians will decide that Opler's notion of a connectivity table for chemical compounds has been the most important concept in this field.[11] Most people seem to think that Sussenguth was the first one to use his concept.[15] But clearly none of these chemical information milestones has had any major discernible impact outside the field, and that is what the historian seeks and seems to find in large-scale citation analyses. This again demonstrates that the microstructure of science is very different from its macrostructure.

So much for the history of mechanical translation of nomenclature. Let me digress now to make some observations on the future of chemical and scientific publication. This has been much in the news these days, that is, *C&E News!* Joel Hildebrand, my freshman chemistry professor, has caused a lot of soul-searching with his rediscovery of the ancient idea of publication by abstract. I've had some contact with him in recent years and I know why he is making these proposals. Unlike James Stemmie who in *C&EN*[16] seems worried that some important ideas will be lost to posterity if we adopt any changed systems, Hildebrand is trying to tell us that the system is overloaded with useless information; he is talking about information pollution on a large scale. I have recently[17] asserted that the abuse of the page-charge system may be aggravating this pollution problem. And I regret to say *Chemical Abstracts* may be equally guilty. CA does this unwittingly in its hopeless aim to be complete. Consider that 25% of the abstracts in CA are of Russian material.[18] From our extensive citation analyses we know that this is absurd in relation to the significance of Russian research. They are polluting the waters of science with a lot of mediocre and unrefereed material. Probably another 10% of CA falls into this category. No doubt others do it too, but the data show clearly that the Russians are the worst offenders. Does anyone anywhere doubt the superiority of the *Journal of the American Chemical Society* over the *Zhurnal Obshchei Khimii*? How would you compare the abstracts of the ACS meetings to the abstracts of unpublished papers that the Russians are now loading into the *Russian Journal of Physical Chemistry*. Undoubtedly it gives the Russians significant political leverage to assert they account for 25% of CA's coverage. Maybe they will even claim CA should pay them a royalty for abstracting without their permission. After all, CA abstracts do constitute a substitute for the original Russian material.

There is an important distinction to be made between unrefereed material appearing in high-priced journals and unrefereed material listed in a depository. Each abstract requires the same space and work. But at least someone was willing to pay for that so-called high-priced journal. If librarians are as indiscriminate as they are accused of being, then why aren't they buying the original Russian journals and abstracts? I'm sure that Earl Coleman would be delighted if libraries bought his translation journals without the slightest evaluation. He knows how hard it is to sell the best that the Russians publish. He would court disaster to publish everything without regard to quality.

It is a rather interesting observation that 10% of CA's budget is about $2 million. If they cut back on Russian material they would find the same $2 million they want the Russians to pay for pirating CA.

At ISI we have very mixed feelings about CA. On the one hand, we resent their high price because a chemistry department is generally apt to say that it can't afford the *Science Citation Index* (SCI) but it must buy CA. If for no other reason, it couldn't get ACS accreditation without it. On the

other hand, the higher CA's price becomes, the more easily we can convince buyers that SCI or CAC is a good value. However, given my choice, I would much rather see CA priced lower. So I have a real concern for their cost-effectiveness. In fact, given my druthers, I would provide for CA a citation index to the chemical literature that would complement CA searches. The combined use of CA and SCI is happening increasingly, but it would be nice if we could accelerate the use of SCI by chemists as was suggested by the Hannay Committee many years ago.[19]

The recent paper by Parry, Linford, and Rich[1] shows a clear trend toward such complementary use of large data bases. This will increase as the cost of on-line services declines.

I recently did a search of the CA data base using our *Permuterm Subject Index* (PSI) to identify pertinent search terms and then followed up the output from CA by checking the items retrieved in the SCI! This is frequently done when people use MEDLINE and SCISEARCH,[2] but obviously the inclination to do so is tempered by the vast differences in per-hour rates.

In closing, I will mention miniprint, which has now come into the limelight. As the cost of paper goes up, CA and ISI may well have to adopt such methods. Whether users will accept miniprint more readily than microform is hard to determine, but there is a whole new technology opening up now that the "Oxford English Dictionary" has become so successful in this medium. Ralph Shaw and Albert Boni experimented with miniprint long ago. I just rediscovered it when I was thinking about ways to cut down on indexing costs. Maybe it's still not too late for CA to try it. After all, the most successful publishing venture of the past decade has been in the miniprint edition of the "Oxford English Dictionary".

1. **Parry A A, Linford R G & Rich J I.** Computer literature searches; a comparison of the performance of two commercial systems in an interdisciplinary subject. *Inf. Sci.* **8**:179-87, 1974.

2. **Garfield E.** ISI's *SCISEARCH* time-shared system trades time for money--but are you ready for this? *Current Contents®* (*CC®*) No. 40, 4 October 1972, p. 5-6.

3. **Watson J D & Crick F H C.** A structure for deoxyribose nucleic acid. *Nature* **171**:737, 1953.

4. **Lowry O.** Personal communication to D.J.D. Price, quoted in: Garfield E. Citation frequency as a measure of research activity and performance. *CC* No. 5, 31 Jan 1973, p. 5-7.

5. **Garfield E, Sher I H & Torpie R J.** *The use of citation data in writing the history of science.* (Philadelphia: Institute for Scientific Information, 1964), 86 pp.

6. **Cole J R & Cole S.** The Ortega hypothesis. *Science* **178**:368-75, 1972.

7. **Ortega y Gasset J.** *The revolt of the masses.* (New York: Norton, 1932), p. 84-85.

8. **Garfield E.** Chemico-linguistics; computer translation of chemical nomenclature. *Nature* **192**:192, 1961.

9. **Garfield E.** Proposal for research in mechanical indexing. Unpublished manuscript, 1956.

10. **Sager N.** Syntactic formatting of science information. *AFIPS Conf. Proc.* **41**: 791-800, 1972.

11. **Opler A & Norton T R.** New speed to structural searches. *Chem. Eng. News* **34**: 2812-14, 1956.

12. **Garfield E.** *An algorithm for translating chemical names to molecular formulas.* (Philadelphia: Institute for Scientific Information, 1961), 68 pp.

13. **VanderStouw G G, Naznitsky I & Rush J E.** Procedures for converting systematic names of organic compounds into atom-bond connection tables. *J. Chem. Doc.* **7**: 165-69, 1967.

14. **VanderStouw G G, Elliott P M & Isenberg A C.** Automatic conversion of chemical substance names to atom-bond connection tables. *J. Chem. Doc.* **14**: 185-93, 1974.

15. **Susenguth, E H.** Graph theoretic algorithm for matching chemical structures. *J. Chem. Doc.* **5**: 36-43, 1965.

16. **Stemmle J T.** Control of scientific papers. *Chem. Eng. News* **53**: 33-34, 1975.

17. **Garfield E.** Page charges; for profit and non-profit journals; and freedom of the scientific press. *CC* No. 7, 17 February 1975, p. 5-7.

18. **Baker D.** World's chemical literature continues to expand. *Chem. Eng. News* **49**: 37-40, 1971.

19. **Anonymous.** ACS report rates information system efficiency. *Chem. Eng. News* **47**: 45-46, 1969.