DOI: 10.1002/pra2.246

Searching Covid-19 by linguistic register: Parallels and warrant for a new retrieval model

G. Benoît

School of Information, University of California, Berkeley, Berkeley, California, USA

Correspondence
G. Benoît, School of Information, University of California, Berkeley, Berkeley, CA 94720.
Email: gbenoit@berkeley.edu; gb@bix.digital

Abstract

Keeping informed given rapid trend in data and resources about covid-19 is a new challenge. Different user groups (researchers/doctors, practitioners, public) vary in linguistic expression and vocabulary so a new retrieval framework might likewise vary to improve retrieval, expose unanticipated concepts, and establish a sustainable research stream. In this project a document collection about covid-19 was created, parsed according to ISO12620's definition of linguistic register, and retrieval sets compared. Results suggest trends from other fields parallel register-oriented criteria; project exposes unexpected concepts across groups, uses of visualization, and warrants ling-register as a sustainable IR research stream.

KEYWORDS

and clustering, ISO12620, linguistic register, visualization

1 | INTRODUCTION

The challenges of producing data resources, sharing them, and sifting actionable data from presidential "sarcasm" raise novel challenges. Certainly the rapid developments of covid-19 provide challenge and opportunity for evaluating different techniques for public awareness. In this paper we suggest an interdisciplinary model for retrieval engine design that is sensitive to the seekers' own linguistic registers and information needs, using covid-19 as the example. The results suggest strongly that applying ISO12699:2019 as part of a retrieval model reveals unanticipated concepts in the collection and exposes fluidity in term-use, as end-user linguistic performance. Applying visualization techniques, we see the impact of linguistic register on distribution by domain and purpose. We conclude that developing linguistic register-oriented IR and visualization is a promising approach to a new IR model.

Recent efforts (Spring 2020) to keep the public and health care practitioners informed with covid-19 data that lead to healthier action in their daily lives have led to a confusing miasma of contradictory facts and mix of research reports, guidelines, blogs and editorials. Looking for ways to improve the comprehensibility and appropriateness of resource location, we consider a "linguistic register" (LR-IR, or "ling-reg") approach, which turns out to be not without some controversy.

The usual interpretation of linguistic register seems to be merely identifying parts of speech as a natural language project or gross-level syntax parsing, shifting IR into computational linguistics. But an ISO standard on "ling reg" seems appropriate also for IR. As an understudied area we feel it is important first to identify briefly parallels between IR practice and LR-IR and next to suggest IR framework models that would be sensitive to a semantic-level linguistic-oriented one. ISO12620:2009 proposes a "Terminology Classification Model" titled

83rd Annual Meeting of the Association for Information Science & Technology October 25-29, 2020.
Author(s) retain copyright, but ASIS&T receives an exclusive publication license

"Specification of data categories and management of a Data Category Registry for language resources" (https://www.iso.org/standard/37243.html). The ideas are explored in a project implementing some of the linguistic behaviors defined in ISO12620:2009 to determine whether LR-IR is a sustainable research program.

We start by describing the theory of register and warrant such an approach for IR. The project to be described integrates "linguistic registers" in retrieval research to explore the impact on retrieval set membership, ranking options, and presentation opportunities to optimize ranking results presented the end-user.

2 | PREVIOUS WORK - THEORIES OF REGISTER AND RETRIEVAL, COMPREHENSIBILITY AND REASONS FOR EXPLORATION

Consider for instance a corpus of medical terms derived from various literatures, representing scientific, practice, and consumer health. The topics are the same across groups and language register naturally varies by group, as discourse markers and long evidenced in corpora studies (Brizuela, Andersen, & Stallings, 1999; Chiu, 1972; Wynne, 2005) and the lexicography of professional practice (Kennedy, 1987; McCrae et al., 2012; David and Gardner (2013). Such studies show how terms are distributed differently based on large-level sociolinguistic and practice domains, such as "Science", "Social Science", "Humanities" (Arazy & Woo, 2007; Hyland & Tse, 2007; Prieto Valesco, 2013; Witt, Heid, Sasaki, & Sérasset, 2009). Domain-less general assumptions on Zipf's Law for determining distribution now seems tenuous (Ferrer-i-Cancho & Elvevåg, 2010). We wonder, then, what parallels are there across interested domains that can improve relevancy retrieval?

Given an otherwise unbound collection of English-language documents, we expect terms to appear with calculable frequencies but in a single domain-oriented collection (such as medicine or a popular topic, such as covid-19), there are significant differences in term frequencies (Liu, 2011). Alonge et al. (1998) and Haspelmath (2019) suggest professional term choices in linguistic register are notably different, affecting retrieval behaviors, that remain an under-evaluated area for retrieval work (Giménez-Moreno & Skorcynska, 2013; Speelman, Grondelaers, & Geeraerts, 2003). The ISO standard for linguistic register (ISO-12069) provides a way to bridge these themes and to create a weight schema as we would in any information retrieval project. Note that linguistic register defined here is *emphatically not* the "parts of speech" or syntactic analysis used in computational linguistics or in natural language processing. Rather it is a sense of the difference between semantic ambiguity of terms, domains of variation, expressivity, and "fuzziness" of comprehension (Heylighen & Dewaele, 1999).

As an information retrieval question tailoring work from the above literatures into the IR model lead to what weighting schemes that are useful in improving retrieval ranking and relevancy. Relevancy might be a measure of "comprehension of the results," something suggested by Heylighen and Dewaele (1999) and expanded upon by Uccelli, Galloway, Barr, Meneses, and Dobbs (2015), by identifying differences between academic language proficiency and its association with reading comprehension. Work at the nexus of comprehension and vocabulary seems to be limited only to general language knowledge (e.g., L1, L2, L3 for language skills) and little in the area of measurable "vocabulary depth" when readers work outside their usual modes of expression and domain.

It is interesting to note en passant the parallels to speech theory's linguistic performance based on a person's normative role. A medical practitioner in her role as doctor assumes a linguistic behavior that will differ when shifting to another role, say parent at a teacher conference.

3 | LINGUISTIC REGISTERS AND INFO RETRIEVAL

Since the 1970s there has been work articulating "registers" as "the linguistic features which are typically associated with a configuration of situational features – with particular values of field, mode, and tenor" (Halliday & Hasan, 1976). "Field" (the total event in which the text functions, indicating the purpose of the speech and subject-matter), "mode" (the function of the text, including the channel and genre, narrative, etc.), and "tenor" (type of role interaction), with "set of *relevant social relations*, permanent and temporary, among the participants involved" collectively define the linguistic features of a text. Register, then, has coherence in respect to the *context* of the situation of its utterance and coherent in itself - that is the utterance in its register enables semantically and syntactically useful phrases that enable message construction and receipt. In this way "ling reg" applies to both the *creation of* and *use* of documents.

By the 1990s "register" was used as a shorthand for mere "formal/informal styles" of speech; increasingly the term "style" is used in general cases, preserving "register" for linguistics, identifying domains of use. For instance, Joos describes register levels of "Frozen, Formal, Consultative, Casual, Intimate" for classifying and analyzing texts. Such levels, though, are more useful in identifying the relationship between speaker/hearer and to identify "back-channel behavior", terms that have no contribution to the text other than to indicate listening, such as "uh huh." It is akin to classification by topic and using subject headings but here the criteria is linguistic behavior.

Two contemporary attempts at standardization of registers are the Open Linguistics project and the ISO 12620:2009 standard. Unlike the earlier discourse-oriented models, OpenLinguistics defines models to associate semantic tokens and phrases with five registers (Open Linguistics, 2017), which map to ideas of the semantic web. The other is an ISO-created standard. The first version appeared in 1999, was updated in 2009, and a replacement ISO/CD 12620 is planned. While a new standard is being discussed, ISO 12620 remains a

TABLE 1 Normative Registers

Normative references	*In the standard, "normative references" indicates other ISO standards used in 12620. It applies in this essay to refer to the context or norms during a speech action.*
Phraseological unit	*"data category term and contains a term or other information treated as if it were a term (e.g., phraseological units and standard text)" (p. 4); "A.2.1.18 phraseological unit*
	Description: Any group of two or more words that form a unit, the meaning of which frequently cannot be deduced based on the combined sense of the words making up the phrase. Note: Although they are made up of more than one word and frequently contain more than one concept, phraseological units can be treated as individual terminological units in terminology databases. In this sense they are grouped together with "terms". They can, however, also be treated as contextual material in some databases. See examples in A.2.1.18.1-A.2.1.18.3." (p. 9)
Set Phrase	a fixed lexicalized phrase, e.g., "handle with care"
Synonymous phrase	Phraseological unit in a language that expresses the same semantic content as another phrase in that same language. Example: The phrases response to open flame exposure and effect of open flame exposure are treated as synonymous phrases in some fire standards.
Registers: neutral	"standard register" - appropriate to general texts or discourse
Academic	[from the 2009 standard]
Int'l scientific term	A.2.1.4 international scientific term [A.2.1.4]
Technical	appropriate to scientific texts or special languages
In-house	"The register of terms that are company-specific and not readily recognized outside this environment. Example: In-house usage at one automotive company for the automotive tuning characteristic gear rattle is crowds. Note: In-house terminology is not necessarily equivalent to bench-level terminology, inasmuch as the former can thrive at very high levels of research and development. In- house terminology is frequently the source of new technical terminology that eventually gains widespread acceptance on a broader scale."
Bench-level	aka: "shop term." "The register of terms used in applications-oriented as opposed to theoretical or academic levels of language. Example: The retrieval end of a broach is commonly called a puller in bench-level usage."
Dialect	
Facetious	
Formal	
Ironic	
Neutral	
Foreign languages	The link between one language's token in a given register to the target language's bench-level
Slang	"An extremely informal register of a word, term, or text that is used in spoken and everyday language and less commonly in documents. Example: In aviation, the phrase fly by the seat of your pants is slang for the more formal fly without instruments"
Taboo	[from the 2009 standard]
Vulgar	"The register of a term or text type that can be characterized as profane or socially unacceptable. Note: Although vulgar register is avoided in formal technical terminology, languages with broad distribution such as English or Spanish can require the documentation of problematic terms that vary in register from region to region."

Note: Definitions of registers from http://semanticweb.kaist.ac.kr/org/tc37/pdocument/standards/ISO12620_1999.pdf.

intriguing model. First because it is an international standard and because the model can be applied to corpora such that we can identify semantic tokens (useful for retrieval models) and registers (suggestive of relevancy ranking and options for displaying results). Moreover a finer-grained understanding of semantic tokens provides more insight into end-user expressions of information needs and topic interests.

ISO 12620 02.03.03 standard states: "ISO 12620:2009 provides guidelines concerning constraints related to the implementation of a Data Category Registry (DCR) applicable to all types of language resources, such as terminological, lexicographical, corpus-based, machine translation. It specifies mechanisms for creating, selecting and *maintaining data categories*, *as well as an interchange format* for representing them." Table 1 enumerates the linguistic units and define the registers that indicate the "relative level of language individually assigned to a lexeme or term or to a text type [A.2.3.3]" (ISO, 2017).

Consider the example of two phrases "heart attack" and "myocardial infarction." The former may appear across the range of linguistic registers while the latter is likely reserved to international scientific and academic registers. Depending on the *users* of the phrase, either may be part of in-house and bench-level registers among health care providers; yet may be the most "technical" to the individual patient. There are few casual terms but they exist as potential search terms: heart attack and heart pangs, which could be classified as bench-level and dialect, respectively.

To an IR system, tokens stripped of other context or linguistic environment lose the benefits of their original articulation (be it oral or textual) and consequently tend to create more false-hits in a retrieval set. Or at least require more disambiguation to determine suitability For example an April 2020 Google search of various tokens returned sets containing expressions that included hits that were antithetical to each other; that is, as equivalent tokens but in linguistic register that were opposites or entirely unrelated. Table 2 presents more examples.

4 | LINGUISTIC REGISTER AND USER BEHAVIOR/LINGUISTIC PERFORMANCE:

As an IR question, documents are more relevant to the end-user's queries and needs when the expression of information need computationally is judged similar enough to the user's context and semantic levels. Interpreting the usefulness of documents is an issue of comprehensibility to be aided or hindered by the underlying IR model. Consequently the goal is to improve the likelihood of smaller, more relevant, that is comprehensible and situationally appropriate retrieval sets. Biber and Finegan (1994), Biber (1995), and Neumann (2014) demonstrate similarities of register use in several human languages and the importance of social contexts on a person's decision to use a particular register and to switch between them as needed in a discourse. As "information seeking" is not performed outside of any social context, it seems reasonable at least to allow this into considerations of information retrieval algorithm design. In parsing texts for retrieval, register variation provides a parallel to the

TABLE 2 Sample registers with weighting examples for retrieval ranking.

Register	Examples			
Technical	*Felis catus*	banknote	Diabetic ketoacidosis	
Academic	feline	currency		
Scientific				*Kalmia latifolia*
Bench-level	cat	pound	diabetes	mountain laurel
Dialect	kitty, gato, minou	pictures of the Queen		sheep's bane
Facetious	pussy, dude			
Formal	cat	Bank of England banknote	diabetes mellitus	
In-house	tang		low blood sugar	calico bush
Ironic	cool cat, tang			
Neutral	cat	pound		
Slang	lynx, chum	quid		
Taboo	moggie			
Vulgar	moggie			

TABLE 3 Example of dividing resources into ISO linguistic registers. Resources are clustered by genre (financial, entertainment, etc.), and media sources

In-house, Bench-level, Slang, Facetious, Ironic	Neutral	Academic; Int'l Sci Term; Technical	Neutral	Dialect, Facetious, Ironic
POP MAG:	NEWS	ACAD	Spoken	FIC
83,275 = 20.7%	79,368 = 19.7%	79,292 = 19.7%	81,690 = 20.3%	78,752 = 19.6%
News/Opinion	Misc-Int'l	History	ABC	GenBooks
Financial	Misc	Education	NBC	Jrnls
Scientific & Technical	Nat'l	Geog/SocSci	CBS	Sci/Fantasy
Soc/Arts	Local	Law/PoliSci	CNN	Juvenile
Religion	Money	Hum	FOX	Movies
Sports	Life	Phil/Rel	MSNBC	
Entertainment	Sports	Sci/Tech	PBC	
Home/Health	Editorial	Medicine	NPR	
African-American		Misc	Inde	

Reviewing the co-occurrence of terms between these sets shows out of 5,345 pairings of 60,000 lemmas		
Popular / News	**855**	**15%**
Fiction / Popular	726	13.5%
Fiction / Academic	627	11.7%
Spoken / Academic	551	10%
Fiction / News	534	10%
Spoken / Popular	502	9.4%
Spoken / Fiction	479	8.96%
Spoken / News	441	8.25%

already-employed metadata applied for weighting schemas and lexical descriptions in the parsed texts.

4.1 | Usual IR Model and potentially more linguistically-sensitive matching

An LR-IR model gains support from literature studies and cross-language IR in addition to those already mentioned, statistically-based ones based on corpus analysis are popular (Baeza-Yates & Ribierto-Neto, 2011; Oakes, 1998). One approach is to identify the collection's genre and to calculate the probable distribution ofterm as an empirical foundation for weighting schemes. In a study of 402,377 documents, David and Gardner (2013) note these differences in distribution (Table 3):

4.2 | And IR Models applied …

In the above project terms are identified by their part of speech and register. Queries are mapped to exact equivalents in each of the corpora's register divisions [academic, …], and a count of matching documents is returned. Terms of "neutral" register are mapped by similarity to terms from other registers. The resulting sum of weights determines strength of the terms' membership in the primary group. While Davies & Gardner's project aimed at genre discrimination, the methods parallel the arguably standard IR methods. We see this kind of work as a foundation to this project.

The usual characterization of IR models is "a quadruple [D, Q, $\mathscr{F}$, $R(q_i,d_j)$] where (a) D is a set composed of logical views (or representations) for the documents in the collection (b) Q is a set composed of logical views (or representations) for the user information needs. … (c) $\mathscr{F}$ is a framework for modeling document representations, queries, and their relationships, (d) $R(q_i,d_j)$ is a ranking function which associates a real number with a query $q_i \in Q$ and a document representation $d_j \in D$. Such ranking defines an ordering among the documents with regard to the query q_i" (p. 23). A ling-reg approach suits $\mathscr{F}$; to support novel concept identification, interactive visualizations are better suited than the usual text-based lists (Benoît, 2019).

4.3 | Building on finer-grain query reformation and revealing user linguistic behaviors

As a transition from the usual model to focus on potential set membership of terms, we see at least one theoretical framework that supports this ling-reg notion. Zermelo–Fraenkel set theory applies here because writing styles, needs of coherence in any written document, and semantic tokens span genres and domain-specific collections. Consequently applying ling-reg as an entry point for creating and reformatting queries allows us to adjust the weighting schemes for finer control over retrieval sets and presentation to the end-user. Moreover, we can imagine a log trace of search terms, longitudinally mapping across different ling-reg of the collections, exposing more about query reformation, shifts in linguistic register while searching, and using the data for adjusting any term-weight feedback scheme. A search log might reveal the users' shifts in register. An individual shifts from performing one normative role to another one and with that an alternative register, appropriate for that role, even though searching for the same concept.

For example, “myocardial infarction” (MI) might be the bench-level register for health sciences while “heart attack” is the equivalent in a casual register. The phrase for MI is also a member of the set of academic register and technical register terms. In this approach the “entry point” may be to match a query to the semantic token (term) bench-level register. Terms in the other registers related to that bench-level have degrees of similarity between. There are no 0 terms because inclusion in the frame requires some rationale for inclusion.

In general, the subset of a set z obeying a formula $\varphi(x)$ with one free variable x is written $\{x \in z: \varphi(x)\}$. The axion of specification holds that this subset always exists because there is one axiom for each φ. Let φ be any formula with all free variables among x, z, $w_1,\ldots,w_n$ (y is not free in φ) then $\forall z \forall w_1 \forall w_2 \ldots \forall w_n \exists y \forall x[x \in y \Leftrightarrow (x \in z \wedge \varphi)]$. y might not be free because of its rarity or by being tightly bound to a particular topic domain. Each of the linguistic registers $\mathcal{F}$ form an axiom; for instance the union over {{“myocardial infarction”, “heart attack”}, {“heart attack”, “heart pangs”}} is {{“myocardial infarction”, “heart attack”, “heart pangs”}. For any set of sets $\mathcal{F}$ there is a set A containing every member that is a member of some member of $\mathcal{F}$: $\forall \mathcal{F}\ \exists A\ \forall Y\ \forall x[(x \in Y \wedge Y \in \mathcal{F}) \Rightarrow x \in A\}$. Given a query term t to be matched in the document collection representation, a retrieval set of sets across a universe of discourse φ matching the query, is $\cup\mathcal{F}: = \{x \in A: \exists Y(x \in Y \wedge Y \in \mathcal{F})\}$.

As noted above, terms for w may be supplied by each of the φ so pairing is not limited to $\{x, y\}$ but the set of $\{w_1, \ldots, w_n\}$ where n is the total number of registers with terms associated with the bench-register.

There does not seem to be any thesaurus that has already classified terms according to ISO12620. Consequently, parsing a document collection requires the preliminary step, creating a preferred term list to identify the bench-level register for the collection. In some domains we imagine the bench-level and the technical or academic registers will be the same and easy to identify. Applying MeSH or other controlled medical vocabulary could identify during parsing the bench-level as well as a set of related terms, yet undistinguished by register.

Similar to some cross-language retrieval techniques, a “dictionary look-up” approach could be applied where each token or phraseological units or set phrases are identified while parsing documents and stored in an appropriate register.

Ironic, slang, taboo, vulgar, and facetious registers may be difficult to disambiguate in collection preparation, recognizing the facetious use of the line would be especially challenging without other evidence from the document context and/or metadata. The ISO standard includes options for normative references. Here metadata about the source document could be usefully integrated in a retrieval system during collection preparation particularly in the display of the results. End-users could get further help understanding the set by contextualizing the references with normative data. For instance, in the sample interface below an end-user could get details-on-demand, exposing the normative references.

5 | PROJECT DESCRIPTION: TESTING THE THEORY

The project explores whether the theory and ISO linguistic register models (a detailed above) may be a fruitful avenue of information retrieval research.

Corpora are usually semantic-level collection representations of tokens or terms (Azabonyad, Shakery, & Faili, 2012; Chen & Nie, 2000; Fung, 1998; Mohammadi, 2010; Rogati & Yang, 2004; Utsuro, Horiuchi, Hino, Hamamoto, & Nakayama, 2003) cleaned of stop words and a normalized numeric equivalent created based on a theory or empirical-based work and finally a weighting scheme for retrieval and relevancy ranking.

Because there is no digital collection of English terms already divided into registers, as described in ISO-12620:2009, that could be immediately applied in full, this project follows the standard but identifies only some of the registers in the test collection. In building the collection terms were compared to other sources, such as MeSH, to identify “international scientific”, “academic” terms; medical and computing resources for "technical", and TESOL resources for neutral terms. Then terms were associated to three fields of linguistic performance (research, practitioner, public).

The test collection of documents are revolve around a single topic (covid-19/coronavirus). While we can argue a professional might choose a domain-specific collection but we investigate an unbounded collection of semantic tokens. The project's small collection (n = 120 documents) was identified register by the document creators' own labeling. Although the ISO standard identifies linguistic units and then the 19 registers that indicate the "relative level of language individually assigned to a lexeme or term or to a text type [A.2.3.3]" (ISO, 2017), here several registers are tested: international scientific term/academic ("Group 1"), covid-19 health care practitioners ("Group 2), and general public resources ("Group 3"). Note that most research and practice journal access is usually restricted to an account subscription, yet to help inform the public, materials about covid-19 are offered gratis by the publishers.

The linguistic registers in the ISO standard define the classes of semantic tokens for the project. Although there are some 19 individual register markers, there is no computerized corpus of English language vocabulary terms that have already been tagged. Consequently only some of the registers were applied in order to test the concept. The graphics represent term distribution by linguistic performance (groups I, II and III); the radial graph shows the distribution of behaviors along eight of the ISO registers.

A "general" collection refers to a non-domain specific one. For example if searching PubMed through its portal https://pubmed.ncbi.nlm.nih.gov the search engine access a collection representation that is already highly biased to a few registers, arguably more technical, bench and in-house registers (for medical practitioners), and international scientific terms. One could argue that a scientist ought to know about and prefer Web of Science, PubMed, and the similar pre-coördinate systems, but this is arguing the ethics of professional behavior, not search engine design. Alternatively a post-coördinate search for similar concepts through a "public" search engine, such as Google with the aim of retrieving scholarly articles, may well have far more representatives of the other registers, too. Should these noisy additions be maintained when ranking or repressed? Moreover, we cannot assume that a URI domain ending (.gov, .mil, .com, .org, etc.) hosts *only* data resources suggested by the domain name. In this project no such assumptions are made; an opportunity came from both searchers' needs and database providers' responses to covid-19 that any person could search *gratis* otherwise for-cost journals, membership-only, and popular press websites, solely about the topic of covid-19. For example, the CDC hosted sites that were tagged specifically for researchers, practitioners, and the general public. The *Journal of Virology*, *New York Times*, *JAMA*, *NEJM*, *ACM Digital Library*, among others, added banners on the websites clearly stating that access to these sites was free for a limited time and for a limited subject. Perhaps long before a covid-10 public health report reached a peer-reviewed journal, these digital venues offered what people needed - the latest information. To underscore the point of this project, it is not to suggest failings in domain-specific resource providers, nor end-users' search behaviors, but to explore what has not been explored before: a novel way of using the semantics of publicly accessible resources to learn more about the potential utility of ling-reg retrieval, warranting further research, and demonstrating how applying ling-reg reveals more about the collections themselves, the use of the collections, and how the retrieval results can be made more useful to end-users of all stripes and any digital resource store, such as a library, research unit, or study of collections/user behaviors.

Parsing the documents from the sources in Table 4, using Rank NL's short stop list, created 24,116 individual terms for Academic, 31,414 for Practice and 44,222 for public sources.

1. Academic - specially biomedical, identified by the publisher for "medical practitioners": 24116 individual terms (or tokens)
2. Practice - health care, nursing, documents identified by the CDC for a "health care practitioner": 31414 individual tokens
3. Public - newspaper reports, popular-press health magazines, and documents identified by the CDC for a general audience: 44222 individual tokens.

It is obvious that some research and precise biomedical terms appear more frequently in the biomedical

TABLE 4 Sources for project corpus

Group 1: Int'l Scientific/Academic	Group 2: Practice	Group 3: Public
Journal of the American Medical Association	British Nursing Assoc.	*New York Times*
New England Journal of Medicine	American Nursing Assoc.	*Atlantic*
Journal of Virology	American Hospital Nurses Assoc.	CDC
Centers for Disease Control (CDC	CDC	

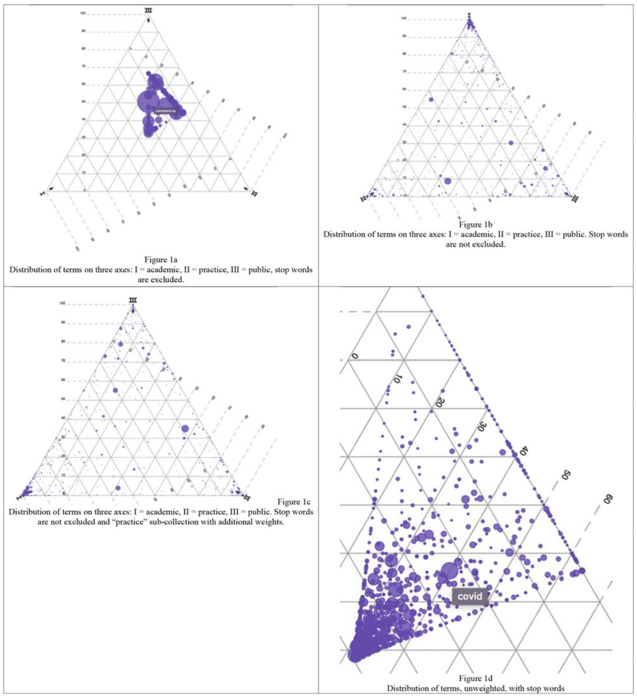

Figure 1a
Distribution of terms on three axes: I = academic, II = practice, III = public, stop words are excluded.

Figure 1b
Distribution of terms on three axes: I = academic, II = practice, III = public. Stop words are not excluded.

Figure 1c
Distribution of terms on three axes: I = academic, II = practice, III = public. Stop words are not excluded and "practice" sub-collection with additional weights.

Figure 1d
Distribution of terms, unweighted, with stop words

FIGURE 1 Plotting by register on 3 axes, one for each Group, with and without stop words. (a) Distribution of terms on three axes: I = academic, II = practice, III = public, stop words are excluded. (b) Distribution of terms on three axes: I = academic, II = practice, III = public. Stop words are not excluded. (c) Distribution of terms on three axes: I = academic, II = practice, III = public. Stop words are not excluded and "practice" sub-collection with additional weights. (d) Distribution of terms, unweighted, with stop words

collection. In a specialized collection, say focusing solely on medical research, the range of linguistic variance will be less than that in a popular press collection. But equally all collections are written in English and naturally share a large number of terms for the sake of composition.

The document collection represents six to eight of the ISO registers that we associated with three areas of linguistic performance (research/academic, practitioner, general public). Such a grouping provides a spectrum of term choice, information needs, and semantic comprehensibility.

The collection was parsed to extract the semantic tokens; one group with stop-words included, one without. The normalized frequencies of terms were plotted on a 2d 3-axis graph, each axis representing a performance area (I, II, and III). Terms shared by all groups cluster in the center of the plot; those shared mostly among only two of the groups appear closer to those vortices. Each term is represented by a circle; the more instances of a term, the larger the circle.

Figures 1a,b show terms plotted without and with stopwords. Without stopwords the terms cluster very closely. Allowing stopwords increases the noise of the data and terms are distributed with more fully across the plots.

In Figure 1c popular terms in one domain (II) were weighted to see if terms shared with another group (I) would be more distinguished but the results were minimal. Unweighted terms with stopwords and weighting "covid" paced covid more centrally in the plot with the resulting shift in academic/international scientific terms drifting to research, technical and academic shifting surprisingly towards public and less towards practitioner. It is already known that some broadly defined domains word choice will vary, measurably so by humanities, social sciences, and physical sciences (Heylighen & Dewaele, 1999; Hyland & Tse, 2007) but a live collection of documents reveal both shifting, shared linguistic registers and shifting themes.

Group 1 literature used highly-specialized terms from biochemistry and also terms that intersected frequently with Group III's (public) literature. The expected concepts (corvid-19, coronavirus) appear in both; interestingly unlike the intersection of Group I and II, the concepts in the I/III intersected area measured by normalized frequency describe public health, [other.] Group I and II intersect more on the themes of practice and hospital care.

Group II's literature was considerably more politicized than the other collections: "And because the large

majority of nurses are women, the nurses are also struggling to get necessary menstruation supplies, as only approved products are being allowed into the city right now—and of course, it's men doing the approving." The most common concepts in the intersections by Groups:

I - IIIemployees, vaccine, people, coronavirus, health, patients, transmissions, treatments, sars, epidemics, business

I - IIventilator, syndrome, chances, yield, ventilatory, patient, disinfectant, specimens, swap, secretion, precaution, cdc, face

II - IIIventilatory, clinicians, guidelines, computed, hospital, control

I-II-IIIwear, cleaning, respiratory, medical, patient, contact, covid, illness, control

Iunique medical terms (terms that are in MeSH), many computing terms, descriptions of research.

Interestingly, a surprising number of otherwise academic/research terms appeared in the public and practice literature; and a number of political terms appeared in the practice literature. For instance, hypernatremia, gammacoronavirus, coronaviridae appeared in the public literature, as well as the other groups, about half as frequently as in the practice-oriented group.

Reviewing the documents associated with these more frequently appearing terms suggest something about immediacy and documentation. A well-thought out,

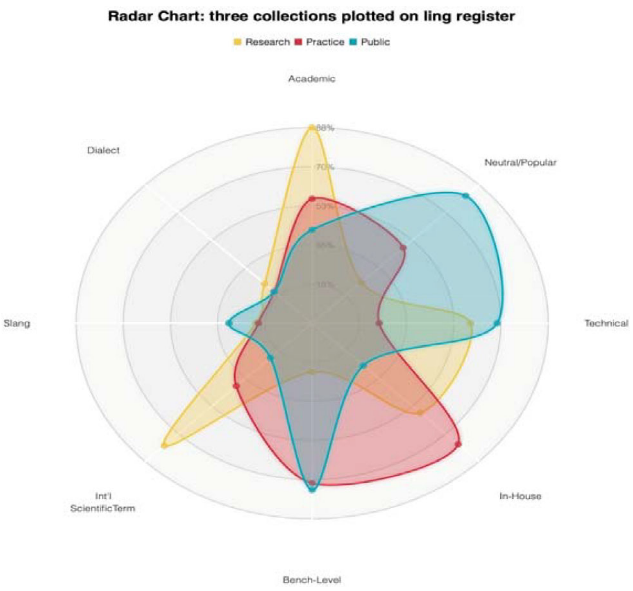

FIGURE 2 Radar chart of three collections represented by ISO Ling Reg

peer-reviewed article takes time to appear in a collection; a more lively production and delivery of articles suggest that otherwise domain-bound terms become unbounded and appear in linguistic register sets with unusual frequency. "Ventilator" for instance appeared along with medical terms more frequently in the public literature/ research than in the practitioner set. Perhaps the free access to medical and practitioner journals as well as popular press, all accessed through the post-coordinate "general" approach of an internet search engine, affected available resources and measurably impacted linguistic behavior (Figure 2).

6 | SUMMARY

The purpose of the project was to draw parallels between linguistic register as defined by ISO12620 and register-oriented work in other fields, to demonstrate commonalities between statistic-, genre- and literature analyses of document collections. Next the program shows traditional and more linguistically-sensitive modeling applies to ISO12620 for IR practice. A small information retrieval project suggests integration of LR-IR can reveal unexpected term clusters for new concepts across work fields (the Groups) and by extension that the choice of register in creating and using documents likely affects comprehensibility and so relevancy. Adding visualization techniques reveal to IR researchers stresses in distribution of concepts within a collection, aiding collection managers and end-users' retrieval.

7 | CONCLUSION

A medical doctor arguably prefers medical research; a practitioner something suited for his or her work, and the general public, lacking the levels of knowledge and technical vocabularies of the other groups, would comprehend non-technical collections. Comprehension + utility go to the heart of relevance ranking. So while current IR approaches are useful, they tend to eliminate terms that otherwise could reveal unanticipated concepts and events. The various registers, based on theories in the literature, offer to the interlocutor of the language A a point of access to the whole collection of recovery … an access point that he or she understands the best or is more appropriate for the given purpose.

We see, too, that work from other fields interested in language and comprehension parallel our interest in IR with genre, statistical trends, and source of production and use of document collections. Moreover linguistic register adapts easily into IR work as the comparison of quadruple and Zermelo–Fraenkel theories imply, and, in fact, can be used to leverage the differences of register to understand query reformation and end-user linguistic preferences.

In this project, the size of the general collection is small but necessitated by the lack of existing corpora based on ISO12620. Group I, medical and research journals (the bulk of the academic/int'l scientific/ group), were provided gratis for the topic covid-19; Group II, practice, was unexpectedly sparse, the majority of the documents being professional-groups, advertisements, references back to a smaller body of documents, often pointing to blogs, brief research notes, and CDC directives. The public group, newspapers, health journals, CDC, were notably richer in concepts, and like the research collection, discussed daily life and business needs. In this case the collection was covid-19 medical-interest oriented, the theory and model tested can be applied beyond to other subject fields.

The results, however modest, suggest a number of implications for retrieval studies. From the perspective of retrieval models, it seems worth investing in creating digital corpora where terms are identified with a ISO ling reg, and tagged with a normative linguistic performance to establish the "neutral" register. From there "domains of use" might become identified more from the trends of actual use than by pre-determined categories.

With the finer level of semantic granularity created by applying ling reg where can be empirical evidence of linguistic behaviors, such as code-switching, as the users' comprehension of the subject and vocabularies evolve, ranging across registers.

Some IR models rely on sources of data for weighting schemes to match more closely how end-users are likely to express their "information need." Setting aside the mathematics of weighting and retrieval, ling reg could be a more responsive, more "just-in-time" updating of matching queries to collections, as the trend to show more technical/scientific terms in the "public" set did in this project.

One can imagine that domain-specific collections, say PubMed, could act as cross-language IR systems often do, with an intermediary mapping across controlled-vocabularies, to more rapidly responsive natural language queries of trending terms.

In an otherwise unbound collection of documents, ling reg mapping may disambiguate terms and phrases. Across literature ling reg creates more accurate retrieval sets. For instance, Shakespeare's line from *Troilus and Cressida* "The plague of Greece upon thee, they mongrel beef-witted lord" has nothing to do with plagues, Greece nor comments upon aristocracy but rather a facetious statement that without ISO ling tagging can be reduced to a set of unrelated semantic tokens. Students writing on the topic of satire in literature would locate more accurate, more relevant sets.

If an IR system supports greater control by linguistic behaviors over retrieval sets, then the data could, perhaps should, be expressed using interactive information visualization techniques. In this project, we apply two graphing forms - a 3-axis chart of performance groups and a radar chart of eight of the linguistic registers - we see terms clustering and being redistributed by term weight manipulation. The radar chart's display an overview of the linguistic trends at the level of a collection. It seems, then, that anyone studying a collection or managing a set of data for a user group could make decisions on a more solid footing of measurable evidence.

As information resources across topics, user-groups, and linguistic-levels - from highly scientific to trendily popular, covering the range of expression - are available through "general" internet search portals and to improve domain-specific collections' utility, we conclude linguistic register-oriented retrieval efforts are worth developing.

REFERENCES

Alonge, A., Calzolari, N., Vossen, P., Bloksma, L., Castellon, I., Marti, M. A., & Peters, W. (1998). The linguistic design of the EuroWordNet Database. *Computers and the Humanities*, *32* (2–3), 91–115.

Arazy, O., & Woo, C. (2007). Enhancing information retrieval through statistical natural language processing: a study of collocation indexing. *MIS Quarterly*, *31*(3), 525–546.

Azabonyad, H., Shakery, A., & Faili, H. (2012). Using learning to rank approach for parallel corpora. *ECAI*, *2012*, 79–84. https://doi.org/10.3233/978-1-61499-098-7-79

Baeza-Yates, R., & Ribierto-Neto, B. (2011). *Modern information retrieval* (2nd ed.). New York, NY: ACM Press.

Benoît, G. (2019). *Introduction to information visualization: from data to meaningful information*. New York, NY: Rowman & Littlefield.

Biber, D. (1995). *Dimensions of register variation: a cross-linguistic comparison*. Cambridge, UK: Cambridge University Press.

Biber, D., & Finegan, E. (1994). *Sociolinguistic perspectives on register*. Oxford, UK: Oxford University Press.

Brizuela, M., Andersen, E., & Stallings, L. (1999). Discourse markers as indicators of register. *Hispania*, *82*(1), 128–141.

Chen, J., & Nie, J.-Y. (2000). *Parallel web text mining for cross-language IR*. RAIO '00: Content-based multimedia information access (volume 1, pp. 62–77). Paris: ACM.

Chen, S., & Nie, J.-Y. (2000, 2015). *Routledge encyclopedia of translation technology*. New York, NY.

Chiu, R. K. (1972). Measuring register characteristics. *TESOL Quarterly*, *6*(2), 129–141.

David, M. & Gardner, D. (2013). *Academic vocabulary lists*. Retrieved from https://www.academicvocabulary.info

Ferrer-i-Cancho, R., & Elvevåg, B. (2010). Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS One*, *5* (3), e9411. https://doi.org/10.1371/journal.pone.0009411

Fung, P. (1998). Statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. Lecture Notes in Computer Science. Conf. of the Assoc for Machine Translation.

Giménez-Moreno, R., & Skorcynska, H. (2013). Corpus analysis and register variation: A field in need of update. *Procedia – Social and Behavioral Sciences*, *95*, 402–408.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman Publ.

Haspelmath, M. (2019). How comparative concepts and descriptive linguistic categories are different. In D. van Olmen, T. Mortelmans, & F. Brisard (Eds.), *Aspects of linguistic variation*. Amsterdam, The Netherlands: De Gruyter.

Heylighen, F., & Dewaele, J.-M. (1999). *Formality of language: definition, measurement and behavior determinants*. Brussels, Belgium: Free University of Brussels.

Hyland, K., & Tse, P. (2007, Jun.). Is there an "academic vocabulary?". *TESOL Quarterly*, *41*(2), 235–253.

International Standards Organization. (2017). *ISO*. Retrieved from https://www.iso.org/standard/37243.html

Kennedy, G. D. (1987). Expressing temporal frequency in academic English. *TESOL Quarterly*, *21*(1), 69–86.

Liu, D. (2011). The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *TESOL Quarterly*, *45*(4), 661–688.

McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., … Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources & Evaluation*, *46*, 701–719.

Mohammadi, E. (2010). Semantic indexing approach of a corpora based on ontology. *ISSRI*, *9*(2), 518–584.

Neumann, S. (2014). *Contrastive register variation*. Berlin, Germany: De Gruyter Mouton.

Oakes, M. P. (1998). *Statistics for corpus linguistics*. Edinburgh, UK: Edinburgh University Press.

Open Linguistics Project. 2017. Retrieved from https://linguistics.okfn.org.

Prieto Valesco, J. A. (2013). A corpus-based approach to the multimodal analysis of specialized knowledge. *Language resources and evaluation*, *47*(2), 399–423.

Rogati, M. & Yang, Y. (2004). Resource selection for domain-specific cross-lingual IR. In SIGIR '04: Proc 27th Annual Int'l ACM SIGIR (pp. 154–161). https://doi.org/10.1145/1008992.1009021

Speelman, D., Grondelaers, S., & Geeraerts, D. (2003). Profile-based linguistic uniformity as a generic method for comparing linguistic varieties. *Computers and the Humanities*, *37*(3), 317–337.

Uccelli, P., Galloway, E. P., Barr, C. D., Meneses, A., & Dobbs, C. L. (2015). Beyond vocabulary: exploring cross-disciplinary academic-language proficiency and its association with reading comprehension. *Reading Research Quarterly*, *50*(3), 337–356. https://doi.org/10.1002/rrq.104

Utsuro, T., Horiuchi T, Hino K, Hamamoto T, Nakayama T (2003). Effect of cross-language IR in bilinguial lexicon acquisition from comparable corpora. *10th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 355–362).

Witt, A., Heid, U., Sasaki, F., & Sérasset, G. (2009). Multilingual language resources and interoperability. *Language Resources and Evaluation*, *43*(1), 1–14.

Wynne, M. (2005). *Developing linguistic corpora: a guide to good practice. Arts and Humanities Data Service*. Cambridge, UK: University of Cambridge.

How to cite this article: Benoît G. Searching Covid-19 by linguistic register: Parallels and warrant for a new retrieval model. *Proc Assoc Inf Sci Technol*. 2020;57:e246. https://doi.org/10.1002/pra2.246