Exploring interdisciplinary relationships between linguistics and information retrieval from the 1960s to today

Engerer, Volkmar Paul

Published in:
Journal of the Association for Information Science and Technology

DOI:
10.1002/asi.23684

Publication date:
2017

Exploring Interdisciplinary Relationships Between Linguistics and Information Retrieval From the 1960s to Today

Volkmar Engerer
Royal School of Library and Information Science, University of Copenhagen, Fredrik Bajers Vej 7 K, DK-9220 Aalborg Ø, Denmark. E-mail: volkmar.engerer@hum.ku.dk

This article explores how linguistics has influenced information retrieval (IR) and attempts to explain the impact of linguistics through an analysis of internal developments in information science generally, and IR in particular. It notes that information science/IR has been evolving from a case science into a fully fledged, "disciplined"/disciplinary science. The article establishes correspondences between linguistics and information science/IR using the three established IR paradigms—physical, cognitive, and computational—as a frame of reference. The current relationship between information science/IR and linguistics is elucidated through discussion of some recent information science publications dealing with linguistic topics and a novel technique, "keyword collocation analysis," is introduced. Insights from interdisciplinarity research and case theory are also discussed. It is demonstrated that the three stages of interdisciplinarity, namely multidisciplinarity, interdisciplinarity (in the narrow sense), and transdisciplinarity, can be linked to different phases of the information science/IR-linguistics relationship and connected to different ways of using linguistic theory in information science and IR.

Introduction: Information Science, Information Retrieval, and Linguistics

The history of the relationship between linguistics and information science (or more specifically, information retrieval [IR]) has yet to be written from either perspective. The first attempts to build bridges between the disciplines were made by information scientists, presumably because of the contrast between the traditionally high status of linguistics as an academic discipline and the young, emerging field of information science in the 1940s and 1950s. However, there are also noninstitutional, qualitative reasons for these one-sided overtures; reasons related to the content, objectives and objects of study in the two disciplines. This is illustrated by a discussion of the work of selected scientists.

Linguists have only occasionally initiated collaborations, and it is perhaps not surprising because of the evident affiliations between text and document in linguistics and information science that most such collaborations stem from the prime of text linguistics in the 1970s and were instigated by researchers such as Petöfi (Petöfi, 1969; Petöfi & Bredemeier, 1977) and van Dijk (e.g., van Dijk's contribution in Walker, Karlgren, & Kay, 1977). Although I am a linguistic specialist, in this article I attempt to take the information scientist's perspective. I trace how information science and IR have absorbed linguistic ideas and theories into specific paradigms over the past 50 years and, importantly, why the process took the specific form it did (see also Engerer, 2012).

Tredinnick (2006) provided an important perspective on the development of information science as a hybrid science-humanities discipline. According to Tredinnick, information science was—and still is—trapped because of its self-image as a science (in the sense of the German term *Naturwissenschaft* [natural science]) and its permanent—and in Tredinnick's eyes, futile—struggle to come to grips with the intrinsic nature of "meaning" in information. Tredinnick (2006, p. 63) argued that information science's persistent doubts about its status as a science and the consequent emphasis on use of scientific methodology in the tradition of Popper and logical positivism was the main obstacle to a robust integration of meaning phenomena into information science. This conundrum, which is one of the major points of departure for linguistic thinking in information science, is particularly evident in IR.

Tredinnick argued convincingly that the tension between the "hard" scientific methodological basis of information science and retrieval and the "soft" nature of meaning, a core, intrinsic aspect of its object of study, was the driving force behind internal developments in information science. In this article I argue that the same fundamental tension between science and meaning also affects the relationship between information science and other disciplines. For a

Received July 9, 2014; revised October 18, 2015; accepted October 19, 2015

 DOI: 10.1002/asi.23684

long time linguistics has been considered as a discipline, which, because it involves the study of language and communication, naturally encompasses meaning and semantic phenomena. It therefore seems an obvious interdisciplinary partner for information science and IR. From the information scientist's perspective collaborations with linguistics researchers should enable information science to tackle issues relating to the meaning of information, while retaining its scientific methodology and status as a fully fledged science. The following sections demonstrate that these hopes have not been completely fulfilled.

How does linguistics relate to information science and to IR in particular? Phrasing the question this way implies that linguistics and information science and retrieval are unitary disciplines, which is clearly an oversimplification. Linguistics is a discipline with many schools, as a cursory inspection of modern textbooks reveals (see, for instance, Aronoff & Rees-Miller, 2007). Although there are some fields of linguistics that are arguably not directly relevant to IR (e.g., phonetics and phonology), others such as morphology (structure of words), sociolinguistics (the social dimension of language structure and use), and discourse analysis (the interactional structure of conversation) are more closely related to the problems of IR (for an overview see, again, Aronoff & Rees-Miller, 2007). In what follows I try to be as specific as possible in linking my claims to the relevant linguistic domains.

I have retained the generic terms *linguistics* and *linguistic* where appropriate. I use the general noun "linguistics" to refer broadly to modern, mainstream structuralist and generative concepts of linguistic description that focus on the analysis of two related core layers of language structure, syntax, and semantics (Borsley, 1999; Butler, 2003; Chierchia & McConnell-Ginet, 2000; Chomsky, 2002; Seuren, 1996). *Syntax* is used in the general sense, to refer to the rules of linear combination of meaningful elements into larger elements, resulting in hierarchical structures (constituent structures), usually culminating at sentence boundaries (although sometimes exceeding them, for example, in text linguistics). The term *semantics* is used to encompass the corresponding rules for combining semantic elements into more complex ones. The combination of parallel syntactic and semantic structures, such that every syntactic rule has a semantic counterpart, is often regarded as a language's grammar. This view on the field of linguistics as grammar, that is, as a rule-governed nexus of syntax and semantics, is motivated by the fact that these two areas roughly correspond to the fundamental distinction between form and meaning. As I explain in the following section, it is this distinction that underlies the "representational problem" in information science. A structuralist account of a syntax-semantics-based definition of a grammar is sufficiently general to encompass both linguistic and information scientific uses of the terms *syntax* and *semantics*. This shared terminology constitutes an important unifying feature of the two disciplines.

The question of how linguistic subdisciplines are connected with information science generally is rather different from the more specific question of the relationship between linguistics and IR and raises the issue of the extent to which information science can be considered a unitary science. Modern definitions of information science as a technological, problem-oriented discipline encompass a large number of somewhat incoherent subdisciplines and methodological approaches (Bawden & Robinson, 2012; Pickard, 2013). This makes it difficult to identify meaningful connections between linguistic and information scientific perspectives. Information science has most often drawn on linguistic perspectives to address the question of how best to represent information, that is, how to represent documents to make them findable by users. This "representational problem" has been fundamental to information science throughout its history (see, for instance, Blair & Kimbrough, 2002; Frohmann, 1990; Fugmann, 2002; Hjørland, 1998a; Larson, 2010; Sparck Jones & Kay, 1973) and has also stimulated a considerable amount of research within linguistics.

The specific linguistic challenge for an IR framework is a communication scenario in which it is possible to consider a given item of information in a document from two perspectives. On one hand we have a representation of the item of information ("metadata"), which is considered (part of) a "description" of the original, complete item present in the document. This representational process is described in indexing theory (Broughton, 2006; Chowdhury, 2010; Frohmann, 1990; Fugmann, 2002; Lancaster, 2003; Mai, 1999; Svenonius, 2000; Weinberg, 2009). On the other hand we have the IR process (Baeza-Yates & Ribeiro-Neto, 2011; Blair, 1990; Chowdhury, 2010; Pandey, 1997, 2003; Ruthven & Kelly, 2011; Warner, 2010), which allows a user to gain access to the represented piece of information via a search statement.

The connection between indexing and metadata and searching, considered both as activities and intellectual approaches to information, is at the heart of information science and makes linguistic reasoning central to information science. For the purposes of this article, I restrict myself to the indexing and IR-based model of organizing and accessing information, as this is a core concept in information science. In the remainder of the article, this model is used as a framework for relating IR—both the indexing and retrieval processes—to linguistics.

A word on terminology. I use the term *information science* to refer to the discipline as a whole and the term *information retrieval* or *IR* to refer to the subdiscipline. The compound term *information science/retrieval* (or *information science/IR*) is used to indicate that an argument refers to both the superordinate discipline of information science and the subdiscipline IR.

As a consequence of its disciplinary concern for meaning, linguistics has accompanied information science and retrieval in its attempts to become a "humanistic science" (some might regard this as a contradiction in terms) from the 1950s and 1960s until today. The following sections tell this

DOI: 10.1002/asi

history from the information scientist's perspective. The arguments made here can be viewed as supporting Tredinnick's more general key points by applying them specifically to information science's interdisciplinary relationship with linguistics. This is done by exploring how the science-meaning dilemma has influenced the changing relationship between information science and linguistics and language.

The article is structured as follows. In the next two sections I take the well-known paradigms of information science, the physical and the cognitive, as a point of departure and explore how linguistics came in to assist information science in those two phases. In the subsequent two sections, I ask the question "Where are we now?" and show how former linguistic concepts such as "semantics" are assimilated into a new technological discipline CPIS ("Computational paradigm of information science") by gradually giving up their linguistic connotations and entering into new disciplinary contexts. The subsequent section discusses two examples of linguistic consolidation in information science, where the autonomy of linguistic concepts is preserved and genuinely contributes to novel ideas in information scientific reasoning and interdisciplinary relationships in general. These interdisciplinary aspects of the information science–linguistics relationship are explored in the succeeding section, where results from interdisciplinarity research are used to describe the different phases of information science's development from a "case science" to a fully fledged discipline and how linguistics was expected to support this emancipating process. I wrap up with conclusions.

The Physical Paradigm in Information Science and Its Linguistic Counterparts

The physical paradigm of information science,[1] which dates from the 1950s and 1960s, defines the discipline predominantly as the science of IR (Chowdhury, 2010, p. 1; Larson, 2010, p. 2553). In this paradigm information is conceived as an objective, real-world phenomenon[2] with distinctive material manifestations, and the objective of information science is thus to uncover objective knowledge about the nature of the phenomenon (Tredinnick, 2006, p. 64). Hjørland, referring to Ellis (1996),[3] described this paradigm as follows (Hjørland, 1998b, p. 610f):

> One approach (often called "the physical paradigm" [. . .]) considers information retrieval as an objective, neutral process, where the solution is a "technological fix" that can be measured by "recall" and "precision." Algorithmic approaches in information science are based on such thinking and on the presumption that the subject of a document is a function of the words in the document (sometimes even that the subject can be described by extracting words from the document). In other words, the "subject" is implicitly regarded as a "semantic condensation" of the document. In my analysis, this view is related to the empiricist view.

The objective, physical perspective or model of IR systems and the related basic claim that the subject of a document is determined by the meanings of the words in it—expressed in Hjørland's definition of the empiricist model of information science—respond clearly to the structuralist, positivistic assumptions that were prevalent in contemporary linguistic theorizing. To be useful to "physical" information science, linguistic theories had to be applicable/employable, readily machine-implementable and, ideally, provide coarse-grained analytical techniques that could be adapted to the bigger units with which information science usually concerned itself (Sparck Jones & Kay, 1973, p. 4f). Possible candidates were theories of formal languages from computer linguistics, artificial intelligence research as well as theories derived from text linguistics and theories with a pronounced emphasis on structural and formalizing properties, for example, the early generative theories. Text linguistics in particular appeared to offer information scientists feasible techniques that could be used on texts as well as sentences, as the text/document was regarded by many information scientists as the most basic unit of their discipline. Fine-grained, sentence-based syntactic analyses did not seem to be any more effective than more rudimentary techniques, as Warner (referring to Sparck Jones & Kay, 1973, p. 197) pointed out (Warner, 2007b, p. 282):

> Linguistically very crude procedures seemed to work quite well for retrieval, with retrieval primarily understood as the transformation of a query into a set of records, and it was not clear what contribution could be obtained from more sophisticated procedures.

In the following section I argue that the kinds of linguistic theories information scientists attempted to borrow and the kind of linguistic support they hoped for were determined by their wish for concrete problem solutions and by technological needs and challenges specific to that time.[4] In the 1950s and 1960s linguistics was also attractive to information scientists because, in science theoretic terms, contemporary linguistic theory offered a similarly positivist view of its object of study, the text. According to orthodox contemporary linguistic theories, texts were delimited, physical entities with distinctive structural features and could thus be analyzed scientifically and objectively (Petöfí, 1969, 1971; Sözer, 1985). This common positivist view of information and text made information science and linguistics "natural bedfellows" (Sparck Jones & Kay, 1973, p. 1), both with respect to their science theoretical assumptions and the presumed conceptual and ontological similarity of their objects of study. One might apply the slogan "Like attracts like" to the first phase of the long-term relationship between information science and linguistics.

According to Tredinnick information science faced (and still is facing) the fundamental conundrum of how to apply scientific methods to socio-cultural phenomena without making sure that observed objects in information science actually are independent of observation and experimentation (Tredinnick, 2006, p. 71). This is in particular problematic in connection with texts. Tredinnick states that "The means

DOI: 10.1002/asi

by which we understand, use or store texts can have an impact on the qualities of those texts" (Tredinnick, 2006, p. 71), targeting the early information scientists' positivist concept of a document. This critique from an information science perspective can also be viewed as an accurate description of a similar problem in contemporary text linguistic theory. The early text linguistic structural definitions of text detached the formal features of texts from their semantics and pragmatic uses, what gave information scientists an independent, linguistic justification for disregarding the meaning and cultural context of information and information activity (cf.Tredinnick, 2006, p. 71). Once again we see that like attracts like.

The Cognitive Paradigm and Its Linguistic Counterparts

The cognitive paradigm in information science can be characterized as follows (Hjørland, 1998b, p. 610f):

> Another approach (often called "the cognitive view" [. . .]) relates the subject of a document to a user's knowledge (or rather to his or her anomalous state of knowledge). Information is here seen as an object, which can fill a gap in an individual person's knowledge. By using cognitive psychology's study of human information processing, it is imagined that it is somehow possible to build information systems, which can relate the content of documents to individuals' needs. In this way, there is a connection back to a rationalist influence.

The cognitive shift in information science reflected a re-evaluation of the positivist approach; it acknowledged that information is socially embedded and that individuals play a role in interpreting information in a meaningful way (Tredinnick, 2006, p. 72). This subjectivist interpretation gave rise to a series of influential concepts and approaches, including Belkin's (1980) concept of a person's information needs ("anomalous states of knowledge") and Kuhlthau's constructivist analysis of information seeking (Kuhlthau, 2004; Tredinnick, 2006, p. 73). The focus of IR shifted from establishing exact correspondences between index data and search queries to a more individualistic, fuzzy, and subjective model of the IR process, as Tredinnick described (Tredinnick, 2006, p. 73):

> Information retrieval therefore becomes a process of matching imprecise representation of information with imprecise representation of need, or in other words matching search statements against surrogates.

Referring to Taylor, Tredinnick described information seeking as a negotiation between an information seeker and an information system (Tredinnick, 2006, p. 74).

One of the main points Tredinnick made in connection with the cognitive paradigm is that it preserves the status of information as objective fact. Hjørland had already made a similar point: "Information is here [in the cognitive view – VE] seen as an object, which can fill a gap in an individual person's knowledge" (Hjørland, 1998b, p. 610f). The cognitive paradigm modified the physical paradigm's notion of "objective information" to emphasize the role of the individual in the processing, reception, understanding and formulating of information and information needs (Tredinnick, 2006, p. 75–77); in other words the notion of "objective information" gave way to the notion of "individualized, personal experience of objective information." Tredinnick concluded that although the cognitive shift represented a move away from naïve positivism, information science was still characterized by scientific, objectivist approaches to information (Tredinnick, 2006, p. 79). In other words, because the cognitive shift left the fundamental conflict between informational, meaning-related phenomena and scientific methods for understanding them unresolved, information science's "meaning problem" survived, albeit in cognitive camouflage.

These observations can be substantiated by considering the specific aspirations that cognitive information science entertained with respect to language and linguistics. Two trends in cognitive information science, each attacking IR systems from a different perspective, drew heavily on linguistic theories and ideas. The following schematic of the IR process illustrates this:

Consider first the representational processes in IR (right side of Figure 1). The cognitive rethinking of the IR scenario promoted a pronounced *meaning skepticism* and a radical critique of the naïve view of the semantic relationship between documents and their representations in indexing theory. Meaning skepticism describes a skeptic attitude towards the proposition that meanings exist in the sense that they are directly linked to linguistic forms (words, sentences, texts), which contain and express them. Meaning skeptics have doubts about texts having a distinct meaning without taking their uses, their authors, receivers and possible contexts into account. The meaning skeptic's strong focus on use and the pragmatic conditions of use, replacing the conventional belief in words as containers for fixed meanings, was a serious challenge for document representation and indexing theory.

Turning now to the consideration of information need and user needs (left side of Figure 1), we can describe the cognitive shift as a move from considering a system user to considering a *language user*, a shift that resulted in a radical reinterpretation of the whole indexing system-user complex. The placing of a communicating user at the center of the analysis of information need is clearly rooted in cognitive assumptions; however the meaning skepticism fostered by the cognitive re-interpretation of representational processes seems to reflect a direct attempt to shift information science from being a purely practical science to being a communication-based humanities discipline. The fundamental aim of meaning skepticism was to integrate a new, more dynamic concept of meaning into the analysis of information processes; this is demonstrated below. Meaning skepticism in information science is very directly related to the information-meaning dilemma and the objectivist legacy of

DOI: 10.1002/asi

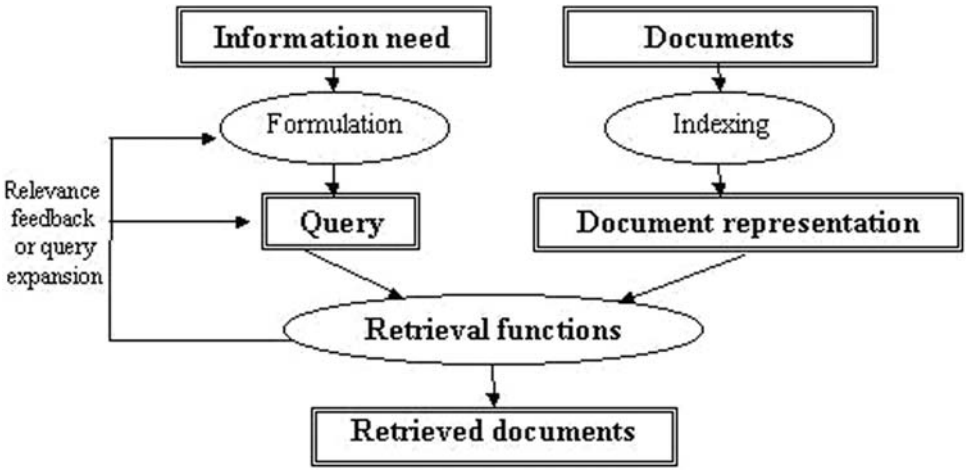

FIG. 1. The information retrieval process (taken from Gharaibeh & Gharaibeh, 2012).

the physical paradigm, as Tredinnick has already pointed out. This is discussed in more detail in the following section.

Meaning Skepticism

The cognitive shift in information science was linked to a strong interest in tracing the scientific assumptions and language theoretic roots of the discipline. These aspirations were due, as mentioned above, to a desire to elevate information science from a purely practical science to a distinctive, theoretically sound discipline. With this aim in mind there was an attempt to tackle the information-meaning dilemma by recognizing the essentially linguistic, semantic nature of information-related activities, whether human or machine.[5]

The skeptical reorientation mainly affected indexation and indexing languages and was focused on the input component of IR systems. Two tendencies can be identified: a conservative, *utilitarian approach* to invoking linguistic theory to rethink the indexing component, and a more *fundamental critique* of it.

The utilitarian approach to linguistic theory argued that the long tradition in information science of conceiving the description of a document as a thematic or topical representation of it could be better understood from a linguistic perspective. The argument was that the indexical relationship is similar to the more general relationship between a text and a summary of that text. The issue of "semantic condensation" was studied extensively by contemporary text linguists. In contrast the more fundamental critique of indexing recognized that linguists and language philosophers, particularly pragmatic theorists of language use and context-oriented theorists such as Searle (1985), Grice (1975, 1989), and the linguist Levinson (2003), had become more critical of the idea that there were simple, structural correspondences between text forms (words and sentences), their "literal" meanings and—this was of particular relevance to information science—a condensate representing the "overall" meaning of a text. Information scientists who endorsed this more fundamental critique were no longer satisfied by mainstream descriptive models that viewed the relationship between document, topical indexation, and semantics as the "meaning" side of linguistic signs (an exception was Beghtol, 1986). This reaction against a naïve semantics was heavily informed by ordinary language philosophy, Wittgenstein's usage-based theory of meaning, and a strong semantic skepticism (Blair, 1990, 2003, 2006; Frohmann, 1990; Hjørland, 1998a).

From System User to Language User

Cognitive information science, which postulates a strong relationship between the meaning of a document—its semantic content—and the information needs of the user, prioritizes language theories that conceive the formerly generic "system user" more specifically as a *language user*. From this communication-oriented perspective a linguistic agent uses an interface (Chowdhury, 2010, p. 265ff; Tedd, 2005, pp. 129–173) and the search algorithm behind it to access document surrogates and, eventually, documents that potentially meet her information needs (Chowdhury, 2010, pp. 5–9). The information science theoretic literature contains several linguistic- and language-inspired philosophical proposals which consider information search from a communications perspective, including Searle's speech act theory (Searle, 1975; cf. Searle, 1985) and Grice's conversational maxims theory (cf. Grice, 1975, 1989). Both speech act theory and, to a greater extent, the theory of conversational maxims tackle a fundamental feature of language use, namely that communicants typically do not say as much as they "should," seen from a naïve linguistic viewpoint; conversational use of language is heavily influenced by the context and the participants' assumptions and knowledge. The communication of an individual user's information needs to a computer interface in the form of a search query, perhaps with the involvement of an information specialist, is an example of the kind of contextualized communication situation that philosophers of ordinary language have in mind

DOI: 10.1002/asi

when analyzing human communication. The information scientist Blair was largely responsible for introducing pragmatic and language philosophical concepts into information science (Blair, 1990, p. 194ff; Blair, 1992, 2003, p. 43f).[6]

The concept of the language-using system user has led to some new perspectives on traditional concepts in information science. From a communications point of view a document description, for example a list of subject descriptors, can be conceived as a component in a communication act between the indexer and the group of potential users of the index (see Blair [1990] and Blair & Kimbrough, [2002] on the concept of "exemplary documents"). Similarly, the cognitive emphasis on language gave rise to powerful arguments against the naïve notion of relevance, according to which documents were considered to be in a "thematic correspondence" with document content, document description, and user queries. New notions of relevance that relate the user's individual knowledge to her cognitive needs have been discussed (Borlund [2003] gives a good survey). Cognitive information science also led to the development of logical approaches, these involved linking descriptions of user knowledge in terms of propositional sets to the propositions implied by document descriptions in a collection (for an early proposal see Cooper, 1971; Van Rijsbergen, 1986). Other examples of linguistic contributions to cognitive information science are information scientific explanations of relevance based directly on Sperber and Wilson's concept of relevance (Sperber & Wilson, 1995) as in Harter's notion of "psychological relevance" (Harter, 1992).

Where Are We Now?

In looking at more recent developments in information science I again draw on Tredinnick's analysis to provide a tentative explanatory frame for the more specifically linguistic aspects of these developments.

Tredinnick argued that information science separated from library science mainly because of librarians' atheoretical, humanities-oriented views, which were grounded in an essentially unscientific literary culture. Tredinnick claimed that in doing so information science missed out on the critical interdisciplinary academic discourse in neighboring disciplines that took cultural phenomena seriously such as psychology, philosophy and, last but not least, linguistics (Tredinnick, 2006, p. 79f).

In the 1960s and 1970s access to computational resources was limited to large organizations such as universities, and the implementation of large IR systems resulted in novel search problems and tasks for information scientists working closely with computer scientists (Tredinnick, 2006, p. 81). These conditions helped ensure that information science remained distinct from computer science, mainly because of the particular nature of the central problem in information science, the search context. All in all, the technological framework of the 1960s and the 1970s was favorable to information science's project of constituting itself as a scientific discipline.

The picture changed radically in the two subsequent decades as the personal computer (PC), and in particular the networked PC, became widespread, and computer resources diversified. Perhaps the most important development was the blurring of the dividing line between computer applications and IR systems as retrieval modules became ubiquitous in software, databases and groupware. This placed information science in competition with computer science (Tredinnick, 2006, p. 80) with the result that "[t]he influence of information science on this development declined" and "[. . .] the center of gravity of information retrieval [. . .] shifted to computer science" (Tredinnick, 2006, p. 81). Tredinnick went on to state that "[t]he computing industry and computer science [. . .] severed their connection with information science and were charting territory on their own" (Tredinnick, 2006, p. 81).

Where does this separation leave linguistics and the rich interdisciplinary relationship with information science that was identified in the discussion of the physical and cognitive paradigms above? And what kind of relationship did the "atheoretical," humanities-oriented discipline of librarianship have with linguistics? Although it is easy to ask these questions it is more difficult to answer them. As an active information science researcher with a strong background in linguistic research and 10 years of professional experience as a subject specialist at a university library I feel more participant than observer in these debates.

These and other limitations notwithstanding, I argue that during recent years there have been two separate developments in the relationship between information science and linguistics (consideration of the relationship between librarianship and linguistics is omitted for reasons of space). First, I consider that there has been a *consolidation* of the relationship based on the cognitive paradigm and that this is reflected in the continuing influence of linguistics on information scientific theory and practice. This consolidation is sustained by continuing efforts to integrate authentic linguistic concepts and theory (such as the paradigmatic/syntagmatic distinction discussed below) into the information scientific frame and thus undermine the antimeaning preoccupations and biases of a still predominantly positivist science. The borrowed linguistic terms and concepts largely retain their linguistic integrity when applied in information science, that is, they retain their discipline-specific meaning and thus contribute to a genuinely interdisciplinary rethinking of information scientific frameworks. Second, there has been an *assimilation* of linguistic concepts into IR that has resulted in the development of a special branch of information science. This latter development, which is discussed in the following section, is completely consistent with Tredinnick's analysis of computer science's monopoly over IR.

The Assimilation of Linguistic Concepts Into Computer Science

The ubiquity of IR technology in all sorts of Internet and software applications brings together practical information

DOI: 10.1002/asi

scientists and computer scientists in the diversified, highly technical and technological field of online IR. Interest is focused on the Internet as the major information medium of our time (Antoniou, Groth, van Harmelen, & Hoekstra, 2012; Chaka, 2010) and research draws on work in fields such as research analysis, research communication, information literacy, information management, interactional design and human-computer communication, culture mediation, knowledge structures, social media, computer-mediated communication, use of natural language in Internet queries, and so on.[7]

In contrast to the above-mentioned consolidation of modern information science, information science researchers working within the computational paradigm have developed a corpus of linguistic terminology that is superficially similar to traditional linguistic terminology (e.g., "semantics," "morphological analysis" etc.) but uses familiar terms to denote new concepts, having moved away from the original disciplinary contexts of use. Most of these new concepts are embedded in the interface between IR and Internet technology, as we see in the investigation reported below. In the following analysis I have distinguished between assimilated linguistic terminology, "linguistic2," and consolidated linguistic terminology that preserves the original meanings of its terms, "linguistic1."

Citation Analysis and Keyword Collocation

To explore how linguistic2 terms are introduced into this emergent field of research, I discuss a small sample of articles that appear to deal with linguistic topics. This exercise examines recent research in information science that has made use of linguistic concepts. Two intertwined methods are used to address the specific research question of whether the linguistic approaches employed in these articles represent the use of linguistic1 concepts (authentic linguistic concepts) or whether linguistic terms are used to denote adaptations of linguistic concepts that are being assimilated into this emerging technological research area.

The first method is based on citation analysis and tentative evaluations of whether the articles in question cite linguistics1 literature and if so, which authors and works. Low proportions of linguistics1 references on reference lists are taken as an indication of the terminological and hence disciplinary self-sufficiency of linguistic terms within a "closed" linguistics2 paradigm.

The second method is a novel technique that might be described as "keyword collocation analysis." The term *keyword collocation* refers to the subject metadata structures of an article, that is, the set of keywords attached to the articles in question (Borgman, 2007; Broughton, 2006; Lancaster, 2003). My argument is that the structure of subject terms attached to a document can tell us something about how linguistic concepts are integrated into the general thematic structure of a article. How so? In the context of databases and information searches an established terminological vocabulary such as linguistics1 typically corresponds to a semantically complex thesaural structure in which concepts are related to one another according to agreed, domain-specific knowledge structures in a way that is consistent with the conventions, norms and practices of the relevant discipline at a given point of time. The Linguistics and Language Behavior Abstracts (LLBA) Thesaurus is used as an example of a disciplinary thesaurus. The complex semantic structure mirroring the terminology of a discipline is partly realized in hierarchical relationships in thesauri; for instance, a descriptor such as SEMANTICS can be linked to a whole subset of semantic schools, theories and approaches (FORMAL SEMANTICS, GENERATIVE SEMANTICS, PROTOTYPE SEMANTICS, . . .). Hierarchical, knowledge-based relationships are directly related to a discipline's terminology; for instance, NOUN PHRASES, PREPOSITIONAL PHRASES and so-called WH PHRASES are all PHRASES, and PHRASE is a member of the category LINGUISTIC UNITS. This nexus of relationships is a product of a discipline's research history; it represents the discipline's accumulated knowledge and is thus subject to ongoing discussion and revision. The noncontingent relationships between knowledge units are coded by the two central types of thesaural relations, hierarchical-generic and partitive relations (Green, 1995b, 2002).

When we turn to the third type of relationship between knowledge units, which is often subsumed into the category of "associative" relationships (Lancaster, 2003, p. 18; Pandey, 2003, p. 31ff), a somewhat different picture of the semantics of thesaural relationships and their arguments emerges. First, the discipline's theoretical-terminological knowledge is structured in terms of generic and partitive relationships rather than associative relationships. Associative relationships primarily code research conventions and good practices in a discipline. In thesauri they are usually expressed by the "Related Terms" category (Bawden & Robinson, 2012, p. 121; Lancaster, 2003, p. 23; Svenonius, 2000, p. 160f; Weinberg, 2009, p. 2285), a term that encompasses a wide array of dimensions, as we see later. Consider the descriptor SEMANTICS as an example again; the following practices and conventions are coded in the list of its related terms (based on LLBA):

1. Grammatical phenomena. Language phenomena that typically (conventionally) are studied in a semantic framework: BINDING, COMPARISON, NEGATION, ANAPHORA, TIME, . . .
2. Concepts. Logical tools and linguistic concepts that typically (conventionally) are used in semantic arguments: ENTAILMENT, IMPLICATURE, TRUTH, AMBIGUITY, POLYSEMY, PROPOSITION, LOGICAL FORM, . . .
3. Levels of analysis. Layers of language description (and their interaction) that can be the target of semantic analysis: WORD MEANING, WORDS, SYNTAX-SEMANTICS RELATIONSHIP, SYNTAX, DEEP STRUCTURE, SYNTACTIC ANALYSIS, LEXICON, . . .

DOI: 10.1002/asi

4. Linguistic fields. Linguistic sub-disciplines and theoretical frameworks that typically (conventionally) study semantic phenomena: COMPARATIVE LINGUISTICS, COMPUTATIONAL LINGUISTICS, HEAD DRIVEN PHRASE STRUCTURE GRAMMAR, COGNITIVE GRAMMAR, . . .
5. Perspective. Study of semantic phenomena is typically (conventionally) carried out either from a developmental, diachronic perspective or from a synchronic one, analyzing at a given point in time: SEMANTIC CHANGE, SEMANTIC FIELDS, ETYMOLOGY, . . .

Let us assume that the structure of subject-indexing terms in documents that take a linguistic1 approach to research is based on linguistics1 conventions and scientific practices (as well as the generic and partitive relationships) and that this is reflected in the document description via the use of terms connected by associative relationships in the set of descriptors. In our example the broad term SEMANTICS could meaningfully be used to refer to aspects of the grammatical phenomenon studied (see Point 1 above), the semantic concepts used (see Point 2 above), layers of language targeted by the analysis (see Point 3 above), the theoretical framework used (see Point 4 above) or to indicate the chronology of the semantic argument (see Point 5 above). Conventional links between thesaurus items, in this example between SEMANTICS and five aspects of its use in research, reflect disciplinary conventions and ideas about what constitutes good practice. The keyword collocation hypothesis posits that these links are reflected in patterns of keyword combinations, that is, collocations, such that combinations of concepts refer to and are referred to by other concepts to which they are associatively linked (some types of connections are specified by pts. 1–5 above). The set of linguistic keywords present in an article's metadata can thus be interpreted as an indexer's coding of both terminological (generic/partitive) coherence and disciplinary practices (associative) through her selection of thematic descriptors. The syntagmatic relatedness of keyword arrays is based on an intact, terminologically consistent, practice-based network of paradigmatic semantic relationships between linguistic descriptors defined in a domain-specific thesaurus. The combinatorial syntagmatic complexity of linguistic index words (Green, 1995a) in a document description can thus be taken as an indication of a document's linguistic theoretical orientation.

Assuming that complex linguistic indexing (use of more than one linguistic keyword) serves as an indication of the extent to which a article's topic is linked to the linguistics1 framework, how should we interpret isolated linguistic descriptors in sets of keywords? Here, obviously, the array of keywords also mirrors scientific practice in terms of associations, just as in the case described above. What does linking a single linguistic concept to several nonlinguistic concepts within a set of subject terms signal about disciplinary practice? The existence and pattern of such hybrid arrangements of keywords appears to reflect the use and nature of hybrid practices and conventions characteristic of a new discipline at a certain stage of interdisciplinarity (interdisciplinarity is discussed further below). A detailed exploration of these emerging, interdisciplinary collocations of linguistic and nonlinguistic subject descriptors is outside the scope of this article; however, I consider some examples that illustrate how a linguistic term becomes connected with other information scientific terms in subject descriptions. This also sheds some light on the interdisciplinary practices of the emerging discipline, which I refer to as the "computational paradigm of information science" (CPIS). The presence of isolated linguistic terms in an article's array of subject terms are taken as an indication of a linguistic2 use in a CPIS environment.

This method is obviously vague with respect to both the argument about the significance of keyword collocation and the assessment of cited literature. A more explicit theoretical account—backed by empirical evidence—of how the syntagmatic thematic structure of a set of attached document descriptors is linked to the paradigmatic semantic structure of the vocabulary is needed. This should include an account of the hierarchical relationships mapping the search vocabulary to a discipline's terminology and an account of how the associative relationships embodied in the search vocabulary reflect disciplinary practices. Arguments based on keyword collocation will remain open to challenge until the theoretical foundations are stronger and clearer. The second technique, evaluation of cited references to determine whether they represent a linguistic1 orientation, is clearly subject to personal bias. However, the assessments made in this article are based on the author's expertise in linguistics1, and the titles in question are named and can be scrutinized by interested or skeptical readers. The primary bases for determining whether a cited publication should be classified as linguistic1 were the title and author(s) and a further criterion was that the publication channel was within the linguistic1 arena. For example, articles presented at information and knowledge management conferences were classified as linguistic2.

This discussion is intended solely to illustrate the emergence of a new discipline through a process of assimilation of concepts from another, in this case linguistics. The sampling method is described in detail below, but there was no attempt to select a representative sample. I describe the procedure by which the 12 sample records were obtained and the criteria used to ensure that selected records were relevant to the research question specified below. The analysis illustrates theoretical points, but does not permit general conclusions about populations of document collections.

Operationalizing the Research Question

To collect an appropriate sample of records it was necessary to operationalize the following phrase: *Recent (1) scientific articles (2) from the area of information science (3) that make use of linguistic concepts (4)*. I assumed that the phrase as a whole would be sufficiently specified if all its

DOI: 10.1002/asi

FIG. 2. Configurations of the descriptor LINGUISTICS in the LISTA thesaurus.

components were defined and operationalized. Component 3 defines the relevant population. I selected the Library, Information Science & Technology Abstract (LISTA) database as a source of records. LISTA indexes information scientific literature, including literature written from both linguistic1 and linguistic2 perspectives. (Note that the Library and Information Science Abstracts, LISA, could also have been used.) To determine whether articles met the Component 4 criterion—*using linguistic concepts*—I used the database's thesaurus to make sure that all articles in the sample dealt with linguistic (linguistic1/2) topics. As a first step I attempted to operationalize linguistic affiliation by the descriptor LINGUISTICS, which, surprisingly, did not exist in the LISA thesaurus (Figure 2).

In LISTA the term LINGUISTICS is introduced with a single subheading, DATA PROCESSING, which is a nondescriptor for the authorized term COMPUTATIONAL LINGUISTICS. LINGUISTICS as thesaurus term also appears as first component in the unauthorized, multiword subject heading LINGUISTIC LIBRARIES. This could be interpreted as a first indication of a weak relationship between LINGUISTICS and *linguistics as a discipline* that is limited to the COMPUTATIONAL branch of linguistics and to librarianship (LINGUISTIC LIBRARIES). A similar pattern of relationships is revealed when we move from the discipline to its object of study, language (Figure 3).

Like LINGUISTICS, LANGUAGE is not a single descriptor in the LISTA thesaurus, the term appears four times as the first component in complex subject headings and three times with additional subheadings (DATA PROCESSING, TRANSLATING, WRITING). All entries are unpreferred strings and refer the user to corresponding entries on COMPUTATIONAL LINGUISTICS, TRANSLATING & INTERPRETING and WRITING. I therefore concluded that the most general object of linguistic study, language, could not be appropriately captured in the LISTA thesaurus, let alone in a free text search.

The next attempt to operationalize the criterion *using linguistic concepts* was based on choosing a descriptor that denoted a concept that was central to both linguistics1 and linguistics2 research contexts. The term SEMANTICS

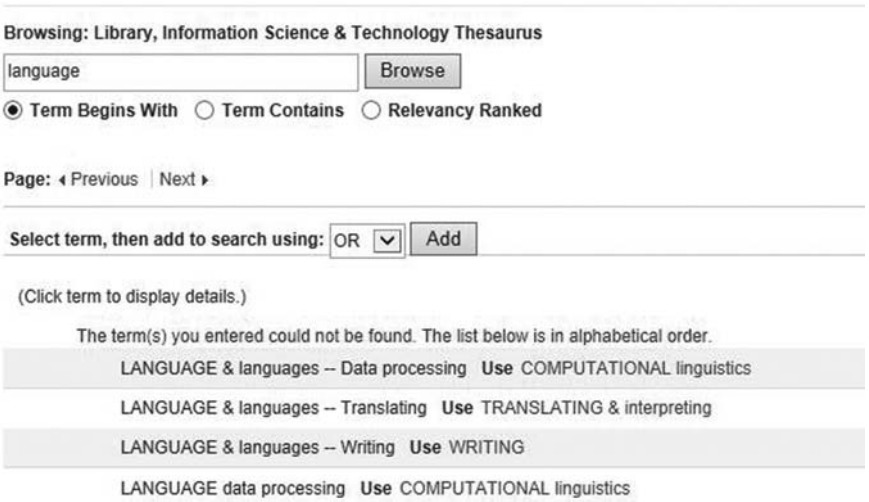

FIG. 3. Configurations of the descriptor LANGUAGE in the LISTA thesaurus.

DOI: 10.1002/asi

FIG. 4. Configurations of the descriptor SEMANTICS in the LISTA thesaurus.

proved to be an appropriate common denominator in the LISTA thesaurus (Figure 4).

The scope note indicates rough coverage of a mainstream, linguistic1 concept of semantics, although an important aspect, sentence meaning, is obviously not part of the definition, and it is also not clear how substituting "meaning" for "semantic" in the definitions would contribute to our understanding. In summary, the term SEMANTICS appeared to be a common, authorized descriptor with an appropriate definition in the scope note and it is a concept that is relevant to both information science and linguistics; I therefore decided that it was a suitable operationalization of an article's linguistic topicality.

The remaining components of the sample description, *scientific articles* (2) and *recent* (1), were easy to define and operationalize. Scientific articles (Component 2) in the LISTA database were defined as "published in scholarly (peer-reviewed) journals" and, according to Component 3, the sample was limited to articles published in the past 10 years as this is a period for which one can be confident that information technological developments have had a measurable impact on the assimilation of linguistic2 terminology into the information science literature. This criterion was operationalized by limiting the search to articles published between January 1, 2005, and December 31, 2014, inclusive using the delimiter "publication date." The search was also limited to publications with accessible links to full text and the option of accessing references in electronic form as well as in a pdf. These two search criteria were implemented to improve the sample and save time; they are unlikely to have influenced the relevance of the records sampled.

The Sample

A search run on July 2, 2015, with the parameters described above produced 53 hits listed chronologically in declining order of publication date. I selected the first and last records (nos. 1 and 53) to make sure that the sample covered the entire time span under investigation. I also selected every fifth record, beginning with no. 5. This procedure resulted in a 12-record, chronologically ordered sample, as presented in the appendix.

The chronological distribution of the sample was rather uneven. There were no records from 2007–2009 and 2013 was clearly overrepresented, with 4 records. The most frequent publication channel was the *Journal of the American Society for Information Science & Technology* that published 5 of the 12 records, almost half the sample.

Results and Discussion

It was clear from a visual inspection of the sample that the linguistics1 paradigm was weakly represented. Only 3 records cited any linguistic1 reference (r4, r6, and r11 with 6%, 15%, and 5% linguistic1 references, respectively). This demonstrates the limited extent to which the linguistic term "semantics" has been assimilated into the CPIS. No chronological trend in citation of references using linguistic1 terminology was detected in the sample.

A qualitative analysis of keyword collocations revealed some weak associations between the presence of the linguistic2 keyword SEMANTICS (an inclusion criterion) and the other, nonlinguistic keywords in the subject descriptions. Five nonlinguistic keywords seem to occur relatively frequently in the 12 records: METHODOLOGY, METADATA, CLASSIFICATION, INTERNET/WWW and INFORMATION RETRIEVAL. The following table charts

TABLE 1. Presence of five keywords in in the exclusively linguistic2 records.

	METH	META	CLASS	WWW	IR	Total
r1 (2014)	X	X				2
r2 (2014)	X	X		X		3
r3 (2013)	X		X			2
r5 (2013)				X	X	2
r7 (2012)					X	1
r8 (2012)		X		X	X	3
r9 (2011)						0
r10 (2010)				X		1
r12 (2005)			X	X		2
Total	3	3	2	5	3	

DOI: 10.1002/asi

TABLE 2. Three-phase integration of the term *semantics* into the CPIS.

Phase 1: 2005–2011	Integration of "semantics" into WWW-settings and the traditional, library-based use of it in classification theory
Phase 2: 2012/2013	"Semantics" is now used in information retrieval contexts, indicating an extension of "semantics" to encompass user-related issues
Phase 3: 2014	The relevance of "semantics" to methodology in information science is recognized. The descriptor METADATA indicates an acknowledgement of the document side of the information retrieval system

the presence of these five keywords in the subject descriptions of the nine exclusively linguistics2 records (Table 1).

In the 2014 publications (r1 and r2) the pair METHODOLOGY/METADATA is a good predictor of SEMANTICS. This suggests that the term "semantics" has been adopted in the interdisciplinary context of methodological discussion about metadata in an online setting (cf. the keyword INTERNET/WWW; the most frequent in the sample, with five mentions). In the 2012–2013 records SEMANTICS is more strongly associated with the INFORMATION RETRIEVAL keyword. In all older records (i.e., records from 2005, 2010, and 2011) no significant association between SEMANTICS and other keywords is evident, although the co-occurrence of CLASSIFICATION and the WWW could indicate an anchoring of semantics in library science and humanities semantic notions in describing classification systems. In summary, there is some evidence for a staged integration of semantics into a new computational IR paradigm CPIS; this comes from indications that there have been three phases in the pattern of associations between the term "semantics" and nonlinguistic descriptors (see Table 2).

I turn now to analysis of the records citing linguistic1 references (r4, r6, and r11). The keyword collocation hypothesis posits that the pattern of co-occurrences of a given keyword and other related keywords in a subject description indicate the terminological system the document relates to. In the analysis of the exclusively linguistics2 records I observed that "semantics" was integrated into other, nonlinguistic domains; the temporal changes in keyword collocations are cited as formal evidence of this process. A similar process can be observed in the records that cited linguistic1 references: Linguistic keywords in one and the same topic description interact with each other and signal by this the document's thematic connection to the linguistics1 domain.

In r4, the SEMANTICS descriptor is linked to a specific broader linguistic1 term, LINGUISTICS-Methodology, indicating the superordinate disciplinary domain from which the term SEMANTICS was derived. In this instance there is clearly a hierarchical relationship between a discipline and sub-discipline. The TIME descriptor belongs to the group of

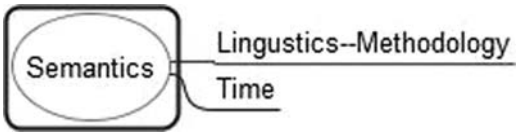

FIG. 5. The SEMANTICS Network in the keyword set of record 4.

linguistic issues typically associated with semantic frameworks (see previous discussion of terminological networks, particularly Point 1, Grammatical Phenomena). The domain-specific embedding of SEMANTICS in r4, involving a hierarchical and an associative relationship in a linguistic1 thesaurus, is visualized below (Figure 5).

A similar pattern can be observed in article r6. In this record the use of SEMANTICS as a linguistic1 term is indicated by a hierarchical connection between SEMANTICS and the term LANGUAGE and languages, whereas the association between SEMANTICS and the keyword COMPARATIVE grammar indicates a specific sense in which the term is used. The list of parameters associated with the term SEMANTICS (see Point 4, Linguistic Fields) describes COMPARATIVE grammar as "linguistic subdisciplines and theoretical frameworks that are typically used to study semantic phenomena." The network of linguistic1 terms for r6, which once again indicates a hierarchical relationship and an associative specification, could be depicted schematically as follows (Figure 6).

The pattern of relationships in r11 is more complex. Here the descriptor SEMANTICS is classified according to both linguistic1 and nonlinguistic terminological systems. Its co-occurrence alongside the SIGNS and symbols keyword indicates an association with the semiotic domain, which traditionally has entertained intensive contacts with the structuralist schools in linguistics1 (Eco, 1991, 1995) and thus shares a significant part of terminology with structuralist linguistics. The descriptor NONVERBAL communication indicates a conceptual extension of the linguistic domain, which is typically restricted to language and verbal communication. The third descriptor, COMMUNICATION of technical information, belongs in a linguistic1 framework but is only loosely associated with semantic phenomena. The descriptor AMBIGUITY, however, is clearly associated with semantics denoting 'semantic vagueness' as one of the most frequent linguistic concepts typically used in semantic arguments (cf. Point 2 in the list of parameters associated with SEMANTICS). The r11 SEMANTICS network, comprising associations involving two linguistic1 terms (one central and one peripheral, left side) and two terms representing an extension

FIG. 6. The SEMANTICS Network for the keyword set of record 6.

DOI: 10.1002/asi

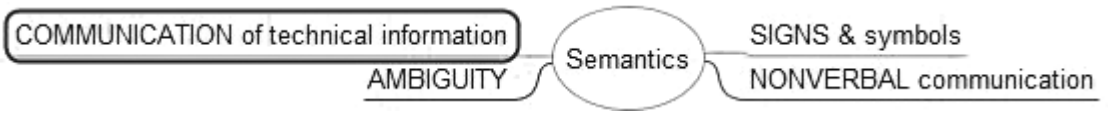

FIG. 7. The SEMANTICS Network in the keyword set of record 11.

of the linguistic1 sphere on the right side, can be depicted schematically as follows (Figure 7).

At this point a brief recap of the analysis and conclusions may be helpful. As a starting point a rather formal citational argument was applied to a sample of information scientific articles dealing with semantic phenomena. It was observed that this sample of research made only very restricted reference to linguistic1 literature. This was interpreted as evidence of a tendency to assimilate linguistic concepts into a new, information scientific context CPIS. I went on to propose a tentative qualitative analysis of the chronology of the assimilation based on tracing the changing patterns of associations involving the term SEMANTICS. In Phase 1 SEMANTICS was linked with a library and classification context; in Phase 2 it was linked to IR contexts and in Phase 3 it was linked with metadata. The keyword arrays for the small number of records in the sample that cited linguistics1 publications (3) reflected this theoretical allegiance, embedding the SEMANTICS descriptor in various aspects of the linguistic1 domain and signaling the article's provenience in a traditional linguistic context.

We can tentatively conclude that the linguistic objects of study in linguistics1 and linguistics2 are not identical. CPIS, a new, linguistically informed but terminologically and conceptually independent branch of information science seems to represent a continuation of the physical paradigm, which now owes its existence largely to and is mainly practiced by computer scientists. From the linguistic perspective this development represents a pragmatic collaboration between applied linguistics (including computational linguistics) and a "technical" branch of information science that has developed from the physical paradigm of information science. The new paradigm appears to be based on the integration of linguistic concepts into a technical IR context and reflects the emergence of a new discipline CPIS. The changes in the relationship between linguistics and information science have gone beyond bridge building; there has been a restructuring of the former physical approach to information and IR to a new discipline CPIS. Below, under the heading of "Perspectives From Interdisciplinarity Research," I discuss these developments from an interdisciplinarity perspective.

The Consolidation of Information Science: Integration of Preserved Linguistic Concepts—Pandey and Warner

The consolidation of information science via the integration of linguistic[8] knowledge is characterized by application of preserved linguistic concepts to IR problems. The discussion of the keyword collocations of articles citing linguistic1 references provided a partial illustration of this process. The consolidatory process essentially reflects the growing acceptance of the cognitive paradigm in information science and represents a continuation of previous attempts to overcome the constraints scientific methodology imposed on meaning-related approaches to information (discussed in an earlier section under the heading of "Meaning Skepticism"). Information scientists working in the humanities tradition do justice to linguistic concepts by appealing to long-established language theoretical concepts such as the semantics-syntax distinction and the well-known Saussurian dichotomy between syntagmatic and paradigmatic relationships (cf. Joseph, 2012; Saussure, 1967/2011). To illustrate this I consider two attempts to apply linguistic theory to information science under the disciplinary banner of linguistics.

Although it is not very well known, Pandey's (2003) investigation *Information Retrieval System: A Linguistic Study* is an elaborate and persuasive argument for the transfer of the idea of a parallel, dichotomous structuring of semantics and syntax to information science. This dichotomy, which represents a central principle of language description in linguistics, can be used in a similar way in information science that is, as the underlying principle in the design of IR systems. Pandey used the semantics-syntax distinction exclusively with reference to the indexing component of IR systems (cf. Frohmann, 1990) and did not, unlike certain other authors, use it as the basis for a linguistic explanation of user-related aspects of IR. From this perspective, which is clearly aligned with traditional library science (Lancaster, 2003; Li, 2009; Tedd, 2005), semantics and syntax can be mapped to two fundamental processes in document content analysis: (a) identification of concepts contained in a document and (b) determination of the relationships between such concepts in text (Pandey, 2003, p. 23). This "dual-process" in linguistic analysis implies the existence of a paradigmatic-syntagmatic axis, a concept that was developed further by Warner (see below).

What makes the indexing process distinctive from the broad, linguistic abilities relevant to comprehension of text in general is, of course, that indexing targets the searchability of index terms, whereas language understanding is not restricted to such a goal. Another condition that distinguishes the indexing process from language understanding as a natural linguistic competency is its requirement for a semantically adequate reduction of text content to its linguistic "essence" that is, a concise summary of document content in terms (words) that function as representatives of concepts (natural language comprehension is not concerned with summaries expressed in language). In the second step of the indexing process these terms—either extracted from

DOI: 10.1002/asi

the document or drawn from an external vocabulary—are connected via syntactic relationships. The resulting chain of syntactically linked terms is applied to the document it represents; it is a "thematic statement" (Pandey, 2003, p. 30) and is treated by traditional indexing theory as a representation of the original document. Thematic statements are—and this is how they are related to linguistic theory and analysis—comparable to natural language sentences denoting certain circumstances in the world (as documents do). Both thematic statements in document description and natural language sentences are interpretable as symbol chains where lexical units (words; terms) are ordered in a syntactic sequence (Engerer, 2014a).

Implementing the syntactic-semantic distinction in the structure of the indexing process makes it possible to draw parallels between the underlying structural principles of indexing and natural languages while still respecting their basic differences (artificial vs. natural; mental location) and their functional divergence (Engerer, 2014b). Pandey, who attributed this theory of indexing language to Ranganathan, summarized the theory's range and limitations as follows:

> Thus the theory of indexing language formulated by Ranganathan is valid even in computerised information retrieval systems at semantic level [sic]. However, syntactic part [sic] of his theory is not applicable to post-coordinate indexing languages. These languages believe in combination of concepts of the search stage. (Pandey, 2003, p. 131)

This raises the question of whether postcoordinate indexing systems—which currently dominate indexing—and the victory of full-text techniques over systematic methods of representing documents through surrogates will diminish the influence of powerful, traditional linguistic concepts on information science. If all syntactic work is outsourced to the user and full-text searching renders the requirements for a shortened representation and indexing obsolete, what role will there be for structural language approaches to IR? Warner has attempted to rehabilitate the use of authentic linguistic concepts in up-to-date IR systems employing postcoordinative indexing on a full-text basis, most recently in his book *Human Information Retrieval* (Warner, 2010). In two preliminary studies (Warner, 2007a, 2007b) Warner linked the syntactic-semantic structural dichotomy, which in Pandey's research was applied only to formal index languages, to the concept of a syntagmatic–paradigmatic axis (Lyons, 1977, p. 270; Saussure, 1967/2011) in IR systems. Warner's model explicitly included the user, or more specifically the user's cognitive system, as a "linguistic agent."

To illustrate his argument Warner used the example of automatic indexing (Mai, 1999, p. 276, 287; Sparck Jones & Kay, 1973, p. 10, 29, 63; Tedd, 2005, p. 170; Weinberg, 2009, p. 2286), where an algorithm extracts searchable description terms from a full-text document. Warner applied Saussure's syntagmatic/paradigmatic dichotomy to this IR setting. In essence syntagmatics describes the linguistic context (the "syntagm") in which a linguistic unit (typically a word) is embedded. There is a paradigmatic relationship between interchangeable linguistic elements for a given syntactic position. The extraction of index terms can thus be understood as the process of detaching an element (typically a word) from its syntagmatic context in a specific document (Warner, 2007b, p. 275). Such words, once transformed into index terms, are semantically "released" into an index and "set free" to assume the maximum possible number of meanings: in their maximal paradigm they become maximally ambiguous terms.

As far downstream as the user query these terms are assembled into novel combinations, a process that Warner interpreted in a Saussurian spirit as the "re-insertion" of isolated terms with multiple paradigmatic meanings into a new syntagm. In the user's syntax, this new syntagm is "re-translated" into multiple syntagmatic instantiations in retrieved documents and texts. In Warner's terminology the user's testing of several syntagmatic environments for a maximally ambiguous search term (or, in plain words, the user's experiments with repeatedly combining a general search term with other terms, specifying and limiting the general term's meaning) instantiates a conversed transformation of a paradigm (Warner, 2007b, p. 275). It is possible that in the reversion the meanings of search terms, which do not become evident until the complete syntagm of the relevant document is checked, will not correspond to the meanings intended by the system user (Warner, 2007b, p. 275f). Only through the retrieval of original text documents can the syntagmatic contexts of query terms be re-constituted, and the coverage of paradigmatic meanings of one single query term can be defined by varying syntagmatic contexts of the word or search term (Warner, 2007b, p. 276):

> Linguistics can, then, contribute a sophisticated understanding of the interaction between signifier and signified enforced by the movement in description from syntagm to paradigm, and from paradigm to syntagm in searching and retrieval, for computational and direct human operations on written language in full-text representation and retrieval.

Without doubt, Warner's language-informed account of "human information retrieval" is an innovative approach. As well as being a novel application of a structural linguistic concept to current information scientific problems it enables one to question the traditional assumptions of the IR community, for example the dogma of the "query transformation" (Warner, 2010, p. 3). More conservative attempts to make use of linguistic ideas and concepts in information science reasoning, such as Pandey's, have rediscovered linguistically motivated structures by working out the language-related principles underlying traditional IR theories; however, Warner has discovered new ways of applying linguistic concepts and language-related analogies and thus opened up new approaches to IR research.

Once again—the two facets of IR, the indexing component on the right side and the linguistic user on the left (see Figure 1), can help to understand the interaction between

DOI: 10.1002/asi

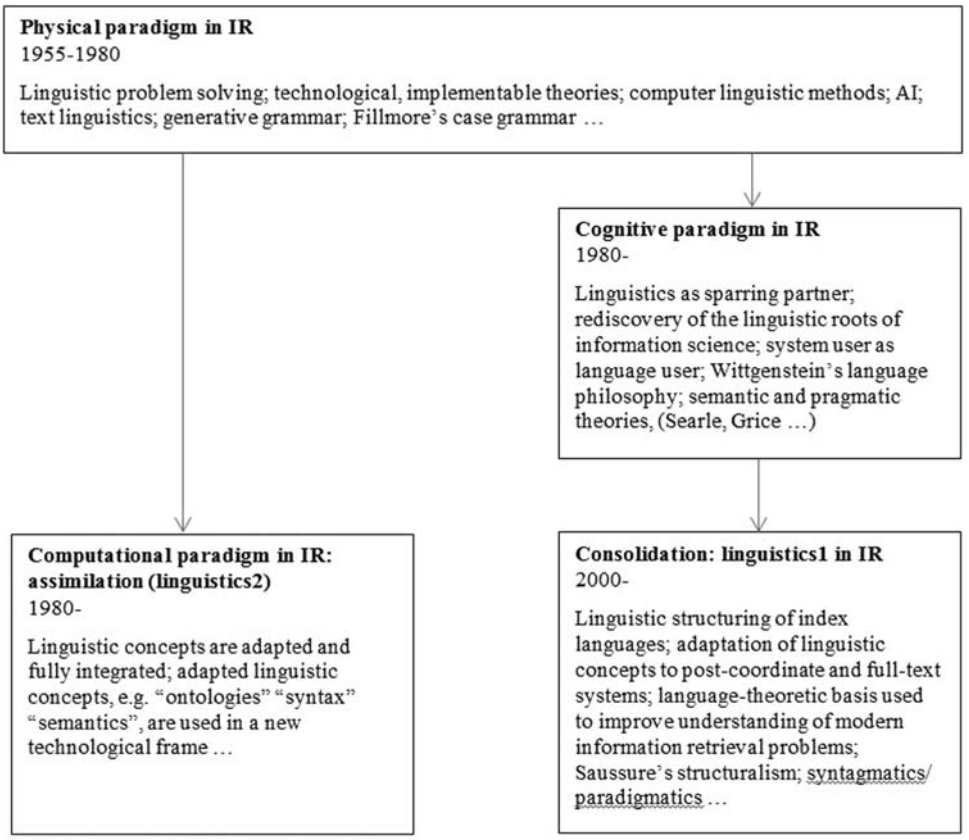

FIG. 8. Relationship between linguistic theory and the paradigms and phases of development in information retrieval.

information scientific problem settings and linguistic thinking better. The metaphoric left-right topology can not only be used to describe the two points of departure for linguistic assistance as demonstrated for the physical and cognitive model of IS (meaning skepticism and "system user as language user") but the same distinction can be used here; Pandey's dichotomy is relevant to the indexing component (right side), whereas Warner's distinction takes the language user's perspective and is relevant to the output component (left side) of the IR system.

Perspectives From Interdisciplinarity Research

The correspondences between linguistics and information science and retrieval rest on the three established paradigms in IR: physical, cognitive, and computational. These differences between paradigms, which are primarily related to developments within the information science and retrieval discipline, are linked to two approaches to the incorporation of elements from linguistics (theories, terminology, etc.) into information science and retrieval that I have called assimilation and consolidation. I have also developed a tentative chronology of the developments in information science and retrieval based on this coupling of paradigms with linguistic assimilation or consolidation. The unidirectional arrows in Figure 8, below, roughly represent the time line, but they are also intended to imply causal processes of the kind "B has developed from A." Each box in Figure 8 contains brief descriptions of a selection of the linguistic elements that can be associated with the relevant IR paradigm.

I now turn to interdisciplinarity theory (cf. Frodeman, Klein, & Mitcham, 2010) to provide a more differentiated, explanatory account of the relationship between IR and linguistics. Interdisciplinarity theory is a very exciting approach that sheds light on the intricate processes that occur when two or more disciplines collaborate or influence each other, or when one discipline is transformed by absorbing ideas from another. The representation of the interdisciplinary relationship between linguistics and information science and retrieval in Figure 8 can be interpreted as an approximation of a transition (not perfectly linear) through three interdisciplinary stages: multidisciplinarity, interdisciplinarity, and transdisciplinarity (this terminology was proposed by Klein, 2010). In the following sections general descriptions of the stages of interdisciplinarity (Klein, 2010, p. 21, 24) are used to develop a more detailed account of the relationship between information science and linguistics.

Multidisciplinarity is a disciplinary relationship of the juxtaposing, coordinating type where theories from different disciplines stand alongside each other but there is no clear integration. From this perspective linguistics has performed a helping, auxiliary function for information science; within information science linguistic theories and concepts have purely instrumental functions. The relationship between the two disciplines can be characterized in terms of the bridge

DOI: 10.1002/asi

building metaphor; scientists (perhaps better: their ideas) can travel between disciplines. The relationship between the physical paradigm in information science and linguistics and linguistic theory can clearly be considered as an example of multidisciplinarity.

Interdisciplinarity is a disciplinary relationship in which ideas from one discipline are integrated into the basic ideas and model of another; in this case linguistic thinking is integrated into basic information scientific problems and models. The links between information scientific and linguistic perspectives mean that they acquire supplementary functions and generate generalizable results. Because of its relationship to linguistics, philosophy of language and semantic theory the cognitive paradigm in information science represents a move towards an interdisciplinary relationship. It represents an important shift from the multidisciplinary relationship associated with the physical paradigm. I tentatively suggest that the consolidatory phase (see Figure 8) could be interpreted as a continuation of the interdisciplinary tendencies of the cognitive paradigm.

Transdisciplinarity describes a situation or process in which a "donor" discipline delivers ideas, theories and concepts to a "receiving" discipline and in which the narrow focus of a specific disciplinary structuring of reality is replaced by an advanced synthesis taking place in the receiving discipline's knowledge system. The receiving discipline changes, essentially through a process of systematically incorporating the methods and theories of the donor discipline into a new knowledge system. In other words, there is a restructuring of the receiving discipline rather than the building of bridges between two disciplines. The computational paradigm's assimilation of linguistic objects and terminology can be viewed as a transdisciplinary relationship: linguistic concepts are being absorbed and a new knowledge system is emerging. The data presented here cannot address the question of whether there was a direct shift from multidisciplinarity to transdisciplinarity as a result of the break from the physical paradigm or whether there was an intervening transition phase of interdisciplinarity. There is certainly no necessary, sequential relationship between the three types of interdisciplinarity.

The reformulation of the information science–linguistics relationship as an interdisciplinary relationship appears to be connected to the concept of information science as a "case science" that is, a discipline with a pronounced emphasis on real-world problems. I conclude this section with some reflections on this interesting aspect of the relationship between information science and linguistics.

A defining feature of real-world cases is the emphasis on holistic phenomena in real-world settings; it is the whole phenomenon or situation, not just selected parts of it, which is the object of study. This insistence on the ontological complexity of the world has to be reflected in the scientific methods used to investigate it, which has the important methodological consequence that all potentially relevant variables have to be considered in the description and investigation of problems. Variables that are not thought to be relevant and variables that do not fall within a discipline cannot be eliminated from the investigation. Research in established disciplines tends to follow a very different approach, perhaps most obviously in the practice of reducing or abstracting problems to eliminate apparently "external" phenomena or phenomena "belonging" to other scientific disciplines. This process of abstraction allows established disciplines to develop precise, discipline-specific models and unambiguous, causal explanations (cf. Krohn, 2010). From a science theoretic view it is, of course, a legitimate strategy.

If we accept a distinction between real-world cases and a "disciplined" (reduced, amenable to precise description) world, then the physical paradigm obviously positions information science (or more specifically IR) as a real-world case science. Modern information science and retrieval has defined itself by reference to practical problems since the 1950s. This trend was accelerated by computerization and the explosion of data, which led to demands for new and more effective methods of IR (Sparck Jones & Kay, 1973, p. 9f). The turning towards linguistics of the physical paradigm in information science was a consequence of this case orientation, because when considered as a real-world problem IR could not be abstracted from its linguistic components. Information scientists had to recognize that dealing with real-world, language-related linguistic phenomena demanded linguistic expertise. Interpreting information scientific overtures to linguistics as a consequence of a case-based approach fits very well with my earlier argument that there is a multidisciplinary relationship between the physical paradigm of IR and linguistics.

If we pursue the distinction between case-based methods and disciplined or reductive methods then the cognitive paradigm (which is characterized by, among other things, the information scientist's questioning of the theoretical basis of her discipline) can be viewed as an attempt to establish information science as a "proper" discipline with discipline-specific models, strictly delimited objects of study and a set of formal criteria by which theories and methods are evaluated. A fully fledged discipline need not be bound to fuzzy, ever-changing real-world cases. The task of linguistics is, seen from this angle, to help the information scientist into a dialog with critical meaning theories and language analytic philosophy. Again, identifying the cognitive paradigm (including the consolidation phase) with the internal stabilization of information science as a "proper," not exclusively case-based discipline fits well with the notion of an interdisciplinary relationship between cognitive information science and linguistics. The cognitive paradigm's acknowledgement of the inherently communicative and linguistic attributes of information and information activities and the interdisciplinary attempt to integrate linguistic concepts into information scientific theory paved the way for information science to develop into a fully fledged discipline because it enabled the field to transcend the case-based approach inherited from the physical paradigm.

DOI: 10.1002/asi

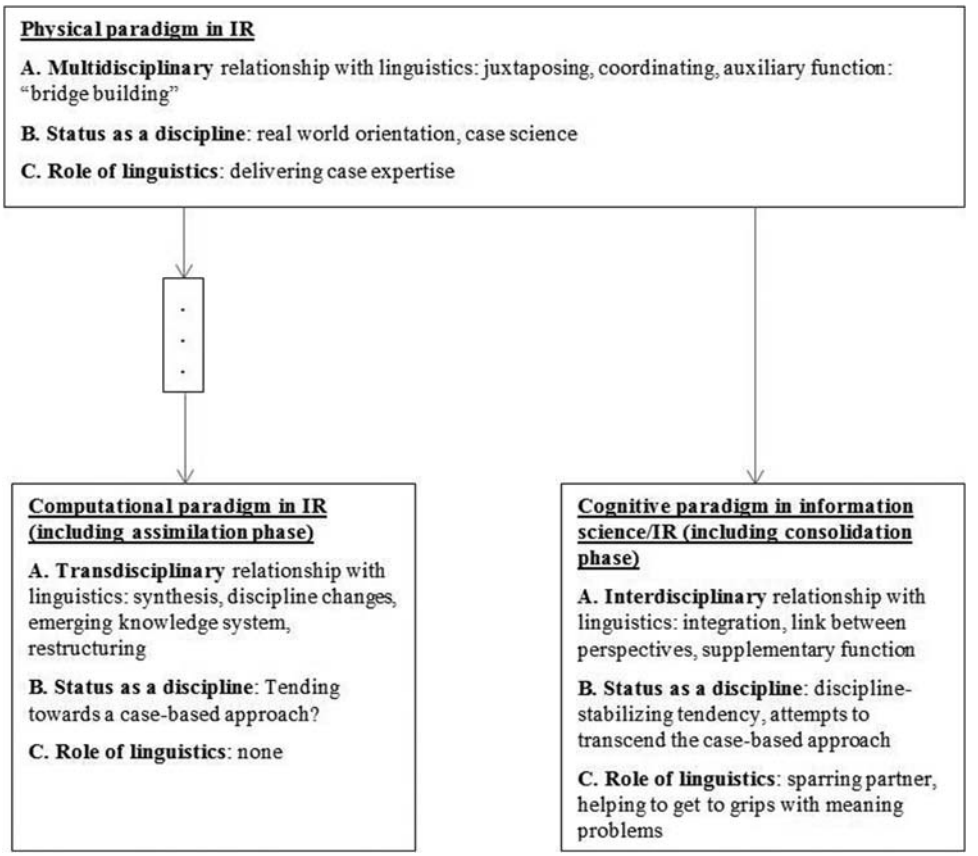

FIG. 9. Characterization of the three paradigms in information science/retrieval according to (A) interdisciplinary status, (B) status as a discipline, and (C) the role of linguistics.

How should we relate the assimilation phase, which sits within the computational paradigm, to the concepts of transdisciplinarity, cases, and the stabilization of "proper" disciplines? There is no doubt that it represents the evolution of a new, linguistics-like technological discipline, thus meeting one of the criteria for transdisciplinarity. However, the assimilation phase retains a certain emphasis on dealing with cases, which suggests a hybrid status. The exact status of the assimilation phase must remain a subject for future research.

The main arguments of this article are summarized in Figure 9,[9] below. The A-fields describe the relationship between information science and retrieval and linguistics in terms of interdisciplinary research, tracing the development from multidisciplinarity to transdisciplinarity (computational paradigm) and from multidisciplinarity to interdisciplinarity (cognitive paradigm). The B-fields indicate a paradigm's status as a discipline with reference to a real-world, case-oriented science. Here we can trace a single line of development, from case science to fully fledged discipline (cognitive paradigm); but the status of the computational paradigm is more debatable. The C-field is intended to capture how linguistics has helped information science to develop as a discipline. The case-based approach of the physical paradigm positioned linguistics as a problem-solver, but as information science endeavored to establish itself as a fully fledged discipline it assumed the role of a sparring partner (Figure 9).

Conclusions

This articles characterizes the scientific structure of information science and describes the disciplinary developments within the field of information science and retrieval in terms of (a) paradigms (physical; cognitive; computational), (b) methodological struggles (scientific and humanities traditions), and (c) significant disagreements and dilemmas relating to understanding of the core concept of "information" (materialist, empirical stance vs. concept of information as a meaning phenomenon). Tredinnick's fairly general claims about internal developments and dilemmas in information science have been substantiated by reference to the specifics of the relationship between information science and retrieval and linguistics.

The discussion showed that all three frames of reference—paradigm, methodology, and stance on the meaning dilemma—have some explanatory power in relation to how the information science and retrieval-linguistics relationship has evolved in the past 60 years. The transition from the physical paradigm to the cognitive represents, from a more general perspective, a significant attempt to tackle the problem of meaning in information both

DOI: 10.1002/asi

methodologically and theoretically; it was accompanied by a distinct shift in the role of linguistics, from provider of technical support to sparring partner promising a solution to the meaning problem in information. The shift to a more intellectual and philosophical approach to the incorporation of linguistic and communicative ideas into information science promised to enable the discipline to come to grips with language-related phenomena in IR systems both in the indexing component and from the perspective of the "new" participant, the "system user as language user." The move from physical to cognitive approaches in information science also constituted a move away from a solely case-based science to a fully fledged discipline, no longer restricted by the exigencies of real-world cases. This shift was reflected in the changing attitude of information science and retrieval to language and linguistics. The shift from a multidisciplinary to an interdisciplinary relationship resulted in a renaissance in the use of language philosophical concepts in information science; this promoted a deeper understanding of the communicative and linguistic roots of what thus became a more humanistic discipline. As an aside I should note that the computer science branch of information science has achieved transdisciplinary status: linguistic concepts have been assimilated into this emerging, new discipline. The emergence of a new discipline as a result of the integration and adaptation of core concepts from another discipline can be followed in the analysis of the sample of publications presented in this article.

To conclude, the investigation presented here confirms the special status of linguistics as a meaning-based partner discipline of information science, assisting information science and retrieval in the ongoing process of rediscovering and revitalizing its roots in communication and language. We should anticipate that linguistic ideas and concepts will continue to catch the attention of the critical information scientist; however it is much less clear what developments in linguistics will be strong and powerful enough to inform information scientific thinking. Two more recent attempts to integrate linguistic principles into information scientific contexts were discussed, the syntax-semantics distinction and the concept of a syntagmatic and a paradigmatic axis in language structure; both draw heavily on well-established, traditional categories in language theory and description. Perhaps this is not surprising; when choosing products in a shop one has not visited before one will tend to select familiar, established brands. However linguistics offers a multitude of theoretical approaches to language that could be used to understand information phenomena. I therefore look forward to exploring modern concepts and theories in linguistics with a view to their potential applications in information studies.

The best way to renew the information science–linguistics relationship and to identify new linguistic inputs to information science is, nevertheless, formal interdisciplinary collaborative research. To return to the shop metaphor, when a family is buying the ingredients for an evening meal it is important that at least one person is familiar with each shop. It is during the cooking process, when the ingredients are combined, that a new and hopefully tasty dish is created. Both the preparation of the ingredients and the enjoyment of the resulting food can be social experiences. This, or something like this, could be a model for fruitful interdisciplinary cooperation between information scientists and linguists.

Acknowledgments

I want to thank my two anonymous reviewers, who contributed enormously to a much better and clearer text. Any faults and errors are solely my responsibility.

Endnotes

1. This is the label used by Tredinnick (2006, p. 63). The approach used in this period has also been described as "technological" or "systemic" (Hjørland, 1998b, p. 610f).

2. Here the term *real-world* is used to describe research in which real-world phenomena are studied in their full complexity, that is, without abstracting them from extradisciplinary variables and influences (cf. Krohn, 2010). The term *case-based*, which is taken from interdisciplinarity research, is used in a similar way. This issue is discussed in detail in a later section, "Perspectives From Interdisciplinarity Research."

3. But note that the physical-cognitive distinction was first proposed by Ellis in "The Physical and Cognitive Paradigms in Information Retrieval Research" (Ellis, 1992).

4. *Linguistics and information science* (Sparck Jones & Kay, 1973) provides the main statement of the information science-linguistics relationship in this early period and illustrates the attempt to transcend disciplinary boundaries.

5. It is worth noting here that information scientists have always taken a skeptical attitude towards their discipline's theoretical and methodological condition, as this quote from Hjørland illustrates: "It is a well-known fact that information science lacks good theories. Most work is of a pragmatic nature, which resists scientific analysis and generalisation" (Hjørland, 1998b, p. 607).

6. Roughly speaking the emergence of the cognitivist preference for a communication-oriented conception of the system user and the return to the principal linguistic fundamentals of information science occurred in the 1980s and 1990s. It was strongly connected with Blair's research, in particular his book *Language and Representation in Information Retrieval* (Blair, 1990). This can be considered the second main statement of the information science–linguistics relationship.

7. A common feature of many, if not all of these research areas is that their research objects (research, information literacy, etc.) have a linguistic–communicative, meaning-related dimension (see Floridi, 2011).

8. The distinction between "linguistic1" and "linguistic2" is abandoned henceforward as it is of no significance to what follows. The term *linguistic* is used in its conventional sense, which approximates "linguistic1."

9. For the sake of simplicity the cognitive paradigm and the consolidation phase are combined in one box as the distinction between them is not important for the three attributes in the box.

References

Antoniou, G., Groth, P., van Harmelen, F., & Hoekstra, R. (2012). A semantic web primer (3rd ed.). Cambridge, MA: MIT Press.

Aronoff, M., & Rees-Miller, J. (2007). The handbook of linguistics. Oxford: Blackwell.

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). Modern information retrieval: The concepts and technology behind search (2nd ed.). New York: Addison Wesley.

DOI: 10.1002/asi

Bawden, D., & Robinson, L. (2012). An introduction to information science. London: Facet.
Beghtol, C. (1986). Bibliographic classification theory and text linguistics: Aboutness analysis, intertextuality and the cognitive act of classifying documents. Journal of Documentation, 42(2), 84–113.
Belkin, N.J. (1980). Anomalous states of knowledge as a basis for information retrieval. Canadian Journal of Information. Science-Revue Canadienne Des Sciences De L'Information, 5, 133–143.
Blair, D.C. (1990). Language and representation in information retrieval. Amsterdam: Elsevier Science.
Blair, D.C. (1992). Information retrieval and the philosophy of language. The Computer Journal, 35(3), 200–207.
Blair, D.C. (2003). Information retrieval and the philosophy of language. Annual Review of Information Science and Technology, 37, 3–50.
Blair, D.C. (2006). Wittgenstein, language and information: "Back to the rough ground!" Dordrecht: Springer.
Blair, D.C., & Kimbrough, S.O. (2002). Exemplary documents: A foundation for information retrieval design. Information Processing & Management, 38(3), 363–379.
Borgman, C.L. (2007). Scholarship in the digital age: Information, infrastructure, and the Internet. Cambridge, MA: MIT Press.
Borlund, P. (2003). The concept of relevance in IR. Journal of the American Society for Information Science and Technology, 54(10), 913–925.
Borsley, R.D. (1999). Syntactic theory, a unified approach (2nd ed.). London: Arnold.
Broughton, V. (2006). Essential thesaurus construction. London: Facet.
Butler, C.T.L. (2003). Structure and function: A guide to three major structural-functional theories. Amsterdam: J. Benjamins.
Chaka, C. (2010). E-learning 2.0: Web 2.0, the semantic web and the power of collective intelligence. In H.H. Yang, & S.C. Yuen (Eds.), Handbook of research on practices and outcomes in e-learning: Issues and trends (pp. 38–60). Hershey, PA: Information Science Reference.
Chierchia, G., & McConnell-Ginet, S. (2000). Meaning and grammar: An introduction to semantics (2nd ed.). Cambridge, MA: MIT Press.
Chomsky, N. (2002). Syntactic structures (2nd ed., with an introduction by David W. Lightfoot). Berlin: Mouton de Gruyter.
Chowdhury, G.G. (2010). Introduction to modern information retrieval (3rd ed.). London: Facet.
Cooper, W.S. (1971). A definition of relevance for information retrieval. Information Storage and Retrieval, 7(1), 19–37.
Eco, U. (1991). Semiotics and the philosophy of language. London: Macmillan.
Eco, U. (1995). The search for the perfect language. Oxford: Blackwell.
Ellis, D. (1996). Progress and problems in information retrieval. London: Library Association Publishing.
Engerer, V. (2012). Informationswissenschaft und Linguistik. Kurze Geschichte eines fruchtbaren interdisziplinären Verhältnisses in drei Akten. SDV – Sprache und Datenverarbeitung. International Journal for Language Data Processing, 36(2), 71–91.
Engerer, V. (2014a). Indexierungstheorie für Linguisten. Zu einigen natürlichsprachlichen Zügen in künstlichen Indexsprachen. In M. Schönenberger, V. Engerer, P. Öhl, & B. Brogyanyi (Eds.), Dialekte, Konzepte, Kontakte. Ergebnisse des Arbeitstreffens der Gesellschaft für Sprache und Sprachen, GeSuS e.V., 31. Mai – 1. Juni 2013 in Freiburg/Breisgau (pp. 61–74). Jena: GeSuS.
Engerer, V. (2014b). Thesauri, Terminologien, Lexika, Fachsprachen. Kontrolle, physische Verortung und das Prinzip der Syntagmatisierung von Vokabularen. Information, Wissenschaft & Praxis, 65(2), 99–108.
Floridi, L. (2011). The philosophy of information. Oxford: Oxford University Press.
Frodeman, R., Klein, J.T., & Mitcham, C. (2010). The Oxford handbook of interdisciplinarity. Oxford: Oxford University Press.
Frohmann, B. (1990). Rules of indexing: A critique of mentalism in information retrieval theory. Journal of Documentation, 46(2), 81–101.
Fugmann, R. (2002). The complementarity of natural and index language in the field of information supply: An overview of their specific capabilities and limitations. Knowledge Organization, 29(3/4), 217–230.
Gharaibeh, I.K., & Gharaibeh, N.K. (2012). Towards Arabic noun phrase extractor (ANPE) using information retrieval techniques. Software Engineering, 2(2), 36–42.
Green, R. (1995a). Syntagmatic relationships in index languages: A reassessment. Library Quarterly, 65(4), 365–385.
Green, R. (1995b). The expression of conceptual syntagmatic relationships: A comparative survey. Journal of Documentation, 51(4), 315–338.
Green, R. (2002). The semantics of relationships, an interdisciplinary perspective. Dordrecht: Kluwer Academic.
Grice, P. (1975). Logic and conversation. In P. Cole & J.L. Morgan (Eds.), Syntax and semantics: Speech acts (pp. 41–58). New York: Academic Press.
Grice, P. (1989). Studies in the way of words. Cambridge, MA: Harvard University Press.
Harter, S.P. (1992). Psychological relevance and information science. Journal of the American Society for Information Science, 43(9), 602–615.
Hjørland, B. (1998a). Information retrieval, text composition, and semantics. Knowledge Organization, 25(1/2), 16–31.
Hjørland, B. (1998b). Theory and metatheory of information science: A new interpretation. Journal of Documentation, 54(5), 606–621.
Joseph, J.E. (2012). Saussure. Oxford: Oxford University Press.
Klein, J.T. (2010). A taxonomy of interdisciplinarity. In R. Frodeman, J.T. Klein, & C. Mitcham (Eds.), The Oxford handbook of interdisciplinarity (pp. 15–30). Oxford: Oxford University Press.
Krohn, W. (2010). Interdisciplinary cases and disciplinary knowledge. In R. Frodeman, J.T. Klein, & C. Mitcham (Eds.), The Oxford handbook of interdisciplinarity (pp. 31–49). Oxford: Oxford University Press.
Kuhlthau, C.C. (2004). Seeking meaning: A process approach to library and information services (2nd ed.). Westport/Connecticut, London: Libraries Unlimited.
Lancaster, F.W. (2003). Indexing and abstracting in theory and practice (3rd ed.). London: Facet.
Larson, R.R. (2010). Information retrieval systems. Encyclopedia of library and information sciences (pp. 2553–2563). New York: Taylor & Francis.
Levinson, S.C. (2003). Pragmatics. Cambridge: Cambridge University Press.
Li, L. (2009). Emerging technologies for academic libraries in the digital age. Oxford: Chandos.
Lyons, J. (1977). Semantics. London: Cambridge University Press.
Mai, J. (1999). Deconstructing the indexing process. Advances in Librarianship, 23, 269–298.
Pandey, R.C. (2003). Information retrieval system. A linguistic study. Delhi: Abhijeet Publications.
Pandey, R.C. (1997). Information retrieval system: A linguistic approach. International Information Communication and Education, 16(1), 14–30.
Petöfi, J.S. (1969). On the problems of co-textual analysis of texts. International Conference on Computational Linguistics, Preprint No. 50.
Petöfi, J.S. (1971). "Generativity" and text-grammar. Göteborg: University of Gothenburg, Department of Linguistics.
Petöfi, J.S., & Bredemeier, J. (1977). Das Lexikon in der Grammatik - die Grammatik im Lexikon. Hamburg: Helmut Buske.
Pickard, A.J. (2013). Research methods in information (2nd ed.). London: Facet.
Ruthven, I., & Kelly, D. (2011). Interactive information seeking, behaviour and retrieval. London: Facet.
Saussure, F.D. (1967/2011). Course in general linguistics (W. Baskin, Trans.). New York: Columbia University Press. (Reprinted from Grundfragen der allgemeinen Sprachwissenschaft. Berlin: Walter de Gruyter.)
Saussure, F.D. (1967). Grundfragen der allgemeinen Sprachwissenschaft. In C. Bally, A. Sechehaye, A. Riedlinger, H. Lommel, & and P.V. Polenz (Eds.). Berlin: Walter de Gruyter.
Searle, J.R. (1975). A taxonomy of illocutionary acts. In K. Gunderson (Ed.), Language, mind and knowledge (pp. 344–369). Minneapolis, MN: University of Minnesota Press.
Searle, J.R. (1985). Speech acts: An essay in philosophy of language (Reprint). Cambridge: Cambridge University Press.
Seuren, P.A.M. (1996). Semantic syntax. Oxford, UK: Blackwell.
Sözer, E. (Ed.). (1985). Text connexity, text coherence: Aspects, methods, results. Hamburg: Buske.

Sparck Jones, K., & Kay, M. (1973). Linguistics and information science. New York, London: Academic Press.
Sperber, D., & Wilson, D. (1995). Relevance: Communication and cognition. Oxford: Blackwell.
Svenonius, E. (2000). The intellectual foundation of information organization. Cambridge, MA: MIT Press.
Tedd, L.A. (2005). Digital libraries: Principles and practice in a global environment. Berlin: Saur.
Tredinnick, L. (2006). Digital information contexts: Theoretical approaches to understanding digital information. Oxford: Chandos Publishing.
Van Rijsbergen, C.J. (1986). A nonclassical logic for information retrieval. The Computer Journal, 29(6), 481–485.
Walker, D.E., Karlgren, H., & Kay, M. (Eds.). (1977). Natural language in information science: Perspectives and directions for research. Stockholm: Skriptor.
Warner, J. (2007a). Analogies between linguistics and information theory. Journal of the American Society for Information Science and Technology, 58(3), 309–321.
Warner, J. (2007b). Linguistics and information theory: Analytic advantages. Journal of the American Society for Information Science and Technology, 58(2), 275–285.
Warner, J. (2010). Human information retrieval. Cambridge, MA: MIT Press.
Weinberg, B.H. (2009). Indexing: History and theory. In M.J. Bates & M.N. Maack (Eds.), Encyclopedia of library and information sciences (3rd ed.) (pp. 2277–2290). New York: Marcel Dekker

Appendix: Sample Details

(Remarks on notation: The sample-defining linguistic keyword SEMANTICS always appears in **bold**; other usage of bold font in the database has been eliminated, although capitalization is preserved. The phrase "X of Y references" indicates the number of references in a reference list that were defined as linguistic1 (X) and the total number of references in the list (Y). For articles where $X > 0$ a list of the presumed linguistic1 records is given. Complete reference lists are available in LISTA or other electronic resources.)

Record no.	Record	Set of keywords	Proportion of linguistic1 references	Details of linguistic1 references*
1	Jörgensen, C., Stvilia, B., & Shuheng, W. (2014). Assessing the Relationships among Tag Syntax, Semantics, and Perceived Usefulness. Journal of the Association for Information Science & Technology, 65(4), 836–849.	ABSTRACTING & indexing services; RESEARCH – Methodology; METADATA; RESEARCH; **SEMANTICS**; SUBJECT headings; IMAGE retrieval; CONTENT mining; CROWDSOURCING; CHI-squared test; PHOTOGRAPHY; PROBABILITY theory; FINANCE; SCALE analysis (Psychology); STATISTICS; DESCRIPTIVE statistics	0 of 67	
2	Gonzales, B. M. (2014). Linking Libraries to the Web: Linked Data and the Future of the Bibliographic Record. Information Technology & Libraries, 33(4), 10–22.	ACCESS to information; CATALOGING; LIBRARIES; LIBRARY automation; METADATA; **SEMANTICS**; WORLD Wide Web; SYSTEMS development; METHODOLOGY	0 of 26**	
3	Tsakonas, G., Mitrelis, A., Papachristopoulos, L., & Papatheodorou, C. (2013). An exploration of the digital library evaluation literature based on an ontological representation. Journal of the American Society for Information Science & Technology, 64(9), 1914–1926.	DIGITAL libraries **–** Evaluation**;** INFORMATION science**;** CLASSIFICATION; **SEMANTICS**; METHODOLOGY	0 of 41	

DOI: 10.1002/asi

Appendix Table. *Continued*

Record no.	Record	Set of keywords	Proportion of linguistic1 references	Details of linguistic1 references*
4	Darányi, S., & Wittek, P. (2013). Demonstrating conceptual dynamics in an evolving text collection. Journal of the American Society for Information Science & Technology, 64(12), 2564–2572.	INFORMATION retrieval; **SEMANTICS**; ELECTRONIC publications; DATA analysis; LINGUISTICS – Methodology; PRESS; STATISTICS; TIME	5 of 79	Baker, A. (2008). Computational approaches to the study of language change. Language and Linguistics Compass, 2(2), 289-307. >< Trier, J. (1934). Das sprachliche Feld. Neue Jahrbücher für Wissenschaft und Jugendbildung, 10, 428-449. >< Eijck, J. & Visser, A. (2010). Dynamic semantics: The Stanford encyclopedia of philosophy. >< Cruse, DA. (1986). Lexical semantics. Cambridge. >< Lehrer, A. (1975). Semantic fields and lexical structure. New York.
5	Choi, Y. (2013). Analysis of image search queries on the web: Query modification patterns and semantic attributes. Journal of the American Society for Information Science & Technology, 64(7), 1423–1441.	INFORMATION retrieval; INTERNET; **SEMANTICS**; SEARCH engines; INFORMATION-seeking behavior; CHI-squared test; COLLEGE students; PHOTOGRAPHY; MEDICAL coding	0 of 71	
6	Muresan, S., & Klavans, J. L. (2013). Inducing terminologies from text: A case study for the consumer health domain. Journal of the American Society for Information Science & Technology, 64(4), 727–744.	CLASSIFICATION; ALGORITHMS; INFORMATION retrieval; MEDICINE – Information services; RESEARCH; **SEMANTICS**; SUBJECT headings; REFERENCE sources; INFORMATION services; LANGUAGE & languages; COMPARATIVE grammar; FINANCE; CONSUMERS	5 of 32	Nirenburg, S. & Raskin, V. (2004). Ontological semantics. Cambridge, Mass. >< Steedman, M. (1996). Surface structure and interpretation. Cambridge, MA. >< Steedman, M. (2000). The syntactic process. >< Joshi, A.K. & Schabes, Y. (1997). Tree-adjoining grammars: Handbook of formal languages. Berlin, 69-123. >< Miller, G. (1990). WordNet: An online lexical database. International Journal of Lexicography. 3(4), 235-312.
7	Guo, L., & Wan, X. (2012). Exploiting syntactic and semantic relationships between terms for opinion retrieval. Journal of the American Society for Information Science & Technology, 63(11), 2269–2282.	INFORMATION retrieval; RESEARCH; **SEMANTICS**; COMPARATIVE grammar; PROBABILITY theory; PUBLIC opinion; FINANCE	0 of 7	

DOI: 10.1002/asi

Appendix Table. *Continued*

Record no.	Record	Set of keywords	Proportion of linguistic1 references	Details of linguistic1 references*
8	1. Groza, T., Aastrand Grimnes, G., & Handschuh, S. (2012). Reference Information Extraction and Processing Using Conditional Random Fields. Information Technology & Libraries, 31(2), 6–20.	COMPUTERS; EXPERIMENTAL design; INFORMATION retrieval; INTERNET; METADATA; LIBRARY reference services; RESEARCH; **SEMANTICS**; FINANCE; SCIENCE	0 of 11	
9	Leung, R., McGrenere, J., & Graf, P. (2011). Age-related differences in the initial usability of mobile device icons. Behaviour & Information Technology, 30(5), 629–642.	EXPERIMENTAL design; POCKET computers; RESEARCH; **SEMANTICS**; USER interfaces (Computer systems); WIRELESS communication systems; QUALITATIVE research; AGING; ANALYSIS of variance; CORRELATION (Statistics); FINANCE; SCALE analysis (Psychology); VISUAL perception; PRODUCT design	0 of 30	
10	Sabucedo, L. Á., & Rifón, L. A. (2010). Managing Citizen Profiles in the Domain of e-Government: The cPortfolio Project. Information Systems Management, 27(4), 309–319.	INTERNET in public administration; RECORDS management – Computer network resources; PERSONAL information managers; CONFIDENTIAL records; ARCHIVES – Access control; DIGITAL preservation; DIGITIZATION of archival materials; **SEMANTICS**; INTERNET	0 of 7	
11	Krull, R., & Sharp, M. (2006). Visual verbs: Using arrows to depict the direction of actions in procedural illustrations. Information Design Journal (IDJ), 14(3), 189–198.	SIGNS & symbols; **SEMANTICS**; COMMUNICATION of technical information; NONVERBAL communication; THREE-dimensional imaging; GRAPHIC arts; AMBIGUITY	1 of 20	Glenberg, A. (2002). The indexical hypothesis: Meaning from language, word, and image.: Words and Images: Working Together - Working Differently. Westport, Connecticut, 27–42.
12	Hudon, M., Mas, S., & Gazo, D. (2005). Structure, Logic, and Semantics in Ad Hoc Classification Schemes Applied to Web-Based Libraries in the Field of Education. Canadian Journal of Information & Library Sciences, 29(3), 265–288.	LIBRARIES & education; LIBRARIES; WEB (Computer program language); **SEMANTICS**; INFORMATION theory; EDUCATION; CLASSIFICATION; LANGUAGE & languages; LOGIC	0 of 29	

*Note: Linguistics1 references in the sample publications have not been adapted to any specific reference style.

**Probably fewer than 26 separate publications as the list seems to include several citations of identical titles, referring to different pages.

DOI: 10.1002/asi