Cross-Textual Cohesion and Coherence

Howard D. White
College of Information Science and Technology
Drexel University
Philadelphia PA, 19104

Theoretical Statement

Let me posit that linguistics, like physics, has four binding forces. The comparison with physics is only fanciful, and I am here interested only in the linguistic forces as they apply to published writings in English. In particular, I am interested in the weakest force, which I call intercohesion and intercoherence—that is, connections *between* or *across* texts produced by different people. These are modeled directly on connections *within* texts, where cohesion means acceptable linkage of sentences through surface structure, and coherence means acceptable linkage of the concepts and relations underlying the sentences (Beaugrande and Dressler, 1981). In the second section of my paper, I will say something about visualizing intercohesion and intercoherence.

BINDING FORCES

	Physics		***Linguistics***
1	Strong nuclear force	*Strongest*	Word uninterruptability (binds morphemes into words)
2	Electromagnetism	*Less strong*	Grammar (binds words into sentences)
3	Weak nuclear force	*Less strong*	Texture/cohesion; coherence (binds sentences into texts)
4	Gravity	*Weakest*	Intercohesion; intercoherence (binds texts into literatures)

Mainstream linguistics traditionally deals with Forces 1 and 2, while discourse analysis and text linguistics are particularly concerned with Force 3. The field most identified with the study of Force 4 is information science.

As a speaker and writer of English, I have internalized the rules and principles of the first three forces; they are primary parts of my competence in the language. I can form old and new words correctly; I can produce grammatical sentences, and I can produce prose with texture. I use "texture" to mean "cohesion" in the sense of Halliday and Hasan (1976)—that is, I know how to signal the connectedness of sentences with explicit markers, such as proforms, verbal substitutions, conjunctions, and repeated vocabulary; and I try to undergird the text with an implicit connectedness of thought that can be followed even when explicit markers are absent—in other words, I try to make the prose coherent. Moreover, you can do all this, too. And because we both can do it, we will agree with a very high degree of regularity that a word is or is not well formed (strongest force) or that a sentence is or is not grammatical (second strongest force). We may lack similarly reliable intuitions of texture (third strongest force); nothing as definite as our sense of correct morphology or syntax helps us decide whether a piece of discourse is ideally well organized. As Halliday and Hasan put it (p. 293), "A text is best thought of not as a grammatical unit at all, but rather as a unit of a different kind: a semantic unit. The unity that it has is a unity of meaning in context, a texture that expresses the fact that it relates as a whole to the environment in which it is placed." But, as everyone who cares about writing knows, it is quite possible to reach high levels of agreement on steps to *improve* texture once they are proposed. It is also quite possible to agree that one draft of a piece is better than another on grounds of improved cohesion or coherence or both.

Force 4 differs from the others in that it is much less embodied in the linguistic competence of individual persons. I cannot extend my sense of well-formedness above the sentence level, and I cannot use texture-creating resources—e.g., anaphoric reference or ellipsis or clausal substitution—beyond the level of my own individual texts; I cannot produce other people's sentences. Force 4 relations are in fact statistical in nature; they emerge across many pieces of discourse over which no single author has control (just as one can be a driver but not traffic). Of course, like everyone else, I have resources for intercohesive writing. Some I may exploit consciously; others, I may benefit from unawares (as when I unintentionally quote or cite the same sources as other writers). But even when I write with knowledge of other texts and deliberately manifest that knowledge through all the means at my disposal, there are almost surely multiple texts cohering with mine that I do not know of, and they, like mine, could well be considered "part of the literature" of my topic. The same is true for any writer. So our intuitions of what constitutes

a literature, based on verbal co-occurrences we have encountered across texts and on the inferences we draw from them, are more idiosyncratic than our intuitions of grammar or even texture, and we are less likely to agree on them. Studies of inter-indexer consistency in information science routinely show only partial agreement in indexers' assignments of terms. If you and I and the indexers lack "text grammars" (internal rules for connecting sentences in texts), we lack "literature grammars" (internal rules for connecting writings in literatures) even more. (Unlike the *Anglo-American Cataloging Rules*, which tell librarians what descriptive elements, such as author and title, to copy from the front matter of texts, a complementary volume called the *Anglo-American Rules for Subject Cataloging and Classification* has never existed and never will, because no one can reduce the divination of subject matter to a formula.) This is not to say that Force 4 does not exist, merely that it is the weakest of four powerful forces.

Of Halliday and Hasan's (1976) list of resources for creating texture—that is, for binding sentences into text through surface-structural ties—the only one that extends *across* texts is what they call "lexical cohesion" or common vocabulary. The key event for intercohesion is the duplication and repetition of major content-bearing terms in two separate texts—more generally, the appearance of the same words (or synonyms, near-synonyms, subordinate, or superordinate terms that replace them). To achieve intercoherence, what the terms *refer to* in the world must be at least roughly the same. When both conditions are met, we can say that writings are "topically related" or "more or less on the same subject. " (Sameness of subject may or may not be the writer's intention.) Authors can also intentionally create explicit ties with other texts by alluding to or quoting from them or by citing them in footnotes or endnotes. Other indicators of intercohesion are added by editors, who associate individual texts with names that span whole subject domains, such as the titles of journals or of monographic series. Either authors or editors may signal broad functional similarities among texts by adding explicit genre markers (e.g., "novel," "letter," "literature review") to the surface structure of presentations. (Below surface structure, many connections besides those of common subject matter or genre await discovery.)

Common vocabulary and citation across texts are the bases of, respectively, natural-language indexing and citation indexing. Because authors work within and not between writings and no controlling intelligence is in force across all texts, there is always a chance for two kinds of linguistic mishap. The first is false cohesion, arising when texts share superficially cohesive word-forms but differ in what the words refer to. This is the problem of terms with multiple senses or homonyms (e.g., in natural-language indexing, "bond" meaning a financial instrument vs. "bond" meaning an adhesive; in citation indexing, "WHITE HD" meaning two different authors, both named H. D. White). The second is false incoherence, arising when words refer to essentially the same thing, but that fact is masked by incohesive word-forms. This is the problem of synonyms (e.g., in natural-language indexing, "lawyer" vs. "attorney," meaning the same occupation; in citation indexing "SPARCKJONES K" vs. "JONES KS," meaning the same author cited in two different ways). Usually such problems can be resolved if enough accompanying text is supplied, but professional indexers have long sought to eliminate the vagaries of authors' natural language through terse additions of their own. Indexers create intercohesion and intercoherence across texts by controlling the vocabulary with which they name the subjects of writings in bibliographies (i.e., the vocabulary of subject headings, descriptors, or classification schedules). They also control variations in how authors might be named in bibliographies with authority lists. The intent is to disentangle writings that accidents of language have falsely joined and to unite writings that accidents of language have falsely separated. So indexers join authors and editors in contributing to the connectedness of texts in literatures.

One must note, however, that even when their subject matter is similar, different authors invariably express themselves differently. Therefore, except in cases of quotation, near paraphrase, or, rarely, plagiarism, it is not authors' sentences, let alone their passages, that are duplicated and repeated across texts. The only things that recur with reasonably stable meaning are their nouns and noun phrases, a mere subset of language. Strictly speaking, what recurs across texts are tokens of word types—words as words. (They could be put in quotation marks or italicized to show they have been nominalized to name themselves.) Force 4 thus reduces language to nothing but noun-phrase indexing terms, and with only noun phrases to work with, one cannot predicate. This means that, while Forces 1 through 3 can jointly be used to make statements about the world, Force 4 language can only be used to make metalinguistic statements about writings, on the order of "such and such a term occurs in such and such a context." If the only language everyone shares is noun phrases as they occur across texts, then we all are limited to communicating like indexers, which most people would regard as an impoverishment. This decline in potential is probably why linguists leave Force 4 pretty much to information science; for example, Levinson's (1983) discussion of "discourse deixis" does not extend to the indexing of literatures even though "deixis" and "indexing" have the same root.

The occurrences of any one noun phrase (e.g., "children") across texts can be counted, and so can the co-occurrences of any two noun phrases (e.g., "children" and "handguns"). In place of predication, Force 4 has co-occurrence counts across texts—repeated associations of noun phrases, common lexis. We cannot know exactly what is being predicated about what until we actually do the reading, but we may have a good idea, based on our world knowledge and our inferential abilities. Co-occurrence places any one of a pair of terms in the context of the other, which sharply constrains meaning (especially if they appear together in a relatively narrow "text window," such as a paragraph). Any pair of terms—e.g., "cotton batting" and "water table"—could co-occur in the same text once, but without some special purpose we would not usually take much interest in single or infrequent co-occurrences of this sort. It is only when counts are increasingly high across texts that we know a literature has formed—for example, one held together by the co-occurrence of "children" and "handguns" (and variations such as "juvenile," "boy," "six-year-old," "revolver," ".357 Magnum") in writing after writing. The explicit appearance of such terms makes for intercohesion of texts. In matters with which we are familiar, we can also make shrewd guesses about their intercoherence—the relations of the underlying concepts—even without reading. It is likely that the literature suggested here deals with children's dangerous access to handguns or something along that line; we do not have to be told that it is not about, say, children and handguns being of different shapes and colors.

Suppose we seek to answer specific questions with passages from writings, or merely want to know what has been said on some topic of interest. In that case, we obviously need to conduct a literature search. Written texts of many kinds multiply quickly enough that individual items one might want to read or consult are lost in the multitude unless they can be quickly found and delivered in response to queries. In practice this means exploiting the intercohesion and intercoherence of texts, since we may want to retrieve explicitly or implicitly connected writings whose unique identities are not foreknown to us. But this in turn means that the texts of our queries must be intercohesive and intercoherent with the texts of the writings we seek (or with texts of representations of those writings). We need matches between our query terms and the bundles of indexing terms representing literatures. Both our query terms and the indexing terms can be of various kinds—natural language phrases that may appear in titles, abstracts, or full text; authors' names; controlled vocabulary descriptors; abbreviations of cited documents that retrieve citing documents, and so on—whatever occurs to us or we can find through look-ups. The following table shows what happens as a result of the matches. It reproduces a table long familiar in information science—the one for figuring recall and precision ratios on the basis of whether relevant documents are retrieved—but improves it by showing *why* the outcomes in the cells occur. Inter*cohesion* of query terms and index terms retrieves, but inter*coherence* of query terms and index terms is what we really want. Counts of documents meeting the four combinations of conditions would go into the cells:

	Intercoherent	Not Intercoherent
Intercohesive	a Hits	b False drops (false positives)
Not intercohesive	c Misses (false negatives)	d Correctly rejected

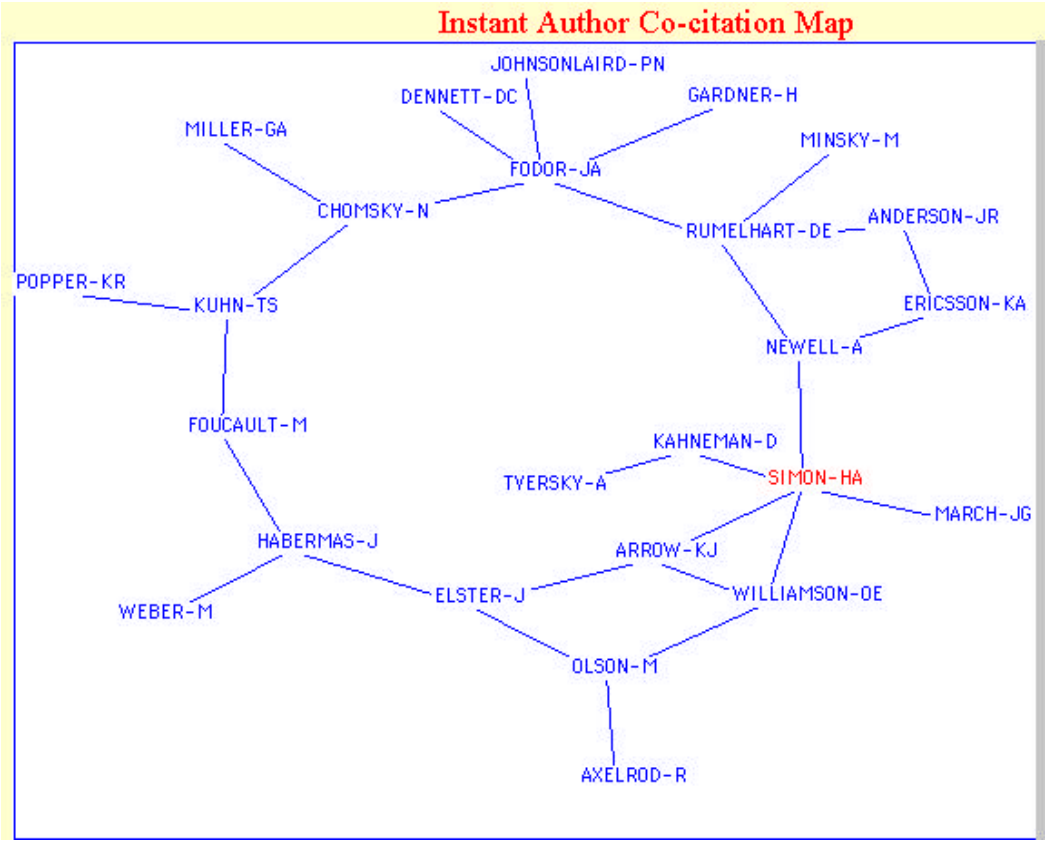
A brief example: if I want newspaper stories on "children's dangerous access to handguns" and I search on "children" and "handguns," any stories that contain those terms in sentences coherent with my interest are hits. Stories containing those terms in ways not coherent with my interest (e.g., an obituary for a collector of *handguns* survived by six *children*) are false drops. Stories I fail to retrieve because my language mismatches theirs (e.g., one about a boy of nine who kills his younger brother with a .357 Magnum, but not mentioning *children* or *handguns)* are misses. The rest I would rightly not want to retrieve.

Visualizing Intercohesion and Intercoherence

Attempts to match query terms and indexing terms are often portrayed as a "dialog" between persons and literatures, especially if computerized systems are used to mediate the transactions. The dialog metaphor is not farfetched. Between person and system there really is a give-and-take resembling that between speaker and hearer, as many have noted. There are utterances (sentences in a context) on both sides—at least one imperative or interrogative prompt by the user and at least one response by the system; the chain of such pairings can grow quite long. The difference, of course, is that this conversational exchange is between a person and a nonhuman auxiliary, a thing. Any term-matching system is designed and maintained by human beings; it partakes of their rationality.

But it lacks intelligence, except for what can be built into its algorithms before persons interact with it and what those persons can supply as the interaction takes place. The person with, say, *average memory, average understanding* augments his or her powers through contact with an entity that has, in effect, *infinite memory, zero understanding*. Interaction with it is most often a solo task (for which a pragmatics of self-service is needed); the person attempts to compensate for the system's ignorance by being not so much its hearer as its editor. The system, that is, shows the person the results of the term-matching and then mindlessly awaits their rejection, revision, or acceptance. That immediately removes what is studied in information science from the types of interpersonal dialogs (including those mediated by computer) studied in several subfields of linguistics.

Until quite recently, dialogs between persons and literature-based information systems were verbal only, but now certain textual relations may be visualized. At Drexel University, Dr. Xia Lin, Jan Buzydlowski, and I have devised a system for visualizing several kinds of intercohesion among documents. One implementation is called AuthorLink (White et al., 2001); let me here simply sketch its relevance to Force 4. AuthorLink is connected to the full 1.26 million bibliographic records from *Arts & Humanities Citation Index* for 1988-1997. When it is given an author's name, it responds with the 24 other authors most often co-cited in humanities journals with the entry author. Clicking a button produces an instant Pathfinder Network (PFNET) map of the most salient co-citation relations among the 25. The map links only those author-pairs with the highest co-citation counts after all 25(24)/2=300 nonduplicative pairs have been considered. To illustrate, the PFNET below is an AuthorLink response in a dialog begun by entering the name of the late polymath Herbert A. Simon. In the big matrix of co-citation counts formed by the AuthorLink program, all of the authors have been co-cited with Simon at least 20 times—in effect, we are looking at "Simon studies" in the humanities—but the map shows that only five of them (Kahneman, Arrow, Williamson, March, and Newell) have their *highest* co-citation counts with him; the highest co-citation count in the matrix for, say, Max Weber is with Jurgen Habermas. (The surname-initials format of names is dictated by the data.)



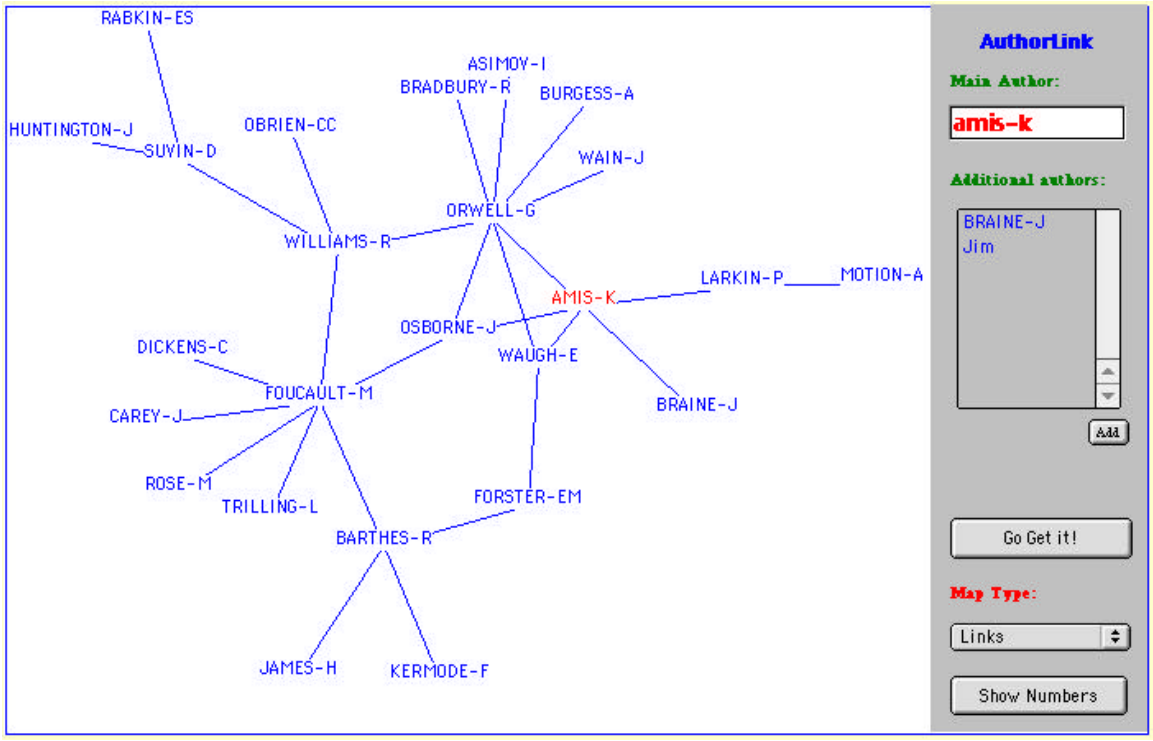
In line with my earlier discussion, note that the map links noun phrases, in this case authors' names that stand for *oeuvres*. The links attest to their intercohesion—that is, various pairs of names on the map, such as Simon and Arrow, are word tokens that explicitly co-occur in not just one or two but many texts. (An option in AuthorLink allows the actual co-occurrence counts to be displayed.) We do not know what is being predicated when the

citations occur, nor how close together they are in the texts, nor what actual works are being cited. We merely know that AuthorLink says that certain authors' works—e.g., those of Simon and Arrow—are repeatedly seen as relevant to the works of the authors who cite them, which implies they are relevant to each other. It is entirely possible, even likely, that we would not find explicit statements of *how* they are relevant to each other in the citing documents. Those we would have to infer, and they might well differ from text to text. But before we descend to that level of detail, it is easy to see Force 4 at work at the relatively high level of the map itself—that is, it is easy to read *intercoherence* as well as *intercohesion* into it if one is noddingly familiar with the disciplines in which Simon's work is used. The test for this I used earlier with "children" and "handguns." It is whether we can instantly infer what is being predicated to make the two terms intercohere (as we cannot with, say, "cotton batting" and "water table"). Obviously my account of "children" and "handguns" did not come from purely *linguistic* knowledge; it is not, for example, like my knowledge that "pistol" and "revolver" are synonyms for "handgun"; the words "children" and "handguns" are in no way substitutable for each other. It is, rather, world knowledge; the two words often co-occur in news stories because, in the world, a whole lotta kids are shooting a whole lotta kids, and this development gets publicly discussed. In the same way, when a linguistics professor saw the Simon map, he immediately inferred that Thomas Kuhn and Noam Chomsky were linked because many academics have used Kuhn's notion of "paradigm shift" to discuss Chomsky's revolutionary transformation of the field of linguistics in the 1950s and 1960s. I agree with him that this is very plausible. And even if Chomsky and Kuhn are co-cited in "Simon studies" for other reasons—in other word, *even if we are wrong*—the fact that we both jump to the same conclusion without having seen any evidence means that Force 4 is with us.

Where literature searches are concerned, we need not confine ourselves merely to inference, of course; we can retrieve writings in which various terms co-occur and see whether they bear us out. This can in fact be done with AuthorLink, which is not only a system for mapping author co-citation data on the fly but also a live interface for retrieving bibliographic data on the articles that are doing the co-citing. When one or more of the names in the map is clicked upon, it is automatically ANDed with Simon's name for a search. I have enough dim knowledge of Simon's intellectual worlds to be able to predict what his name implies when it is ANDed with some of the other names on the map. I know, for instance, that Simon is one of the founders of cognitive science , which has close connections with artificial intelligence. (Howard Gardner describes this tie in *The Mind's New Science,* and his is one of the names on the map.) If any name among the 25 connotes artificial intelligence, it is Marvin Minsky's; and, sure enough, when I retrieve the articles from humanities journals that cite Simon with Minsky, I see titles like "Artificial human nature; design, artificial intelligence, computers," "The functional architecture of adaptive cognitive systems with limited capacity," "Artificial intelligence and sign theory," "The computer as the artist's alter ego," "Cognition of systems of artificial intelligence," and "The prospects for building truly intelligent machines." But Simon was also an economist (he and Arrow are both Nobel laureates in economics); the cluster of names below his, from Elster through March, connotes (to me) the intersection of economics, organization theory, and theories of rational behavior. When I retrieve articles that co-cite Simon and Oliver E. Williamson (whose work I do not know at all), I see such titles as "The economics and politics of regulation," "Instrumental stakeholder theory; a synthesis of ethics and economics," "Contested exchange versus the governance of contractual relations," "Power and wealth in a competitive capitalist economy," "Small firms in economic theory," and "The transaction costs and benefits of the incomplete contract of employment," along with histories of particular economic enterprises. Thus, names like Minsky's and Williamson's are constraining the manifold implications of Simon's name in fairly intelligible ways.

Drilling deeper—that is, actually reading the co-citing articles and noting the contexts in which Simon's and other co-cited authors' works are invoked—would probably create a much more fragmented impression. Co-citing documents that tightly link Simon's name with someone else's in a single assertion, a particular knowledge claim, would most likely be quite rare. But that does not nullify the broad "rightness" of our intuitions about reasons for linkage at the more global level. The intercoherence we sense among AuthorLink names is the *subject* force that, among other things, holds specialties and disciplines together—Force 4. It emerges because of strong intercohesion among words, as seen in their co-occurrence counts. That, interestingly, is quite easy to visualize through software.

The analysis of intercoherence can be taken a step further in AuthorLink. In the next figure, I show a bit more of the interface along with a different PFNET map, this one for Kingsley Amis, the British novelist. Again, one has to know something of what co-cited authors' names connote in order to interpret the map readily. But since I am familiar with many of Amis's books, I can immediately predict why he appears with certain writers. For example, on seeing the PFNET, I am fairly sure that he is co-cited with George Orwell, Ray Bradbury, and Isaac Asimov because he wrote *New Maps of Hell,* a critical study of modern science fiction, especially dystopias. I am also



fairly sure that that is *not* why he is co-cited with authors like John Braine, John Osborne, or John Wain. They would be associated with Amis because they were all lumped together in the 1950s and 1960s as the "angry young men" of post-WWII British letters. Amis's novel *Lucky Jim* and Braine's novel *Room at the Top* would be mentioned in any history of the period; both were made into movies, and both still attract readers. Knowing such things, I can engage AuthorLink in a (very limited) dialog to see whether the co-citing articles confirm my guess. Since this is Amis's map, he is automatically entered as "Main Author" at top right. I have clicked on the label "Braine-J" to place Braine's name in the search box labeled "Additional authors." When the "Go Get It!" button is clicked, Braine and Amis will be automatically ANDed so as to retrieve the articles that co-cite them. But I have also used the small "Add" button below the search box to let me enter the term "Jim" in the box with Braine's name. I am predicting that at least one of the co-citing articles will specifically cite Amis's *Lucky Jim* along with something by Braine. If I am wrong about ANDing in "Jim," no articles will be retrieved, because my final condition will not be met. Thus, I am betting on a specific piece of intercohesion. After doing the search, AuthorLink does not exclaim, "Oh, Master, what a clever surmise!" but the result is the same. As it turns out, all four of the articles that cite both Amis and Braine cite both *Lucky Jim* and *Room at the Top.* The same exercise with "Hell" added to Amis as main author and Orwell, Bradbury, and Asimov as additional authors produces two articles, both citing *New Maps of Hell* in the context of dystopian works such as *1984* and *Fahrenheit 451*.

The retrievals with both the Amis map and the Simon map show intercoherence underlying intercohesion. There is, however, a key difference. With the Simon map, I *recognized* that the titles of co-citing articles seemed coherent across broad subject areas like artificial intelligence and economics. But I did not say in advance which of Simon's works would be co-cited with Minsky's or Williamson's. With the Amis map, because I know more about the subject matter, I *predicted* the very works that would be intercohesive, and my predictions were confirmed by the search results. I take this as evidence of intercoherence under Force 4. AuthorLink can be used to study it further.

References

Beaugrande, Robert de, and Wolfgang Dressler. 1981. *Introduction to Text Linguistics.* London: Longman
Halliday, M. A. K., and Ruqaiya Hasan. 1976. *Cohesion in English.* London: Longman.
Levinson, Stephen C. 1983. *Pragmatics.* Cambridge, England: Cambridge University Press.
White, Howard D., Xia Lin, and Jan Buzydlowski. 2001. The Endless Gallery: Visualizing Authors' Citation Images in the Humanities. *Proceedings,* ASIS&T Annual Meeting, November 2-8, Washington, D.C. 182-189.

Dr. Howard D. White
Field: Information Science

After taking his PhD in librarianship at the University of California, Berkeley, in 1974, Professor Howard D. White joined Drexel University's College of Information Science and Technology. He co-authored *For Information Specialists: Interpretations of Reference and Bibliographic Work* (Ablex, 1992) with Marcia J. Bates and Patrick Wilson, and his latest book is *Brief Tests of Collection Strength* (Greenwood, 1995). He has also published on bibliometrics and co-citation analysis, evaluation of reference services, expert systems for reference work, innovative online searching, social science data archives, library publicity, American attitudes toward library censorship, and literature retrieval for meta-analysis and interdisciplinary studies. In 1993 he won the Research Award of the American Society for Information Science for diverse distinguished contributions in his field. In 1998 he and Katherine McCain won the best JASIS paper award for Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972-1995. He is a Drexel Distinguished Professor for 1998-2002. More at: http://www.cis.drexel.edu/faculty/HUD.Web/HUD.html

Publications on AuthorLink

White, Howard D., Xia Lin, Jan Buzydlowski. (2001).
The Endless Gallery: Visualizing Authors' Citation Images in the Humanities. Proceedings, ASIS&T Annual Meeting, November 2-8, Washington, D.C. 182-189.

Lin, Xia, Howard D. White, Jan Buzydlowski. (2001). Associative Searching and Visualization. The Book of Abstracts, SSGRR 2001, International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet. August 6-12, L'Aquila, Italy. 16 (full text on cd-rom).

Buzydlowski, Jan B., Howard D. White, Xia Lin. (2001).
Term Co-occurrence Analysis as an Interface for Digital Libraries. Proceedings, Visual Interfaces to Digital Libraries, a Workshop of the First ACM + IEEE Joint Conference on Digital Libraries, June 28, Roanoke, VA. http://vw.indiana.edu/visual01/buzydlowski-et-al.pdf

White, Howard D., Xia Lin, Jan Buzydlowski. (2001).
Co-Cited Author Maps as Real-Time Interfaces for Web-Based Document Retrieval in the Humanities. Conference Abstracts, ACH/ALLC 2001. 2001 Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing. June 13-17, New York City, NY. 127-129. http://www.nyu.edu/its/humanities/ach_allc2001/papers/white/

Lin, Xia, Howard D. White, Jan Buzydlowski. (2001).
AuthorLink: Instant Author Co-Citation Mapping for Online Searching. National Online Proceedings 2001, May 15-17, New York City, NY. Medford, NJ: Information Today. 233-241.
http://faculty.cis.drexel.edu/~xlin/presentations/NationalOnline2001.ppt

White, Howard D., Jan Buzydlowski, Xia Lin. (2000).
Co-Cited Author Maps as Interfaces to Digital Libraries: Designing Pathfinder Networks in the Humanities. Information Visualization 2000, IEEE International Conference on Information Visualization, July 19-21, London, England. Los Alamitos, CA: IEEE Computer Society. 25-30.
http://www.computer.org/Proceedings/iv/0743/07430025abs.htm

Lin, Xia, Jan Buzydlowski, Howard D. White. (2000).
An Interactive System for Co-Citation Visualization. Proceedings, Annual Meeting of the Classification Society of North America, June 8-11, Montreal, Canada.
http://www.gerad.ca/CSNA/A018.php3

Buzydlowski, Jan, Xia Lin, Howard D. White. (2000).
Comparison of Graphical Techniques for the Presentation of Co-Occurrence Data. Proceedings, Annual Meeting of the Classification Society of North America. June 8-11, Montreal, Canada.
http://www.gerad.ca/CSNA/A019.php3