Orthography as a Fundamental Impediment to Online Information Retrieval

Terrence A. Brooks

Graduate School of Library and Information Science

University of Washington,

Seattle, Washington 98195-2930

tabrooks@u.washington.edu

Abstract

Orthography is the linguistic study of written language: elements of text such as letters, punctuation marks and spelling. Information retrieval systems operate in the orthographic realm matching some text strings (i.e., index entries) from documents with other text strings (i.e., query terms) from patrons. During the early history of information retrieval, it has been convenient to assume the rationality and uniformity of orthography in order to concentrate effort building information retrieval systems. Fundamental orthographic problems have persisted into modern information retrieval systems, however, where white-space normalization and the arbitrary treatment of punctuation have exacerbated the orthographic impediment to information retrieval.

Orthography as a Fundamental Impediment to Online Information Retrieval

Buckland distinguishes several types of information, one of which is "information as thing." The significance of this distinction is that "information retrieval systems can deal directly with information *only* in this sense" (Buckland, 1991, p. 352). Information things such as data and documents are themselves composed of other things: the elements of written language. Orthography is the branch of linguistics that studies the elements of written language such as letters, punctuation and spelling. This link between information retrieval (IR) and orthography leads to the thesis of this essay, a modification of Buckland's assertion: *information retrieval systems can deal directly with information only in an orthographic sense*.

It has been convenient to assume the rationality and uniformity of orthography during the formative years of IR, although the maladroit handling of text has been evident (Borgman, 1986, 1996). Written language is not simply recorded speech, but has been evolving for at least four hundred years as a cultural artifact with its own structural and representational features (Nunberg, 1990). New information technologies spur the evolution of written language. For example, Grabowski (1996, p. 87) reports a form of spelling used on the Internet that can only be understood if pronunciation is imagined (“4 2sday nite” instead of “for Tuesday night”). Such a novelty makes text unpredictable (e.g.: Does *Tuesday* appear in some records and *2sday* in others? What other form might it take?). Unpredictable text in a database is hard-to-find text, and hard-to-find text undermines retrieval. The average searcher has neither the training nor the temperament to struggle with unpredictable word forms, preferring "normal" text:

> I just hate it when people do 'keyword' indexing, as if normal text is written in keywords rather than . . . well, normal text. Did you ever wonder, as you wander from library catalog to library catalog, about all the different ways that keyword indexing has been implemented? (Coyle, 1993)

The argument of this essay is that written language is idiosyncratic, culturally determined and continually evolving: that is what makes it expressive and useful. However, these same characteristics make orthography a fundamental impediment to IR.

The Context of Critique

As a field report by a participant observer, this essay reflects my own online experiences. While a comprehensive review of online vendors would be ideal (Saffady lists more than twenty in his 1996 survey), I shall focus on the query languages of only three: Dialog and DataStar (The Knight-Ridder Corporation, Mountain View, CA) and EPIC (OCLC Online Computer Library Center, Dublin, OH).

Dialog is the world's largest multidisciplinary online information service (Saffady, 1996, p. 343), while DataStar is the largest European database vendor. Both offer access to more than 400 databases. EPIC is a medium-sized multidisciplinary online vendor (Saffady, 1996, p. 355) offering access to approximately 70 databases. EPIC features a command-oriented interface based on the NISO ANSI Z39.58 Common Command Language for Interactive Information Retrieval (American National Standards Institute, 1994).

One can search the ERIC (Educational Resources Information Center) database on all three of these vendors. The OLUC (Online Union Catalog) database is available on EPIC. Most of my examples come from these two databases. The ERIC database, 30

years old in 1996, is the world's largest source of educational information with approximately 800,000 documents and journal articles. It is dwarfed, however, by the OLUC that contained 31.1 million bibliographic records on June 30, 1995. The OLUC "is the most consulted database in academe" (Smith, 1996, p. 1).

The issue of the magnitude of the orthographic challenge to IR is addressed below. Setting the context for the calibration of this magnitude is the recognition that the focus of this essay is established, influential database vendors, and pre-eminent databases that are searched by thousands of people daily.

The basis for my critique is the pedestrian ambition of the average IR user: "What are you looking for?" "Just type it into the system." The purpose of IR systems is to help people find information; therefore, "selectivity" (Meadow, 1992, p. 2) is important. Selectivity implies that a searcher can formulate a query to pinpoint a specific bibliographic record. This requires one to reproduce, say, the title of the desired item. Ordinary people will use an everyday orthography; they do not possess a subset of written language just for database systems. They will use apostrophes in possessive forms, include periods in acronyms, hyphenate words and include commas in inverted phrases (Drabenstott and Vizine-Goetz, 1994, p. 126). They will not be able to anticipate arcane interpolations that catalogers might insert into a bibliographic record:

> Note that a minimum of interpolation is wanted because those searching the machine catalog cannot very often be expected to 'second-guess' the cataloger in this respect, i.e., users will normally formulate search keys that necessarily do not take interpolations into account (Cataloging Distribution Service, 1995, p. 7)

They will assume that they will not have to transform their request into a tricky neologism or employ special insider knowledge:

> Without further instruction, people assume catalogs are arranged like a telephone book or some other well-known filing sequence. . . . but the real difficulties lie in the fact that cards may be filed based on information that is not on the card. (Borgman, 1996, p. 498)

A critique of IR systems can be justified by illustrating the distance between the ordinary orthography of the IR searcher and the orthography necessary to use an IR system. The following sections of this essay examine how IR systems normalize text to produce "words" based on the presence of white-space, how punctuation is arbitrarily altered in those "words", and how far the result lies from the orthography of the average IR user.

Finding "Words" in Text

Although IR searchers are said to be “searching a database” or “searching for documents”, these metaphors obscure the reality of the more mundane task of matching query term to index term. In an IR system hosting unrestricted text, the task of matching one string of characters to another string of characters would be very difficult unless there was a normalizing algorithm that processed both the document text and the query text. Normalizing text before attempting a match helpfully shields searchers from issues such as the upper case/lower case distinction, which can be safely ignored for most searching. On the other hand, one profound consequence of normalizing text is the mechanical discovery of "words" in both query text and document text.

Normalization should ideally produce lexical words or lexemes. A lexeme is the smallest distinctive unit of a language, usually consisting of a single word but sometimes multiword phrases (Chalker & Weiner, 1994, p. 223; Crystal, 1992, p. 226). The multiword lexeme "brother in law" may appear in text as one word, *brotherinlaw*; three words, *brother in law*; or ambiguously as *brother-in-law*. A normalizing process that treats hyphens in an arbitrary fashion, or considers "in" as a word unworthy of indexing throws the searcher into uncertainty. *Brother in law* could produce either two or three index terms and *brother-in-law* may produce one, two or three. The ideal normalization process would be based on a natural language parser for unrestricted text,

> but, in fact, despite centuries of research we have no complete grammar or lexicon for any natural language. For English, surely one of the most studied languages, despite the existence of impressively large grammars and dictionaries, there is a seemingly endless supply of linguistic questions whose answer is not known. Given this lack of available linguistic description, it is not surprising that no adequate parser exists. It could hardly be otherwise. (Hindle, 1994, p. 105)

Any mechanical process that attempts to find words in text faces the fundamental problem of “what counts as a word or token in the indexing scheme?” (Fox, 1992, p. 103). This can be seen in the Brown Corpus, a one-million-word database of American English, assembled at Brown University in 1963-64 (Francis & Kucera, 1982) that defines “word” variously as:

- Graphic Words: strings of letters demarcated with spaces that may contain only hyphens and/or apostrophes,

- Compound Words: strings composed of other words, either appearing as a continuous string, in a hyphenated form or linked with spaces,
- Merged Words: two words merged to reflect a reduction in spoken language (i.e., he'll, 'tis, gonna), and
- Pseudo Words: normally independent words that have been linked together by a hyphen (i.e., "Charles MacArthur-Helen Hayes" contains the pseudo word "MacArthur-Helen").

One can easily imagine the impatience of pioneer IR designers who were eager to build IR systems, but stymied by the problem of parsing unrestricted text. Assuming the rationality and uniformity of orthography was a necessary convenience. The practical IR compromise is to chop text on white space thereby producing sets of non-blank characters, or orthographic words (Crystal, 1992, p. 419). Defining a word by white space is a rather recent innovation. In the manuscript culture, letters followed each other in a continuous stream without spaces. Only with the introduction of printing did it become conventional to use spaces to define words (McArthur, 1992, p. 1120). Given this arbitrary relationship between white space and lexemes, it is not surprising that white-space normalization produces many text fragments that do not look like "words." For example, consider the preceding McArthur reference. It produces the odd orthographic words *left parenthesis McArthur comma, 1992 comma,* p *period* and *1120 right parenthesis period*.

My guess is that few people would seriously regard such arbitrary formulations as words. However, this generalization is contingent on the present example; in fact, there are many non-blank strings with embedded punctuation widely regarded as words.

Chalker and Weiner give the example of "They're for my mum" (1994, p. 275) as containing either four or five words depending on one's taste. "Opinions vary as to whether certain compounds are in fact one word or two (e.g.: half way, half-way, halfway), and whether such forms as *don't* and *I'll* are single words or not." (Chalker & Weiner, 1994, p. 426).

Undeterred by these wordy conundrums, IR systems further process orthographic words. Salton and McGill (1983) describe a radical process where both prefixes and suffixes are removed before reduction to word-stem form. "This reduces a variety of different forms such as analysis, analyzing, analyzer, analyzed, and analyzing to a common word stem 'analy'." (p. 71) The three commercial systems under review take a more benign approach of merely transforming punctuation. Punctuation marks may be (a) elided (i.e., *Alzheimer's disease* is indexed as *Alzheimers* and *disease*), (b) retained (i.e., *at&t* is indexed as *at&t*), or (c) used to break orthographic words apart (i.e., *Charleston Hop'n John* is indexed as *Charleston*, *Hop*, *n*, and *John*). Some or all of these methods may be employed in the various fields of the same bibliographic record.

While vendor normalization processes are proprietary, the following general descriptions sketch how both document text and query text are processed:

- Many types of normalization are performed on a search key including removing or replacing non-alphabetic characters, converting into either upper- or lowercase depending on the standards established, and removing leading, trailing, and extra blanks. (Cooper, 1996, p. 333)
- Normalization, an automatic process that includes several operations: (1) reordering the words in subject access points into alphabetical order, (2) eliminating stopwords,

(3) disregarding capitalization, and (4) disregarding punctuation. (Drabenstott & Vizine-Goetz, 1994, p.138)

- Users don't need to know about normalization because the system normalizes their search terms using the same rules as it used to create the index entries. . . .The normalization used in RLG databases follows rules developed from analyzing the data and thinking about the ways searchers might enter search terms. Some of these rules were designed years ago; more are developed for new kinds of data as we encounter it. . . . We have used our own experience for our primary guide. (Stovel, 1995, 20 July)

Some Consequences of Normalization

White-space normalization and the idiosyncratic treatment of punctuation combine to present a serious challenge to the ordinary IR searcher. Given below are the descriptions found in vendor manuals of what is, in effect, their normalization process. Each is followed by some examples illustrating its application to real data. (Each example includes the name of the source file and the unique record identifiers.)

Dialog

To Dialog, a "word" is a series of alphabetic and/or numeric characters surrounded by either punctuation or blank spaces. . . . . When a word containing punctuation is word-indexed, the punctuation is ignored and the word is divided: X-RAY is word-indexed as the two entries X and RAY; 1,100 is word-indexed as the two entries 1 and 100 (p. 3-8) . . . . any term containing punctuation is broken apart, the punctuation marks are ignored, and each part of the term is entered in its own alphabetic position in the index. For example, the term ALZHEIMER'S is word

indexed as the two words ALZHEIMER and S. (p. 4-10) (Dialog Information Services, 1991)

- Apostrophe breaks word apart

  Alzheimer's Disease indexed as three words: *Alzheimer, s,* and *disease.*

  EJ521205 CG548331 File 1 ERIC

- Apostrophe and comma retained

  O'Toole, Richard indexed as: *O'Toole, Richard*

  EJ519452 CG548167 File 1 ERIC

- Hyphen breaks word apart

  old-fashioned indexed as two words: *old* and *fashioned*

  ED231606 RC014224 File 1 ERIC

- Double hyphen disappears

  Assurance--A Laboratory indexed as three words: *assurance, a,* and *laboratory.*

  EJ511224 CE528348 File 1ERIC

- Double hyphen splits one subject heading into two parts

  Cinderella (Legendary character) - - Drama indexed as two separate fields.

  Can be retrieved with *Cinderella (legendary character) (l) drama*

  2308164 LCCN: 87131633 File 426 LC MARC - Books

- Internal periods break word apart

  I.O.O.F. Cemetery indexed as five words: *I, O, O, F* and *Cemetery.*

  6624920 LCCN: 87180107 //r96 File 426 LC MARC - Books

- Internal question mark breaks word apart

  Truyen c?o indexed as three words: *truyen, c* and *o.*

6634453 LCCN: 96948519 File 426 LC MARC - Books

- Question mark disappears

  9 Easy (?) Steps indexed so that *easy* is adjacent to *steps*

  EJ138479 PS504716 File 1 ERIC

- Ampersand breaks word apart

  P&O Prepared Foods indexed as four words: *P, O, Prepared* and *Foods*.

  EJ521145 CE529111 File 1 ERIC

DataStar

Note that punctuation is not stored in the dictionary file and is not normally searchable. (Some special terms which include punctuation, such as chemical formulae, are searchable. . . . (p. 1.3) Finally, a note about words which in normal English usage are hyphenated, e.g. x-ray. As discussed earlier, punctuation is not stored in the dictionary file, so it should not be used in a search. Treat such terms as two words, and use the ADJ operator between them. (p. 3.9) (DataStar Guide: System Reference Manual)

- Apostrophe breaks word apart

  Alzheimer's Disease indexed as three words *Alzheimer, s*, and *disease*

  AN ED389846 D-S Update: 960530 File ERIC

- Apostrophe terminates surname, punctuation stripped, words reduced to single letters, hyphens inserted

  D'Alli, Richard, ed indexed as one word *d-a-r-e*

  AN ED239921 D-S Update: 920000 File ERIC

- Hyphen breaks word apart

CD-ROM indexed as two words *CD* and *ROM*

AN ED394234 D-S Update: 961009 File ERIC

- Hyphen binds words together

  Decoding (Reading) indexed as one word *decoding-reading*

  AN EJ534694 D-S Update: 970513 File ERIC

- Various punctuation marks ignored

  9 Easy (?!!) Steps indexed as three words *9, easy* and *steps.* Only *9* and *easy* are adjacent.

  AN EJ138479 D-S Update: 920000 File ERIC

- Ampersand breaks words apart

  P&O Prepared indexed as three words *P, O,* and *Prepared*

  AN EJ521145 D-S Update: 960923 File ERIC

EPIC

A keyword is a group of one or more letters or numbers surrounded by punctuation or spaces. (p. 100) In general, punctuation such as , . ; ? and symbols such as $ * % are not indexed. Exclude them when you enter a search. Enter a space in place of a / (slash). For example, to search for *brother/sister*, enter find brother sister. In keyword indexes, & and - (hyphen) within a keyword are indexed and should be entered. However, at the beginning or end of a keyword, they are considered as spaces and should be excluded. Search for terms containing & or - in different ways to retrieve records that may vary in the use of & and -. For example, to search *multi-family dwelling*, enter find multi-family dwelling and find multifamily dwelling. Do not enter diacritics and characters such

as accent marks, apostrophes, and other marks used within words. (EPIC User's Guide, 1991a, p. 103)

- Apostrophe elided, slash breaks word apart

  Alzheimer's/psychiatric disorders indexed as three words *Alzheimers, psychiatric* and *disorders*.

  AN: 35851016 File: 23 OLUC

- Hyphens retained, comma removed

  Drama in English, 1945- - Texts indexed as a phrase

  su=*drama in English 1945- - texts.*

  AN: 35824159 File: 23 OLUC

- Hyphens removed, comma removed

  Drama in English, 1945- - Texts indexed as four words: *drama, English, 1945, texts.*

  AN: 35824159 File: 23 OLUC

- Hyphen binds words together

  medicine, and psychology-are indexed as two words *medicine* and *psychology-are*

  NO: EJ102302 File: 1 ERIC

- Hyphens replaced with spaces

  A-- B-- sea indexed as phrase *ti= A space space B space space sea*

  AN: 32923296 File: 23 OLUC

- Hyphens removed; *a* treated as initial article

  A-- B-- sea indexed as two words *B* and *sea*

  AN: 32923296 File: 23 OLUC

- Three hyphens retained

  (R-(-)-deprenyl) indexed as *r---deprenyl*

  AN 17201943 File: 23 OLUC

- Apostrophe elided, ampersand retained

  Charles Harper's Birds & words indexed as phrase as ti=*Charles Harpers Birds & words*.

  AN: 1217464 File: 23 OLUC

- Apostrophe elided, ampersand removed

  Charles Harper's Birds & words indexed as four words *Charles, Harpers, birds* and *words*

  AN: 1217464 File: 23 OLUC

I claim that these examples illustrate the extraordinary ad hoc, arbitrary treatment of language in three influential database vendor systems. These variations in the use of punctuation are not found in the orthography of ordinary searchers. Even Grudin (1989) who argues that inconsistency may have a place in user interfaces would deprecate systems that defy users to build a mental model of, say, how to use a hyphen.

The degree of shock or ennui provoked by these examples is a factor of habituation: long-time users of all these systems tend to regard such idiosyncratic treatment of punctuation as normal--like background static on a radio that becomes less noticeable over time. It is not my intent to bully up some vendors. Dialog, DataStar and EPIC are useful and successful systems that satisfy thousands of searchers daily. While their normalization processes may introduce some static into IR, the real source of the problem lies in the nature of text itself. As writers seeks novel expression, as corporate

names and product names seek to differentiate themselves, etc., the nature of text is revealed to be unpredictable, exceptional, in short . . .creative.

Language as a Cultural Artifact

What you are reading right now is a highly stylized art form. Consider how this text is formatted, the presence of centered headings, how a series of words ends with a graphic element called a period. The textually conservative use style manuals such as the Publication Manual of the American Psychological Association to assert language norms. Normative textual behavior can be analyzed just as one can analyze the iconography of Picasso. For example, Dillon (1982) studies "titular colonicity" - the presence of colons in the titles of scholarly articles. On the other hand, the textually avant-garde struggle against the forms, shapes and the spelling of words. Melville Dewey changed his name to Melvil Dui (Wiegand, 1996, p 63), internet users consult dictionaries of :-) to express emotions in their electronic mail ("Smiley Face Dictionary"), and *Toys r Us* uses a backward letter *r* to indicate a child's writing:

> There are an interesting number of cases where we would have to accept that individual letters, and the way they are presented in typography or handwriting, do permit some degree of semantic or psychological interpretation, analogous to that which is found in sound symbolism, though the element of subjectivity makes it difficult to arrive at uncontroversial explanations. (Crystal, 1995, p. 268)

Verbal artists use the keyboard as their palette. Distinction is sought through font variation: *ConneXions, InformationWEEK* and *net*, or by combining letters, numbers and punctuation: *.exe, RElease 1.0, Soft*letter, T.H.E. Journal, I.T.1 Magazine*. These latter

risk malformation by any normalization process that breaks words apart based on punctuation. Sometimes font and spelling changes become one as in this advertisement: “GRAB THIS VNIQUE BVSINESS OPPORTVNITY” (Harris, 1986, p. 107). How far from an ordinary orthography is the substitution of *v*'s for *u*'s? Textual creativity is limited only by human imagination. Here is a short list:

- *Grant$ for women and girls, 1993/1994* can be retrieved with the query term *grants*, but not *grant$* (OLUC an 31483793)
- *;Login:* can be retrieved by ignoring the leading semi colon and trailing colon (OLUC an 10959450)
- *** must be retrieved with *asterisk asterisk asterisk* (OLUC an 29357394)
- *?* must be retrieved with *question mark* (OLUC an 28740285)
- *(!) yeah: cover and poems* must be retrieved with *exclamation mark* (OLUC an 3459474)
- *Output options: The .WHERE and .WHY of FreeStyle* can be retrieved with either *where* or *why*, but not *.where* or *.why* (Dialog File 148 aa=14928374) *.Where* and *.why* are Lexis/Nexis commands. These commands were created by prefixing periods thereby distinguishing them from "normal" words. Ironically, in this title they are being used as normal words. That is, text that was specifically designed not to be like normal text is being used here as normal text.

Catalogers can modify orthrography in creating bibliographic records. They use rule #2.14 E1 in the *Anglo-American Cataloguing Rules* (Gorman & Winkler, 1988) to simulate gothic capitals. Thus, *The Chronicle Historie of Perkin VVarbeck* (OLUC an 32529424) is written with two capital letter *v*’s for one letter *w.* How many searchers

would think to substitute two *v*'s for a *w*? The *ISBD (M)* (International Federation of Library Associations and Institutions, 1978, p. 12) encourages the introduction of corrective text, giving the example *The world in [d] anger*. None of the systems described above would compose this correction into the word "danger", but all would produce the text fragment *d*.

There is no end to the illustrations of the creative, arbitrary quality of text. The following sections, however, details several challenging aspects of text that have specific implications for IR systems.

The Absence of Text

> One is tempted to compare the introduction of the space as a word boundary to the invention of zero in mathematics; but the parallel is superficial. The space is not one additional alphabetic character, but the absence of any character." (Harris, 1986, p. 113)

The space is an invisible character, but still a character. It epitomizes a fundamental orthographic challenge to IR systems: The perception of text by a human reader and the parsing of text by a computer are different processes. A human reader reads this sentence, for example, as sets of non-blank letters separated by spaces. But *space* a *space* computer *space* parses *space* this *space* sentence *space* as *space* a *space* continuous *space* stream *space* of *space* characters *space* with *space* the *space* space *space* character *space* between *space* them. This difference in perception poses a difficult question for the average IR user: What happens if I enter two words? (i.e., how are multiword arguments interpreted?) because to enter two words I have put at least one space between them. What is the function of spaces? or, Does the absence of

text have a functionality? or, Is an invisible character playing an extremely important role in the interpretation of my query argument? Here are some examples of the various interpretations of the invisible character, the space, as it appears between two single words "school" and "accidents":

- [Dialog] *Select school accidents/de* This is a multiword ERIC descriptor. This space is a hard space, a real character. If one were accidently to type *school space space accidents/de,* you would have committed a syntax error resulting in zero retrieval.
- [Dialog] *Select school accidents/ti* Here the searcher is apparently following the foregoing example and desires to find these two words as a phrase in a title. Unfortunately this is not the way to indicate the proximity of these two words in a title. It is a syntax error that results in zero retrieval.
- [DataStar] *School accidents.ti.* Here the user is also seeking a title phrase. But this space is interpreted as a Boolean OR operator so this query is equivalent to *school or accidents.ti. School* is searched throughout the record (which includes the title field) but *accidents* is searched only in the title field. Consequently, many of the records captured by this query will not have *accidents* in their titles because of the effect of the implicit Boolean operator.
- [EPIC] *Find ti school accidents* Here the space acts as an adjacency operator. EPIC will retrieve those records with the phrase *school accidents* in their titles.

The point of these examples is that the human searcher perceives the same words in all: *school space accidents.* It is only in the context of a specific IR system that the space character takes on different meanings.

Invisible characters will trouble searchers (Drabenstott & Vizine-Goetz, 1994, p. 126), but database searchers are forced to recognize and negotiate it. Consider the example of *brooks , glynnis* (OLUC an 31619383) which has been unfortunately indexed as *brooks space comma glynnis*. It files alphabetically in an unexpected place and probably hides from all but the luckiest searchers. Or, consider *williams, a 1731-1776* (OLUC an 35068619) which is really *williams comma space a space space 1731 hyphen 1776*. How many searchers would anticipate two hard spaces before the *1731*? Spying more than one contiguous invisible character would be impossibly difficult for most ordinary searchers.

Verbal Caste Systems

Not all words have equal status. Early IR systems labored under computer storage constraints so it was desirable to eliminate informationless filler words such as "a", "an", "the", "of" and so on. The Boolean operators "and", "or" and "not" commonly join the group of stop words. However, reducing the universe of potential text arguments is always dangerous. Searchers tend to use initial articles in title searches (Drabenstott & Vizine-Goetz, 1994, p. 127) and authors use articles and pronouns as titles (for example, Stephen King's *It*, and Joyce Carol Oates' *Them*). Furthermore, what is an informationless word in one language can be very useful in another language or context: thus "an" is an article in English but means "year" in French, "or" is a Boolean operator but is also the postal abbreviation for Oregon and the French word for "gold", etc.

Consider the common English article *a*. DataStar considers *a* an informationless word in French (DataStar Guide, n.d., p. 18.2), while EPIC bans it from English. This

makes searching for “vitamin a” difficult, as well as Andy Warhol’s novel *a* (OLUC an 442253). Consider searching for the *A. A. T. E. Guide to English Books, 1973* (EPIC ERIC no ED088079). Here the leading *a* is not an article but an abbreviation. The searcher who recognizes that punctuation is removed before indexing is now in a quandary (that is, *A period space A period space T period space E period space* has become *A space A space T space E space*). Is the leading *a space* now equivalent to the article *a*, which would also be *a space*? In EPIC the answer depends on whether one is keyword searching (yes!) or phrase searching (no!). These subtleties are at a level of complexity beyond the capabilities of most ordinary searchers.

Searching difficulty increases in direct proportion to the number and contingent application of stopwords. DataStar uses two stopword lists for the ERIC database. Fourteen stopwords are applied throughout the record, 57 others are applied to the title and abstract fields. The practical consequence of two stop word lists for the same database is that adjacency becomes field dependent. Consider the phrase “attitudes toward diversity” which appears in the record DataStar ERIC EJ525443 in both the title and identifier fields. Are “attitudes” and “diversity” adjacent? Yes, in the identifier field; no, in the title field. Imagine the confusion of the ordinary searcher.

<u>The Hyphen</u>

McIntosh (1990, p. 3) distinguishes two uses of the hyphen in the English language: joining two elements of a compound word (the link hyphen) and signaling that a word is being split at the end of a line of printing (the break hyphen). He lists (p. 30) many examples of nonstandard hyphenations that have appeared in English newspapers. What happens when these newspapers are digitized? The normalization

process breaks words on the hyphens resulting in odd text fragments. Text fragments that can be found in the Dialog files 622 (Financial Times Fulltext), 710 (Times/Sunday Times (London)), and 711 (Independent (London)) include “Europeans” broken into *Europe* and *ans*, “distinguishing” broken into *distingu* and *ishing*, “occurred” broken into *occ* and *urred*, “successful” broken into *succ* and *essful*, “positions” broken into *posit* and *ions*, “accuracy” broken into *acc* and *uracy*, “everyone” broken into *ever* and *yone*, “asked” broken into *ask* and *ed*, “important” broken into *imp* and *ortant*, “owners” broken into *ow* and *ners*, and “where” broken into *whe* and *re*. The point of these examples is their unpredictability. How many searchers would be omniscient enough to search for *successful* and, as well, *succ adjacency operator essful*?

The Apostrophe

The conventional use of the apostrophe has been to indicate the possessive form, and missing letters in a contraction but its application has never been exact (Little, 1986; Sklar, 1976). For example Little surveyed American signage and found widely varying uses of the apostrophe including some that appear to be nothing more than textual decoration:

> Today it [the apostrophe] frequently fails to appear where it is expected. “Dads Favorite Child” (child’s night shirt), President Reagans Address (ABC-TV), Chelsea Mans Shop (commercial sign), Violators will be towed at owners expense (No Parking sign) At the same time it is showing up in the strangest places: “Chez Lee’s” (restaurant sign) “Knoxville Welcome’s Big John Tate” (banner), “First 200 Mom’s get a free rose” (restaurant marquee), “Lar’ry 66 Service” (service station sign) (Little, 1986, p.16).

Room (1989, p. 23) urges dropping the apostrophe altogether from the English language. Some deapostrophised words would produce ambiguous homographs such as *hell, shell, wed, well* and *wont*, but these could be perhaps understood through context. American professional educators have already taken the bold step of expunging the apostrophe. Both the *Thesaurus of ERIC descriptors* (Houston, 1990) and the ERIC *Identifier Authority List* (Weller & Houston, 1992) are deapostrophized. They offer the descriptor "Childrens Theater" and the identifier " Americas Competitive Challenge." In a personal telephone call Jim Houston, lexicographer of the ERIC database, explained that this strategy saved the searcher from the difficulty of dealing with the apostrophe. The devaluing of punctuation echoes an earlier age when authors gave their work to compositors and printers to punctuate according to the printers' own standards (Bruthiaux, 1993, p. 35).

Punctuation and Semantic Effects

There are many instances when punctuation has a semantic effect (Meyer, 1987, p. 41). For example, the ERIC descriptors *Letters (Alphabet)* and *Letters (Correspondence)* are differentiated by their parenthetical qualifiers. Punctuation is crucial in disambiguating author and personal names. Surnames are not always obvious, consider a name such as *Richard Jardine Leon* (OLUC an 32604949). It is only by viewing the author field that I discover the surname is *Jardine Leon*. Furthermore, there are some words whose meaning is dependent on punctuation: "re-form" (to form again) from reform (to improve), "re-signing" a document from resigning an employment, "re-cover" a chair from recover a chair (McIntosh, 1990, p.

68). The following two passages (Hockett, 1987, p. 5) have the same words in the same order, but contrasting meaning due to punctuation.

Version one

That bright red rose – I see its thorn. I disregard the scent it gives off – that's nothing! I hate the scratches I got before; I fuss about the pain, too. Much I think of the beauty!

Version two

That bright red rose I see; its thorn I disregard. The scent it gives off, that's nothing I hate. The scratches I got; before I fuss about the pain too much, I think of the beauty.

Punctuation in these passages creates meaning. An example is the period at the end of the first sentence that clarifies that it is the scent that is disregarded (version one) or the thorn that is disregarded (version two). The sentence boundary is a fundamental element of orthography that serves to associate certain words meaningfully together. Both Dialog and EPIC, however, use adjacency operators that ignore sentence boundaries, thereby permitting "thorn" and "disregard" to be retrieved by adjacency operators. Most ordinary searchers would be surprised to find that words in different sentences could be connected by a proximity operation. Such a definition of adjacency destroys the integrity of the sentence, and most ordinary searchers respect sentences.

Indicators of Magnitude

Unfortunately, the multifaceted problem of orthography resists reduction to the single proof that would satisfactorily convince everyone of its magnitude. Many readers will be convinced that orthography is a fundamental impediment to IR by the

numerous examples given above, but others will resist, remaining skeptical in the belief that the examples presented so far are unique or at least low frequency phenomena. The following section gives a few indicators of the magnitude of the orthographic impediment to IR.

Finding the Name of a Company

Dialog has, in effect, recognized the challenge of orthography by introducing the Company Name Finder (file 416), Journal Name Finder (file 414) and Product Name Finder (file 413) databases. The Company Name Finder database is a search aid database designed to locate company information across the several hundred Dialog databases.

Place Table 1 about here

Table 1 presents data downloaded in May 1997 for the company “L.A. Gear”. There are several high-frequency variants: about 1,500 records use “inc” in the name, but more than 2,000 don’t; 1,600 prefer “L. A.” but 1,900 do not. Librarians would recognize this as a species of name authority problem. Such control generally does not exist in commercial online vendors, or the Internet, so searchers must cope with the many ways that "L", "A", "Gear"and "Inc" can be written.

Finding the Name of a Person

Table 2 shows what happens to an author’s name that begins with an *apostrophe O*. In Dialog, Terry O’Banion is recorded without a comma in one document and is an editor in two documents. DataStar reduces his name to *O-B-T* and conflates him with other authors whose names reduce to these three letters. His name as an editor reduces to the four letters “O-B-T-E”.

Place Table 2 about here

While the frequency counts of one author's name may not be impressive, by extension this treatment applies to all authors with embedded punctuation in their names. Surely the Irish, just to name one group with many last names possessing punctuation, should be concerned about orthography. The reduction of a personal name to *o-b-t* would simply astonish most ordinary searchers. That's not the way they handle personal names.

Examples from the TREC conferences

The Text Retrieval Conferences (Harmon, 1995) represent large, ongoing experiments with IR systems. Orthographic problems have been finessed by "the virtually universal use of stemming and a stoplist." (Sparck Jones, 1995, p. 302) Nevertheless, it is possible to use sample words and phrases from the TREC questions and apply these to databases used by the TREC studies. In this way we can estimate the orthographic problems that would arise if stemming was not used.

Place Table 3 about here

Table 3 presents a sampling of words and phrases from several TREC questions (available at ftp-nlpir.nist.gov) searched in three of the databases used in the TREC studies: AP News (Dialog file 258: 1984 through 1997), The Federal Register (Dialog file 669: 1988 through 1997) and San Jose Mercury News (Dialog file 634: 1985 through 1997). For some of these terms, variants number in the hundreds; for others, variants make up a third of all occurrences. Although orthography may not be a focus of the TREC studies, orthographic issues would play a crucial role in the real-world searching of their questions.

The Dirty Database Test

Table 4 presents the Dirty Database Test (Beall, 1991) applied to Database 23, the Online Union Catalog with frequency comparisons from two dates: February 7, 1993 and May 16, 1997. In this four year period there has been an annual percentage increase in misspelling of these ten words of 38.8%. Even a modest extrapolation from these data would indicate a vast number of misspellings in a database of more than 30 million records.

Place Table 4 about here

Some Potential Solutions

This essay has described several ways that orthography impedes IR and given some indications of the dimensions of the problem. While it is merely a field report with the goal of alerting the IR community to the gravity of the problem of orthography, responsible scholarship compels at least an attempt at solution if for no other reason than to dodge the common criticism that pointing out a problem is easier than solving it. While it is difficult to suggest a solution to such a multifaceted problem, the following suggestions for further research are offered:

The 30-41 Solution

A 30-41 normalization would limit text to the numbers zero through nine and the letters *a* through *z*. The Unicode Standard (The Unicode Consortium, 1996) defines the character zero as U+0030 and the capital letter *A* as U+0041, hence the name for this solution. To implement this solution, unconstrained English text would be broken on white space, and the resulting orthographic words limited to zero to nine (Unicode values

U+0030 to U+0039) and the letters *A* through *Z* (Unicode values U+0041 through U+005A).

One implication of a 30-41 normalization is that difficult punctuation marks would be described in words as is the current practice, for example, *upside down exclamation point* (OLUC an 761582).

My cataloging friends have already derided this solution as hopelessly naïve, yet I persist in believing that there may be specific input situations where it could be effective. It also illustrates how impoverished is our understanding of the orthographic challenges to IR in a polyglot world. A 30-41 index presumes that the languages of the world are like English; unfortunately, "many assumptions about character rendering that hold true for English fail for other writing systems." (The Unicode Consortium, 1996, p. 2-2).

The Regular Expressions Solution

One can be confident that verbal artists will continue to express themselves with new and unexpected uses of the elements of orthography. An example would be giving additional meanings to the period, the asterisk and so on. As a stream of text is analyzed, a normalizer would not be able to differentiate these new uses for punctuation unless they were flagged. The regular expressions solution has already been employed in the book Mastering Regular Expressions: Powerful Techniques for Perl and Other Tools (Friedl, 1997) where special text formulations are flagged by corner brackets ⌜ and ⌟. In the following sentence, the corner brackets flag the special use of the period and asterisk, while indicating that the comma has no semantic meaning: *At first, ⌜é.*û⌟ matches the entire string.* (Friedl, 1997, p. 97) The essence of this suggestion is that any text element -

punctuation or stop word - that has been used semantically could be foregrounded by a similar device. Thus one could search for titles that are composed of three asterisks, etc.

Spoken Language Normalization

Is it possible to create a normalization algorithm that would parse text, not as it is written but as it is read? "Writing, when compared to speaking, can be seen as a more standardized system which must be acquired through special instruction." (Grabowski, 1996, p.75) When we read text many orthographic elements disappear.

> When you wish to refer in conversation to "a man's hat," you don't say *a man apostrophe s hat*, but when you write the phrase that is precisely what you convey: *a man's hat*. We say *mans*, but if we wrote that word that way we would be thought childish, or ignorant, or both" (Shaw, 1963, p.4)

What Shaw may be unintentionally illustrating is that spoken language, which would include reading text, is a form of simplification of language. If I could search for text as I would read it, I may be able to avoid many orthographic problems.

Conclusion

This essay is not a comprehensive survey of the impediment that orthography presents to IR, but merely a survey of several problems that I have noticed in using three commercial vendors. However, even within its limited scope, this survey has illustrated that naïve assumptions about the uniformity, consistency and predictability of language are no longer tenable. On the contrary, it appears that language itself is the fundamental problem of IR. Given that language is a cultural phenomenon, not a deductive or mechanical phenomenon, this should not surprise us. Nor should we be surprised at the obstinate lack of progress in IR as long as we continue to ignore orthography. Perhaps

the time has come to focus less on language as a thing and more on language as a cultural artifact, to move from the mechanical deconstruction of language and capture instead more semantics in our database systems. I recognize that this is not a novel suggestion, but it has become more urgent as our databases grow in size and become increasingly opaque with an orthographic fog.

One safe prediction is that orthographic challenges to IR will mount with the development of Internet repositories of unrestricted text in many languages. "As the non-English-speaking world comes online and preserves their full character sets in their online catalogs and other retrieval systems, matching filing order, keyboard input, and display, will become ever more complex." (Borgman, 1996, p. 499). Construction of large IR systems without more regard to fundamental language problems will inevitably result in the construction of electronic Towers of Babel. The end users of IR systems are still ordinary searchers who want to use an ordinary orthography. Without attending to the language problems of IR, we are casting more people into situations like the following:

> I ran into an interesting problem with diacritical marks and meta tags.
>
> We're beginning to use Dublin Core meta fields for our webpages, and last week I was doing some one-on-one instruction with someone with a Spanish name that has several acute-accented vowels. HTML does have ways to express these marks, of course, so we were just ducky when it came to the webpages, but we were stumped by the META fields. Should we use the escaped HTML characters to write the name? Should we write it in Americanized English? Should we repeat the field and do both? What about the abstract? (We're planning bilingual abstracts.) How are the meta

fields searched and displayed--raw, "webified," or (as I suspect)

both/either, depending? (Schneider, 1997, May 24)

Author Note

I gratefully acknowledge everyone who commented on earlier drafts: Allyson Carlyle, Jeffrey Cinnamond, David Crystal, Karen Drabenstott, Ed Glazier of RLG, Stephen Harter, Carol Hert, Charles Hildreth, Bill Jordan, Jack Kessler of FYI France Online Service, Joseph Kiegel, Edmund Mignon, Sam Oh, Kris Rayment, Katherine Sotol, Jan Spyridakis, and Robert Thomas. I am indebted to my students who brought many of these examples to my attention.

References

American National Standards Institute. (1994). ANSI/NISO Z39.58 - 1992 Common command language for online interactive information retrieval. Bethesda, MD: NISO Press.

Beall, J. (1991). The dirty database test. American Libraries, 22, 197.

Borgman, C. L. (1986). Why are online catalogs hard to use? Lessons learned from information-retrieval studies. Journal of the American Society for Information Science, 37, 387-400.

Borgman, C. L. (1996). Why are online catalogs still hard to use? Journal of the American Society for Information Science, 47, 493-503.

Bruthiaux, P. (1993). Knowing when to stop: Investigating the nature of punctuation. Language & Communication, 13, 27-43.

Buckland, M. K. (1991). Information as thing. Journal of the American Society for Information Science, 42, 351-360.

Cataloging Distribution Service, Library of Congress. (May 1995). Library of Congress rule interpretations, 1.0E, page 7

Chalker, S. & Weiner, E. (1994). The Oxford dictionary of English grammar. Oxford: Clarendon Press.

Cooper, M. (1996). Design of library automation systems: File structures, data structures and tools. New York, NY: Wiley.

Coyle, K. (1993, October 1). "WAIS software." PACS-L@UHUPVM1.BITNET.

Crystal, D. (1992). An encyclopedic dictionary of language and languages. Oxford, UK: Blackwell.

Crystal, D. (1995). The Cambridge Encyclopedia of the English language. Cambridge, UK: Cambridge University Press.

DataStar guide: System reference manual. (n.d.) Philadelphia, PA:

Dialog Information Services. (1991). Searching Dialog: The complete guide. Palo Alto, CA.

Dillon, J. T. (1982). In pursuit of the colon: A century of scholarly progress: 1880-1980. Journal of Higher Education, 53, 93-99.

Drabenstott, K. M. & Vizine-Goetz, D. (1994). Using subject headings for online retrieval: Theory, practice, and potential. San Diego, CA: Academic Press.

EPIC user guide. (1991a). Dublin, OH: OCLC Online Computer Library Center.

Fox, C. (1992). Lexical analysis and stoplists. In W. B. Frakes & R. Baeza-Yates (Eds.), Information retrieval: Data structures & algorithms (pp. 102-130). Englewood Cliffs, NJ: Prentice Hall.

Francis, W. N. & Kucera, H. (1982). Frequency analysis of English usage: Lexicon and grammar. Boston, MA: Houghton Mifflin.

Friedl, J. E. F. (1997). Mastering regular expressions: Powerful techniques for perl and other tools. Sebastopol, CA: O'Reilly.

Gorman, M. & Winkler, P. W. (Eds.). (1988). Anglo-American cataloguing rules, Second Edition, 1988 Revision. Chicago, IL: American Library Association.

Grabowski, J. (1996). Writing and speaking: Common grounds and differences toward a regulation theory of written language production. In C. M. Levy & S. Ransdell (Eds.), The Science of writing: Theories, methods, individuals differences, and applications. Mahwah, NJ: Lawrence Erlbaum Associates.

Grudin, J. (1989). The case against user interface consistency. Communications of the ACM, 32, 1164 - 1173.

Harmon, D. (1995). Overview of the Second Text Retrieval Conference (TREC-2). Information Processing & Management, 31, 271-289.

Harris, R. (1986). The origin of writing. London: Duckworth.

Hindle, D. (1994). A parser for text corpora. In B. T. S. Atkins, & A. Zampolli (Eds.), Computational approaches to the lexicon (pp. 103-151). Oxford, UK: Oxford University Press.

Hockett, C. F. (1987). Refurbishing our foundations: Elementary linguistics from an advanced point of view. Amsterdam: John Benjamins.

Houston, J. E. (Ed.) (1990). Thesaurus of ERIC descriptors. Phoenix, AZ: Oryz Press.

International Federation of Library Associations and Institutions. (1978). ISBD(M): International standard bibliographic description for monographic publications. London: IFLA International Office for UBC.

Little, G. D. (1986). The ambivalent apostrophe. English Today, 8, 15-17.

McArthur, T. (1992). The Oxford companion to the English language. Oxford: Oxford University Press.

McIntosh, R. (1990). Hyphenation. Bradford, UK: Computer Hyphenation Ltd.

Meadow, C. T. (1992). Text information retrieval systems. San Diego, CA: Academic Press.

Meyer, C. F. (1987). A linguistic study of American punctuation. New York, NY: Peter Lang.

Nunberg, G. (1990). The linguistics of punctuation. Menlo Park, CA: Center for the Study of Language and Information.

Room, A. (1989). Axing the apostrophe. English Today, 19, 21-23.

Saffady, W. (1996). The Availability and cost of online search services. Library technology reports, 32, 337-454.

Salton, G. & McGill, M. (1983). Introduction to modern information retrieval. New York, NY: McGraw-Hill.

Schneider, K. G. (kgs@bluehighways.com). (1997, May 24). Diacritical marks and meta tags. E-mail to Web4Lib (web4lib@library.berkeley.edu).

Shaw, H. (1963). Punctuate it right! New York, NY: Barnes & Noble.

Sklar, E. S. (1976). The Possessive apostrophe: The development and decline of a crooked mark. College English, 38, 175-183.

"Smiley Face Dictionary." http://users.nbn.net/%7Ekimle/smile.html (November 15, 1996).

Smith, K. W. (1996). OCLC - Moving toward the next stage of the electronic library. In Proceedings of the Fourteenth Annual Conference of Research Library Directors. Tomorrow's Access-Today's Decisions: Ensuring Access to Today's Electronic Resources (pp. 1-5). Dublin, OH: OCLC Online Computer Library Center.

Sparck Jones, K. (1995). Reflections on TREC. Information Processing & Management, 31, 291-314.

Stovel, L. (BL.MDS@RLG.Stanford.EDU). (1995, July 20). Term normalization. E-mail to Terrence A. Brooks (tabrooks@u.washington.edu).

The Unicode Consortium. (1996). The Unicode Standard, Version 2.0. Reading, MA: Addison-Wesley.

Weller, C. R. & Houston, J. E. (1992). ERIC identifier authority list. Phoeniz, AZ: Oryx Press.

Wiegand, W. A. (1996). Irrepressible reformer: A biography of Melvil Dewey. Chicago, IL: American Library Association.

Table 1

Variations in Representation of the Company Name "L A Gear"*

Name Variations	Dialog File Number	Record Count
L.A. GEAR	545	77
L.A. GEAR INC.	111	482
L.A. GEAR, INC.	545	752
L.A. GEAR, INC.	226	308
LA GEAR	18	331
LA GEAR	16	720
LA GEAR	570	506
L A GEAR	573	419

* Gathered from file 416 Dialog Company Name Finder May 5, 1997

Table 2

Variations in Representation of the Personal Name “Terry O’Banion”*

Dialog		DataStar	
Name Variations	Record Count	Name Variations	Record Count
au=O’BANION TERRY	1	O-B-T.au.	50
au=O’BANION, TERRY	39		
au=O’BANION, TERRY, ED.	2	O-B-T-E.au.	2

* Data collected May 9, 1997

Table 3

Frequencies of TREC term variants *

TREC Questions	Variants	Frequencies
101	ANTIMISSILE	335
	ANTI(W)MISSILE	3215
101	REENTRY(W)VEHICLE	28
	RE(W)ENTRY(W)VEHICLE	68
107	JAPANS	21
	JAPAN(W)S	42421
111	NONNUCLEAR	139
	NON(W)NUCLEAR	2688
111	NONPROLIFERATION	867
	NON(W)PROLIFERATION	2848
122	RDTE	12
	RDT(W)E	67

* Data collected May 10, 1997

Table 4

The Dirty Database Test applied to the Online Union Catalog (Database 23)

Misspelling	Frequency (February 1993)	Frequency (May 1997)	Percent Increase
Febuary	83	112	34.9%
Guatamala	58	80	37.9%
Misssion	0	3	300%
Goverment	112	274	144.6%
Fransisco	36	69	91.6%
Grammer	86	340	295.3%
Recieve	26	27	3.8%
Wensday	7	9	28.5%
Seperate	19	51	168.4%
Conditons	33	181	448.4%