H5

Practical Guide Series

10 Quick Tips from a Linguist

Optimizing Keyword Searches

Optimizing Keyword Searches: Quick Tips from a Linguist

Keyword search can be tricky. Whether you're negotiating with the other side or conducting your own keyword searches for litigation, you want the keywords (or "queries") you create (or agree to) to find what you intend them to find, without inadvertently including a bunch of irrelevant stuff that you'd otherwise have to weed through.

> ***Keyword*** *[kē'wərd]*
>
> *1. A word that serves as a key to a code or cipher.*
>
> *2. A significant or descriptive word.*
>
> *3. A word used as a reference point for finding other words or information.*

You've spoken with the subject matter experts who have provided insight into the vocabulary (read shorthand, code words, slang) that may have been used about the subject at hand. So creating keywords for the search shouldn't be too difficult, right? Well, read a few tips from a linguist and *you* be the judge. (And remember, at some point you may have to let the *judge* be the judge.)

1. Know your search engine.

- Just as a start, know that not all search engines are created equal. They may vary in the way they index information or parse your queries.
- Indexing is a process that determines which elements within a document are made available to a search engine, and there are choices that can be made about what gets indexed. If you get an unexpected result, is the problem in the index or the query? Familiarity with the tool will help you troubleshoot a query that didn't produce expected results so that you don't waste time barking up the wrong tree. Knowing what was or was not indexed may affect the way you create effective search strings.
- Numbers are a good example of elements that may not be indexed (or may not be indexed as expected) and are thus not searchable. If searching for numbers is critical to meet your goals (such as being able to search for patent numbers, currency amounts or dates), it is important to ensure that these elements are indexed the way you expect.
- "Stop words," common words that are ignored during indexing, are also important to know about. They may surprise you, comprising the very words you're searching for.

2. Pay attention to how you use common words as keywords.

- Consider whether any of the terms you're thinking of using as part of a search string are extremely common in everyday speech or in the particular data population you're searching. If they are, you'll likely retrieve a large number of documents that aren't related to what you really want to find.

- Instead of using common words as single keyword search terms, use a series of "general interest" topic keywords that can be combined with them to form more narrowly targeted two- or three-element search strings. If you can, use Boolean or proximity operators (see #4 and #5) to help target what you want. This strategy should reduce the yield of irrelevant material and result in a more focused result set.

Example: Say you're looking for documents related to an increase in price and attendance at a sports event. Since "increase" is a common word, it doesn't lend itself well as single element in a string. Instead, pair it with topic specific words to avoid over-capture as in (price* w/5 increas*) or (attendance* w/25 price* w/25 increas).

3. Beware of words with multiple meanings.

- Consider if any of the terms in the search string have multiple meanings—as so many words in English do (think “graft” or “check” for example.)
- Replace a multiple-meaning term with more specific terms or a combination of terms that will help home in on what you really need and avoid inadvertently targeting irrelevant material.

Example: Say you need to find all documents related to construction cranes in a defective product lawsuit. Depending on the nature of the data collection, searching for “crane” may bring you off-topic discussions about people with a similar last name or even discussions about the bird variety. Try replacing the term with an array of anchored terms such as **crane* w/25 equipment**, **crane* w/25 machine**, and **crane* w/25 construction**.

4. Make effective use of Boolean operators.

- Carefully consider how you use Boolean operators such as AND, OR and NOT. The AND operator, for example, is useful when you're looking for the coexistence of words in a document no matter where they're located or when targeting several different topics within a document. But if two keywords need to appear within the same section, paragraph or sentence in order to be on topic, using the AND operator alone to join them will bring in a lot of irrelevant results.
- The OR operator is useful to specify when one keyword can be substituted by another, for example "Chicago or New York." It should not be used to combine distinct concepts into a single search string, like "California or Denmark." The NOT operator is generally used to exclude results that you think you may inadvertently retrieve. Be careful, because there is a higher chance of missing documents of interest when you use NOT.

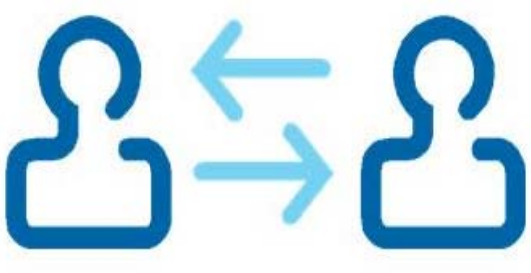
Boolean *[boo-lee-uhn]*

1. *Of or relating to a logical combinatorial system treating variables, such as propositions and computer logic elements, through the operators AND, OR, NOT, and XOR*

2. *Of or relating to a data type or variable in a programming language that can have one of two values, true or false.*

5. Make effective use of Proximity operators.

- When using a proximity operator, consider whether it is sufficiently large to capture the concepts represented in the string.
- If words can appear several sentences away from one another and still communicate the targeted concept, a larger operator may work. If the keywords must appear near each other to capture the concept then a large operator would be overly inclusive.



Proximity** [prok-sim'-i-tee] **Operator

***1.** nearness in space or time*
***2.** nearness or closeness in a series*

Proximity operators allow you to locate one word within a certain distance of another. The symbols generally used in this type of search are w(for within) and n (for near).

Example: Say you're looking for documents relating to charitable organizations. The string: **charit* w/250 (organization OR organizations)** may be over-inclusive, since the two words that far apart may not relate to one another. A smaller proximity operator would be a better choice in this case, as in **charit* w/25 (organization OR organizations).**

6. Make effective use of “wildcards.”

- First, make sure you’re using the wildcard operator in a way that is consistent with the search engine you’re using. Search engines vary in their application of wildcards.
- Consider whether wildcard operators are attached at the appropriate place in the stem of a word. If it’s not in the best place, it might target too many unrelated words or it might omit words you want to capture.

Wildcard (wild kärd)

1. an unknown or unpredictable factor

*2. a symbol (as ? or *) used in a keyword database search to represent the presence of zero, one, or more than one unspecified characters*

Example: Say you’re looking for documents related to forging documents. Using the term **Forg***, in addition to capturing variations of the word “forge”, will also capture variations of the common word “forget.” It will likely return many documents, a large number of which will be off-topic.

Example: Say you’re targeting everything related to refrigerators. **Refrigerate*** would seem to be a good choice. However, that will only target “refrigerate”, “refrigerates” and “refrigerated”. Placing the wildcard earlier in the stem, as in **refriger***, will capture other desired variations such as “refrigerating”, “refrigeration”, “refrigerator”, “refrigerators”, “refrigerant”, etc.

7. Don’t conflate multiple concepts in one string.

- Consider whether the search string conflates multiple distinct concepts. Trying to kill a few birds at once may not work well in a search string and you run into the problem of having to sort out what is related to which concept when you get the results.

- Rather, segregate distinct concepts into separate search strings so that you can evaluate the resulting sets independently, making it easier to identify other promising search strings and refine the flawed ones.

Example: Say you’re trying to find documents related to partnership agreements across entities in a particular market. The string: **(agree* or accept* or announc* or accord* or align* or coord* or coop*) w/10 (product* or market* or distribut* or pric* or cost*)**
targets different combinations of multiple concepts. Instead, break search strings of this type into multiple separate searches, such as **announc* w/10 product*, announc* w/10 market*, agree* w/10 product*, agree* w/10 market***. That way, you can better assess both the documents that are returned and the search string that identified them.

8. Consider “all and only” variations of the search elements.

- Words are notoriously pesky when it comes to variations—most have several as a result of tenses and noun and verb representations.
- Not paying close enough attention to the possible word variations in all of the elements in the string runs the risk of both under- and over-capture. Ask yourself: does the search string target all and *only* the desired variations of the elements represented in the string?

Example: Say you wanted to know about funds that were allocated for the purchase of a box that was used for an illegal shipment. The string:
allocate w/3 (volume* or product* or box* or ship*) would not capture all of the desired variations of the word “allocate,” since it does not target forms such as “allocated,” “allocates,” “allocation,” etc. Conversely, **box*** may target unwanted forms such as boxing or boxed when all you want is box or boxes, so you may retrieve more that you want.

9. Optimize synonyms or near-synonyms.

- Does the search string include a reasonably exhaustive set of synonyms or near-synonyms to capture the concept that is being targeted? It may help to treat synonyms or near-synonyms as single elements in search strings to enhance their coverage.

- Be as exhaustive as possible in order to avoid missing relevant material that is conceptually related to the material being targeting by the search string.

Example: Say you wanted to begin locating documents about the European Union's Commission's increasing regulation in a particular market area. The search string: **(Euro* Comm* or EU Comm* or EC) w/25 regulat*** does not contain an exhaustive set of synonyms for the "regulat*" element. Other forms might include requir*, order*, decree*, interven*, etc.

Synonym *syn'o·nym'*

1. *A word having the same or nearly the same meaning as another word or other words in a language.*

2. *A word or an expression that serves as a figurative or symbolic substitute for another.*

10. Try to capture common orthographic errors.

- Does the search string target common spelling errors or typos of keywords within the string? The spelling of names, for example, may vary. And human spelling errors comprise a significant source of variation within documents.

- Common misspellings should be incorporated into the search string in order to avoid missing relevant documents.

Examples:

(judgment OR judgement) w/15 trial* will capture a common spelling error.

(John OR Jon) w/5 White may capture a common misspelling of a first name.

Orthography (ȯr-ˈthä-grə-fē)

1a: the art of writing words with the proper letters according to standard usage

b: the representation of the sounds of a language by written or printed symbols

2: A part of language study that deals with letters and spelling.

If you're interested in improving your ability to find the documents you need in litigation, investigations, and regulatory response, H5's linguists and experienced search analysts can help.

Email us at info@h5.com to request a no-obligation consultation with a member of our team, or visit us at www.H5.com.