Current Comments

Automatic Indexing and the Linguistics Connection

Number 8 February 23, 1981

A few years ago, I described some of the difficulties in explaining to people that I am an information scientist.[1] The problem of describing how I make a living is only compounded when I mention that I obtained a doctorate in structural linguistics after having studied chemistry and library science. First of all, most people don't know what structural linguistics means. And even if they do, the connection between linguistics and information science is not at all obvious. The purpose of this essay is to make that connection more apparent. A recent article by Joseph Greenberg, Stanford University, describes the use of linguistic models in several other disciplines.[2]

It was by no means obvious 30 years ago that linguistics and information retrieval research shared common ground. A few theoreticians like Bar-Hillel may have been overtly aware of the connection. But linguists like Zellig Harris certainly were not. So it was only after two decades of a gradual evolution that Christine Montgomery could say, "Information science is concerned with all aspects of the communication of information, language is the primary medium for the communication of information, and linguistics is the study of language as a system for communicating information."[3]

In a talk I gave at the American Chemical Society in 1975 (which was published later that year [4]), I told some of the story about the difficulties I had in merging linguistics and chemical information science. My doctoral dissertation dealt with an algorithm for the computer translation of chemical nomenclature into molecular formulas.[5] Recently, I've described the application of linguistics to the machine translation of scientific texts.[6] But now I want to discuss how linguistic analysis is used by information scientists to develop methods for automatically indexing scientific texts. I'll use ISI®'s *Permuterm® Subject Index* (*PSI*) and *Key Word/Phrase Subject Index*™ (*KWPSI*™) as specific examples.

I became interested in linguistically based machine methods in information science even before I began the formal study of library science at Columbia University. But after acquiring a master's degree and enough credits to satisfy the basic requirements for a PhD, I still could not find a Columbia faculty member who would help me shepherd my proposed dissertation topic through a multidisciplinary faculty committee. As a consequence of this and economic considerations, I accepted a consulting assignment with Smith, Kline & French (SK&F) laboratories in Philadelphia.

My old friend Casimir Borkowski was already in Philadelphia and had known about my frustrations in trying to complete a dissertation on "Machine methods of scientific documentation." Cas and I shared an interest in mechanical

translation and similar problems in linguistics. By 1956, Cas was studying structural linguistics under Harris at the University of Pennsylvania. He introduced me to Harris, and over lunch we talked about my interests in information retrieval. I described to him the process of human, that is, cerebral, indexing of scientific papers. We agreed that structural linguistics was relevant to automatic analysis of scientific texts. And I suggested that he could receive support for such research from the National Science Foundation. Not much later, that in fact occurred.

In the summer of 1954, I left Columbia and moved to Philadelphia. I was able to keep up my contacts with Harris while I worked as a documentation consultant for SK&F and several other clients. In 1958, the same year Cas got his doctorate, I decided to try for a PhD in structural linguistics at the University of Pennsylvania. I worked out a deal with Harris to take one additional year of formal courses in linguistics combined with a reading program he would supervise.

During this time, I had a contract to index and code thousands of new steroids for the US Patent Office. From this experience and from earlier experience as an abstractor for *Chemical Abstracts*, I learned that the same chemical compound could be named in many different "dialects." From my first contact with chemical nomenclature at the Johns Hopkins University Indexing Program in 1951,[7] it was common to talk about the "language of chemistry." But no one had really given any serious consideration to the idea that chemistry, or its nomenclature, could be treated formally as a language.

Any "systematic" name of a chemical compound contains enough semantic information so that a chemist could draw its structural diagram. If this is true, then it certainly contains the even less information found in a molecular (empirical) formula. Therefore, it should be possible to compute the formula directly from the name. After Allen Day, professor of chemistry at Penn, agreed to serve as an additional faculty advisor, Harris gladly agreed to accept this problem as the basis of my PhD thesis.

By the end of the year, with programming help on Univac I from John O'Connor, I succeeded in generating a molecular formula from a systematic name on a computer for the first time. Of course, I'd tested it hundreds of times manually. Just at that time, however, Harris had gone on sabbatical leave. It took only ten pages to describe the theory behind the algorithm as well as the actual procedure. I'd been taught by my old boss Louis P. Hammett that the brief description of complex ideas was an ideal in science. But my substitute dissertation advisor wouldn't accept such a short manuscript as a PhD thesis! It was very annoying to have my manuscript judged on length instead of content.

However, I'd invested too much time and effort already to stop at this point. I went along with my advisor's recommendations to "fill it out." The result was a 110 page thesis on "An algorithm for translating chemical names to molecular formulas,"[5] which was approved in 1961. This was later reduced to 68 pages when set in type. The editor of *Nature*, however, was perfectly willing to have the ideas published in one page.[8] Readers interested in further details of my experiences in applying linguistics to chemical information science can refer to an earlier publication.[4]

All during this time and in later years I observed the research going on at Penn in the application of linguistics to information retrieval. That work and other linguistic research over the last 20 years has been dominated by transfor-

mational grammar (TG) theories. The original version of TG was developed by Harris in the early 1950s.[9] Noam Chomsky, Harris' student, developed his own version of TG a few years later.[10] It was Chomsky's version that became widely popular and much discussed in the field. One of Chomsky's important contributions was his idea that there is a "deep structure" on which all languages are based. This idea challenged the old structural linguistic claim that all languages are unique to themselves. John Lyons, University of Edinburgh, points out that "the effect of Chomsky's ideas has been phenomenal. It is hardly an exaggeration to say that there is no major theoretical issue in linguistics today that is debated in terms other than those in which he has chosen to define it, and every school of linguistics tends to define its position in relation to his."[11]

At the same time that Chomsky's influence spread among linguists and others, information scientists were pursuing a variety of theoretical and pragmatic approaches to automatic indexing and retrieval of information. And the field of mechanical translation was also quite hot due to the post-Sputnik interest in Soviet science.

Before we consider the various linguistic strategies they applied to machine indexing, it is useful to describe what *human* indexing involves. The human indexer analyzes the "natural language" of a document and tries to choose indexing terms that represent its main semantic content. Ideally, this derived "index language" should provide a description of text content that is so accurate that it is unnecessary for the index user to see the original paper to determine if it is relevant. Susan Artandi, Rutgers University, New Brunswick, New Jersey, says, "Indexing implies the understanding of the meaning of the text and the ability to make value judgments concerning its information content relative to the perceived interests of the expected information seekers."[12]

If computers are to translate natural language texts into a formal indexing language, they must be "taught" how to identify meaning in scientific texts. Anyone with experience in indexing knows this is an incredibly difficult task even under ideal circumstances. And if one looks at some of the work done on artificial intelligence one realizes, in a formal sense, how extraordinarily difficult the automatic indexing problem really is.

In general, research on automatic indexing has been guided by two linguistic strategies: syntactical analysis and semantic analysis. Syntactical analysis concentrates on the grammatical *structure* of sentences. Semantic analysis focuses on the *meaning* of sentences or words. Of course, syntactical and semantic analyses are not two exclusive strategies—elements of each are combined in many or most automated indexing systems.

Karen Sparck Jones, Cambridge University, England, explains that simple semantic strategies uncover the meaningful content of a text by analyzing word *frequencies*.[13] The computer identifies words as character strings separated by blanks. Words are ranked according to their frequency of occurrence, and index terms are derived from this list.[12] Usually, a suffix stripping dictionary is used so that the computer tabulates "molecules" as "molecule" or "retrieval" as "retrieve."

The most frequently occurring words on these lists are parts of speech that don't contribute much to the text's semantic content—prepositions, conjunctions, and articles, for example. Automated semantic analysis uses an "exclusion list" or "stop list" dictionary to eliminate the "dead weight" in

documents. Exclusion lists specify which words are to be excluded from processing either by grammatic function or by simple word-length.[12] Words like "compare" or "describe" can also be specified as "fluff" words to be excluded.[13] The output is a "simple list of all the non-trivial words in the original text"[13] ranked in order of frequency of occurrence.

The Keyword-in-Context (KWIC) indexing system, initiated by H.P. Luhn at IBM in 1958, uses a stop list to exclude such "obviously non-significant"[14] words as "report," "analysis," "theory," "method," etc., from article titles. However, this ignores the reality that some users may be interested in whether any methods or theories are discussed in the document, or whether it is an analysis or review. Indiscriminate exclusion of such words reduces the information content of the subject index.

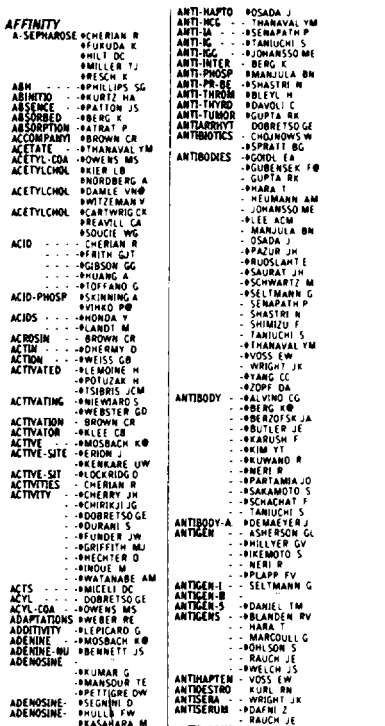
Also, KWIC indexes are bulky and cumbersome to use. After excluding all words matching those on the stop list, the remaining "significant" title words are rotated. For example, the title "Analyzing methods for protein determination by sephadex gel filtration: a review," may be reduced to "Protein determination sephadex gel filtration." This abbreviated title is "rotated" to be indexed under protein, determination, sephadex, gel, and filtration. Thus, if a title is reduced to five significant terms, the document will be indexed under all five terms in rotation. A modified version of this kind of simpleminded rotative indexing is used to produce the subject index to *Current Contents*® (*CC*®) each week.

When Irv Sher and I designed a subject word index to augment the citation and author indexes for the *Science Citation Index*® (*SCI*®), we kept in mind the shortcomings of full-stop lists and rotation indexing. The *Permuterm Subject Index* (*PSI*),[15] developed in 1964, uses a *small* full-stop list and a *semi-stop* list. The full stop list excludes prepositions, conjunctions, articles, and so on. The semi-stop list excludes words like "describe" or "method" from being primary index terms, but they are retained as secondary sub-entry terms.

Instead of just rotating them, *PSI* completely permutes title words to produce all possible pairs, including the inverse of all pairs. Thus, if a title is reduced to five significant words, 20 word pairs are generated—n(n-1), where n is the number of different significant words. When the index is printed, all word pairs are arranged in alphabetical order by primary term. Co-terms associated with the primary term are indented and arranged in alphabetical order under the primary term. The authors who used the word pairs in their titles are indicated by dashes leading from the indented co-term. (See Figure 1.) The full titles and bibliographic citation can then be located in the Source Index of the *Science Citation Index*, *Social Sciences Citation Index*®, and *Arts & Humanities Citation Index*™.

Also, two- or three-word phrases are statistically analyzed to determine frequency of occurrence. Compound terms that occur with great frequency, like "birth control," "guinea pig," and "*Escherichia coli*," are hyphenated and treated as *single* words. Thereafter, these and all other semantically useful compound terms are used to create word *phrases* by permuting all title words that occur with the hyphenated "word." This greatly reduces the volume of the *PSI* by decreasing the number of permutations. At the same time, this increases the specificity and retrieval speed of the *PSI*. ISI has created a dictionary of about 8,000 two- and three-word phrases that occur over a given statistical frequency threshold. The

Figure 1: Example of *Permuterm® Subject Index* (*PSI*) entry for "Affinity," taken from the *Science Citation Index®* (*SCI®*).



computer detects these compound terms in article titles and automatically lists them in hyphenated form.

Of course, ISI's "phrase" dictionary requires human intellectual effort to keep it current—new compound terms often meet and surpass the frequency threshold. This is particularly true of recently coined terms like "opiate-receptors." Also, we'd like to increase the average length of compound terms to enhance specificity and further reduce the size of the *PSI*. Unfortunately, compound terms consisting of four or more words don't occur frequently enough to warrant special treatment. But one could display additional terms with each two- or three-word phrase to make each entry more specific. An alternative strategy that would also generate longer and possibly more useful phrases is based on *syntactical* analysis.

The basic strategy in syntactical analysis is to *parse* sentences. That is, sentences (or titles) are broken down (parsed) into their component parts of speech and each component is described grammatically—noun, verb, adverb, adjective, and so on. The computer uses a number of grammatical clues to automatically recognize word sequences, depending on the component parts into which the sentence is parsed. For example, Borkowski identified case citations in legal texts by programming a computer to recognize "v." (as in John Public v. State).[16] On this simple parsing level, the component word phrases aren't characterized grammatically—they are simply identified as units containing potential index terms. Also, the phrases are still only two or three words in length.

In higher level parsing strategies, the computer is programmed to recognize punctuation marks, prepositions, or conjunctions as sentence "dividers."[13] Whatever occurs between these divisions is isolated as phrases. The computer then analyzes the *relation* between different phrases in a given sentence. Usually, the computer is programmed to consider *noun* phrases. Noun phrases can be characterized according to their function—subject, object, and/or modifier.[3] Or noun phrases can be related to the "verb environments" in which they appear.[13] In either case, the result is a list of "canonical components" which represent the logical relations linking noun phrases in the document.[13]

Sophisticated parsing strategies for automatic indexing sometimes rely on Harris' theory of string analysis.[9] Harris' theory provides for the "decomposition" of a sentence into several component strings. One of these strings is a "kernel

sentence" to which all other strings are directly or indirectly joined. These strings can then be transformed or "articulated" to produce syntactically equivalent phrases. For example, if you want to retrieve all documents on "information retrieval systems," the computer should recognize documents on "systems for the retrieval of information" as also being relevant to the search request.

In 1967, J.E. Armitage and Michael Lynch developed an algorithm which automatically articulates a single title-like phrase into several useful index phrases.[17] Based on this work, researchers at *Chemical Abstracts* recently developed techniques for processing natural language phrases to produce subject index entries for *CA*.[18] However, a human analyst had to precode the phrases before the computer processed them. We at ISI felt it was possible to generate index entries from *unprocessed* natural language titles.

In 1977, George Vladutz, now ISI's manager of basic research, suggested that syntactical analysis could be applied for this purpose. Our goal was to develop a *Key Word/Phrase Subject Index* (*KWPSI*)[19] that will be even more subject specific and compact than the *PSI*. In order to achieve this goal, we first had to break down a title into its component phrases in order to successfully provide entry points for an automatic indexing system. One possible approach was to apply parsing techniques.

At this stage, we visited New York University, which is funding a Linguistic String Project. The aim of the project is to develop methods for producing semantic representations of scientific text content. Naomi Sager, formerly associated with Harris at Penn, was kind enough to parse a sample of titles taken from ISI's data base to see if noun phrase identification would be a useful indexing strategy for automation. The results of the parses were very encouraging. But, as in any standard procedure of syntactical analysis, each word processed by the computer must be present in the system's dictionary already, along with appropriate morphological and syntactical information. While this might be possible for a particular specialty, the effort to update such a dictionary for our purposes would be prohibitive. ISI processes too broad a spectrum of information to enter every word we encounter.

Instead, Vladutz developed an algorithm that uses a smaller dictionary of words having syntactic function only—prepositions, conjunctions, articles, and so on. Ironically, this dictionary coincides with the list of stop and semi-stop words in the *PSI*. The dictionary is small because it is aimed at *titles* or title-like text. Scientific article titles have a relatively simple structure with a very limited number of verbs. So our syntactical analysis works quite well. Whether it would work on extended text remains to be tested.

Our procedure is called Multilevel Substring Analysis (MLSSA) because the product is four different substrings of the natural language titles we input. The substrings range from main word phrases in the title to the individual title words themselves. Each meaningful word in a substring is processed to produce syntactically equivalent variations. Meaningful words are identified as non-stop and non-semi-stop words. The substring variations have a large enough context around each meaningful word to be semantically self-contained. When *KWPSI* is printed, meaningful words are alphabetically sorted and all substring phrases associated with a given meaningful word are indented under it. (See Figure 2.)

Although the multilevel procedure takes twice as long to parse a title than the *PSI* takes to permute, *KWPSI* is

Figure 2: Example of *Key Word/Phrase Subject Index*™ (*KWPSI*™) entries as they would appear in *Quarterly Index to Current Contents®/Life Sciences* (*QUICC*™/*LS*).

*AFFINITY
ADSORBENT; PREPARATION and PROPERTIES of *(ISOLATION and PURIFICATION of BIO-POLYMERS. AFFINITY CHROMATOGRAPHY. POLYSACCHARIDE SHACER. PURIFICATION of PROTEOLYTIC-ENZYMES)* 40 056 0556
ALTERED * *(METHOTREXATE-RESISTANT CHINESE HAMSTER OVARY CELLS CONTAIN. DIHYDROFOLATE-REDUCTASE METHOTREXATE)* 40 078 4321
of AROMATIC and D-RING HALOGENATED ESTROGENS; SYNTHESIS and RECEPTOR-BINDING * *(ESTROGEN. RECEPTOR BASED IMAGING AGENTS)* 40 128 0994
CHROMATOGRAPHIC INTERACTIONS of PROTEASES *(LOW. MOLECULAR-WEIGHT SOYBEAN PROTEASE INHIBITORS)* 40 060 0385
CHROMATOGRAPHY *(ISOLATION and PURIFICATION of BIO-POLYMERS. PREPARATION and PROPERTIES of AFFINITY ADSORBENT. POLYSACCHARIDE SHACER. PURIFICATION of PROTEOLYTIC-ENZYMES)* 40 056 0556
CHROMATOGRAPHY *(SOLID SUPPORT. COVALENTLY BINDS THIOL GROUPS VIA. CLEAVABLE CONNECTOR ARM)* 40 087 0774
CHROMATOGRAPHY *(SUBSTRATE INDUCED DISSOCIATION of GLYCERALDEHYDE-PHOSPHATE DEHYDROGENASE DETECTED. STUDY of SUBUNIT INTERACTIONS. AFFINITY SORPTION)* 40 093 0285
CHROMATOGRAPHY of PORCINE PANCREAS DEOXYRIBONUCLEASE I *(DNA BINDING SEPHAROSE. NON-DIGESTIVE CONDITIONS. SUBSTRATE-BINDING SITE)* 40 070 0797
ELECTROPHORESIS; DOLICHOS-BIFLORUS PLANT using * *(STUDY of BINDING-PROPERTIES of ISOLECTINS)* 40 081 0237
of HEMOGLOBIN; OXYGEN MULTISTAGE REGENERATION PROCESS and INVIVO REDUCTION of *(LIPOSOMES of ALLOSTERIC EFFECTORS. OXYGEN)* 40 120 0502
LABEL; NEW * *(ADENOSINE 5'-(2-BROMOETHYL)-PHOSPHATE. ADENINE-NUCLEOTIDE SITES. PROTEINS)* 40 072 7517
of SIDE-CHAIN HALOGENATED HEXESTROL DERIVATIVES; SYNTHESIS and RECEPTOR-BINDING * *(ESTROGEN. RECEPTOR BASED IMAGING AGENTS)* 40 128 1002
SORPTION *(SUBSTRATE INDUCED DISSOCIATION of GLYCERALDEHYDE-PHOSPHATE DEHYDROGENASE DETECTED. AFFINITY CHROMATOGRAPHY. STUDY of SUBUNIT INTERACTIONS)* 40 093 0285
*AFFINITY-PURIFIED
GUANINE-NUCLEOTIDE REGULATORY PROTEIN *(RESTORATION of GUANINE NUCLEOTIDE-STIMULATED and FLUORIDE STIMULATED ACTIVITY. ADENYLATE CYCLASE-DEFICIENT CELL LINE)* 40 061 0439

smaller than *PSI* by between 25-40 percent. Also, if you compare Figures 1 and 2, *KWPSI* is more content-specific than *PSI*. As a printed index, *KWPSI* should be easier to use than *PSI*, and should retrieve the articles that are really relevant to one's interest. *KWPSI* may also be transferred to an online system. However, *KWPSI* does not have some of the generic searching advantages of *PSI* due to its format of precoordinated word pairs. And these advantages of *PSI* have yet to be built into any existing online system.

Although there is still work to be done on *KWPSI*, I believe we're making significant headway toward more responsive yet fully automated indexing systems. It should be obvious by now that linguistic research is closely related to this effort. The theoretical models of syntactic and semantic analyses, as well as the set of transformational grammar rules, were developed by linguists. Information scientists have now applied this theoretical linguistic research to the practical problem of automatic indexing. Borkowski reminded me of Gerry Salton's very useful work at Cornell University in automated indexing.[20] If we hope to extend automatic indexing techniques to process abstracts or full text, as well as titles, even more intensive linguistic research is needed. However, the *PSI* and *KWPSI* demonstrate that automatic syntactic and semantic analysis of article titles is more than adequate to produce informative and content-specific indexing terms.

This is not the place to discuss the advantages of such indexing in conventional and online systems. As the cost of computer time goes down one can seriously contemplate using methods of text analysis that would produce "deep" indexing or *a posteriori* indexing implied in the pioneering research of people like John O'Connor at Lehigh University.[21] Such procedures might even identify papers that report information on the toxicity of drugs even though the authors have never used such an expression to characterize the work. In the meantime we have to do a systematic and thorough job in dealing with the explicit words used by authors. It's the complementary task of citation indexes to deal with implicit or *a posteriori* meanings they attribute to the works they cite.

* * * * *

My thanks to Alfred Welljams-Dorof for his help in the preparation of this essay.

REFERENCES

1. **Garfield E.** What do you do for a living? *Current Contents* (6):5-7, 5 February 1979.
2. **Greenberg J H.** Types of linguistic models in other disciplines. *Proc. Amer. Phil. Soc.* 124:35-40, 1980.

3. **Montgomery C A.** Linguistics and information science. *J. Amer. Soc. Inform. Sci.* 2:195-219, 1972.
4. **Garfield E.** Citation analysis, mechanical translation of chemical nomenclature, and the macrostructure of science. *J. Chem. Inform. Compul. Sci.* 15:153-5, 1975.*
5. ----------. *An algorithm for translating chemical names to molecular formulas.* PhD dissertation, University of Pennsylvania, 1961.
6. ----------. Can machines be scientific translators? *Current Contents* (33):5-9, 18 August 1980.
7. **Himwich W A, Garfield E, Field H G, Whittock J M & Larkey S V.** *Final report on machine methods for information searching: Welch Medical Library Indexing Project.* Baltimore, MD: Johns Hopkins University, 1955. 38 p.
8. **Garfield E.** Chemico-linguistics: computer translation of chemical nomenclature. *Nature* 192:192, 1961.
9. **Harris Z S.** *Methods in structural linguistics.* Chicago, IL: University of Chicago Press, 1951. 384 p.
10. **Chomsky N.** *Syntactic structures.* The Hague: Mouton, 1957. 116 p.
11. **Lyons J.** Linguistics. *Encyclopedia Britannica.* Chicago: H.H. Benton, 1974. Vol. 10. p. 992-1013.
12. **Artandi S.** Machine indexing: linguistic and semiotic implications. *J. Amer. Soc. Inform. Sci.* 27:235-9, 1976.
13. **Sparck Jones K.** Automatic indexing. *J. Doc.* 30:393-432, 1974.
14. **Luhn H P.** Keyword-in-Context Index for technical literature. *Amer. Doc.* 11:288-95, 1960.
15. **Garfield E.** The *Permuterm Subject Index*: an autobiographical review. *J. Amer. Soc. Inform. Sci.* 27:288-91, 1976.*
16. **Borkowski C, Cepanec L, Sperling Martin J, Salko V & Treu S.** Structure and effectiveness of *The Citation Identifier*, an operational computer program for automatic identification of case citations in legal literature. *J. Amer. Soc. Inform. Sci.* 21:8-15, 1970.
17. **Armitage J E & Lynch M F.** Articulation in the generation of subject indexes by computer. *J. Chem. Doc.* 7:170-8, 1967.
18. **Cohen S M, Dayton D L & Salvador R.** Experimental algorithmic generation of articulated index entries from natural language phrases at Chemical Abstracts Service. *J. Chem. Inform. Comput. Sci.* 16:93-9, 1976.
19. **Vladutz G & Garfield E.** *KWPSI*—an algorithmically derived *Key Word/Phrase Subject Index.* *Proc. Amer. Soc. Inform. Sci.* 16:236-45, 1979.
20. **Salton G,** ed. *The SMART retrieval system: experiments in automatic document processing.* Englewood Cliffs, NJ: Prentice-Hall, 1971. 556 p.
21. **O'Connor J.** Automatic subject recognition in scientific papers: an empirical study. *J. Assn. Comput. Mach.* 12:490-515, 1965.

*Reprinted in: **Garfield E.** *Essays of an information scientist.* Philadelphia: ISI Press, 1980. 3 vols.

In *Current Contents®/Social & Behavioral Sciences* **12**(51):18, 22 December 1980. *Citation Classic.* **Wilensky H L.** The professionalization of everyone? *Amer. J. Sociol.* **70**:137-58, 1964.
The last sentence in the sixth paragraph should read: "What makes long training necessary and persuades the public of the mystery of the craft is both intellectual and practical knowing, some explicit (learned from books and demonstrations), some implicit (intuitive understanding acquired from supervised practice and observation)."

Reference 2 should read: **Galbraith J K.** *The new industrial state.* New York: New American Library, 1968. Chapter 25.
