Linguistics is an often overlooked resource for indexing because it is little understood. However, linguistics is a powerful mechanism to establish indexing accuracy and support consumer understanding. The goal of indexing is to guide consumers in retrieving information. Understanding how to validate your indexing for job opportunities and projects will give you a tool that other indexers may not have. One powerful way to validate indexing, or return on investment (ROI), is through linguistics.

Whether you are a back-of-the-book indexer, a taxonomist, or another type of information specialist, we all must use logic to classify information. Linguistic analysis demonstrates that indexing is a judgment based on solid linguistic and cognitive science—essentially a judgment based on probability of accuracy. Linguistics can be harnessed to provide indexing accuracy probability measures—an ROI instrument I wish to help you craft. In order to use linguistics effectively—for stronger indexing as well as ROI—I will deconstruct the three main tenets of linguistics so that you can measure the accuracy of your indexing.

Three Linguistic Indexing Components

Linguistics does not have to be overly complicated, but it can become so if information professionals are unfamiliar with the basic tenets of linguistics. Here, I will focus on the components most relevant to indexing. Throwing most of the pleonasms out, the three universal components of linguistics, as can be seen in Figure 1, are form, meaning and context.

- **Information form** is the basic grammar of a language set—processing the construction of information.
- **Information meaning** is connecting the semantic interpretation of word association and concepts—word structure/concept structure.
- **Information context** is form and meaning sown into a collection of informative communication—how the consumer understands the information.

**Form, meaning, and context** are the building blocks of all information.

Each linguistic component builds upon the other. It is not necessary to deconstruct information to the root level; rather, the tenets of linguistic analysis should be customized to the content type you are working with. For instance, when indexing an entire book, it would be absurd to analyze every root and word—in which case pulling information from structure pieces like chapters may be more suitable—or if only a paper title is indexed a more detailed analysis may be used.

**Knowledge Bridge ROI**

Whatever content information professionals are working with, they seek to find a way of bridging the gap between information creators and the consumers of information. Essentially, linguistics...
should help communicate knowledge effectively and as universally as possible, creating a knowledge bridge between content and content consumers.

Creating a knowledge bridge starts with identifying two classes of information: common and uncommon knowledge. The common knowledge that both the creator and consumer share is based on common societal inferences, such as the Sun rises in the morning and the days of the week are Monday through Sunday. There are always exceptions that should be taken into account depending on the index topic. An example would be a book on the 20 day cycle Mayan calendar. The Mayan calendar negates the supposed common knowledge that the days of the week are Monday through Sunday. The Mayan calendar is an example of uncommon knowledge which the creator understands but the consumer may not. Accurately indexing uncommon knowledge so that consumers can understand where to find that information is essentially what creating a knowledge bridge seeks to do.

The understanding of each person is what I call linguistic fingerprints—all of the experiences and learning that each person accumulates to form their conception of ideas and their cognitive abilities. The way consumers interpret indexing, and how to measure the accuracy of indexing, is filtering all information through their unique linguistic fingerprint.

How can the strength of a knowledge bridge be quantified? Linguistic analysis can be used to create accuracy probability to substantiate your indexing precision. While nothing can replace the experience indexers draw upon to index information, having accuracy statistics from previous and current work will help validate the cost of your work through ROI.

ROI accuracy probability is calculated by:

1. Indexing the content as you normally would.
2. Next, taking all the knowledge gained in each linguistic tenet, make a reference list.
3. The reference list serves to identify the main points of the information. The main points will be the dominant references.
4. Once the main points are identified, having been already linguistically analyzed in the first step, the information can be validated for accuracy.
5. Comparing the linguistic analysis to a concentrated synopsis of the content, such as a table of contents or abstract—will generate the accuracy probability as shown in Figure 2 and described in Step 6.
6. As shown in Figure 2, the context gleaned from the linguistic analysis is that this paper is primarily about Testing of Fuel Filters. Validating the linguistic analysis, based on the paper abstract, shows that there are 14 forms of test, 5 forms of fuel, and 3 forms of filter. All three of these terms have a high probability of accuracy due to the linguistic analysis compared to the abstract. If a more detailed analysis were done the terms diesel and choking could also be analyzed.
7. Comparing your original indexing classification to the linguistic accuracy probability will show what the indexer thinks the thrust of the content is and what the semantics indicate the content is about. The threshold of human indexing should be 90 percent or higher. By identifying the linguistic accuracy probability and your own indexing accuracy probability overtime you can generate an average accuracy probability of your work.

Looking at the Figure 2 example many indexers may be able to properly index this information without linguistics—and I am not proposing that this should replace tried-and-true indexing—however, analyzing your indexing accuracy through linguistics will arm you with ROI validation of your work.

The following sections will help guide indexers through linguistically validating their classifications.

Semantic Underpinnings Of All Three Tenets

Semantics is another way to describe linguistic analysis of context. Ranging from textual to conceptual, linguistics are the tools to enable semantic interpretation. Semantics relates to the landscape created from all of the tenets of linguistics. In effect, and as shown in Figure 1, it is the study of the linguistic bits-and-pieces that form information.

Semantic form, word/term, and sentences combine to create semantic structures. Semantics can be analyzed from the context level down to the root level, as can be seen in Figure 1, depending on the indexing level of specificity. Consumers will translate indexing classification mentally and connect the mental picture to an associated meaning or concept. Semantically relating the information meaning and the indexing meaning is one example of building a bridge of knowledge.

The components of linguistics often can overlap in subtle ways. While outlining these components I have attempted to give the most straightforward application of the nomenclature. However, they all interconnect to form the foundation in which we communicate ideas, and therefore while defining one aspect, another may be referenced.

Information professionals create bridges by analyzing the information semantics, through linguistic deconstruction, and establishing natural language concepts that consumers will recognize. Using the semantic components gathered through linguistic analysis, the main points can be reverse-engineered to form the context of the content. Information professionals can build an accurate bridge between textual information and consumers through understanding the linguistic pieces of the semantic whole.
Language Tenet 1: Form

Linguistic form is, and the morphology of, the grammar that information is composed of. Based on indexing classification, morphology is used to deconstruct the composition of a term. It essentially means the study of analyzing words to find out what they mean, similar to how people study word parceling for standardized test vocabulary exams.

The root, or morpheme, is the main term that is static—meaning it is the most basic form of the word. The root is most often used in the creation of vocabulary lists, categories, or preferred terms unique to the information. Layers of meaning, called bound morphemes or affixes, can be added to the root. Some examples are:

- Prefixes: ante-, post-, un-
- Simulfixes: man  men, mouse  mice
- Suffixes: -able, -age, -less

Bound morphemes can be added to change the root—creating facets, or different shades, of a term. There are bound morphemes that are unique to particular languages as well, such as an inflix, which can be used in creating bilingual classifications. Figure 3 shows an example of morphology.

By adding the prefix and suffix the term ambiguity changes—becoming a negative (dis-) and past tense (-ed) term. Within text this also changes the context of the information. A root word is the most pure example of a term where words/terms have morphemes connected that change the definition of the root. The difference between roots and words/terms is important to note because the morphemes add new levels of knowledge that must be extracted for the linguistic analysis.

Language Tenet 2: Meaning

Psycholinguistics is a dramatic way of describing how written language is represented mentally—or the meaning of the information. Indexers use psycholinguistics to connect content information to the cognizance of consumers. Sometimes architecture to organize data is already in place—for instance a content management system or taxonomy—while in others it is not. In either case, it is up to indexers to choose the best organization system for classifying and relating information to consumers.

Through classification, information professionals would like to make all information easily accessible in a logical and structured manner. The fact is that information is messy, language is messy, and most of all human interpretation is messy. So how do information professionals establish order? In order to build up one must break down—which is the function of the meaning linguistic tenant. Breaking information down will enable indexers to parcel out the meaning of the linguistic forms. Meaning takes the roots and words/terms and analyzes how they are structured into the sentence; this is to identify the concepts of information.

Meaning is identified by taking a word/term and connecting it to related, or facet terms, with prepositions. Facets are not only variations of a term, but they can also be quantifiers of a term.

For example, in Figure 4, prepositions serve as relationship distinctions. In conjunction with facets and terms, prepositions create concepts. Figure 4 shows an example of concept relationships created through terms, facets, and preposition links.

For clarification, I would like to take a moment to explain the difference between a term and a facet. A term identifies a word root and a facet is usually a variation, or IS-A relationship, of the word. For instance, Canis lupus familiaris is a term while doggy, hound, and dog are variations. In indexing this would usually manifest as a see also or use for entry. There are many different types of data; therefore, the difference between the two depends on the detail of your classification method.

For instance, using Figure 4, buggy can be a facet of vehicle because buggy is a type of vehicle. Conversely, copper is not a type of study so it cannot be a facet of study, but it can be a facet of piping.

The meaning of information is semantically built into the sentence...
structure, or syntax, and can be gleaned by identifying the forms and concepts in the information. As can be seen in Figure 5, changing the structure of a sentence changes the meaning drastically.

Language Tenet 3: Context

Context is the tenet on which indexing is generally based on because it ties form and meaning together semantically. Context is then filtered through the cognizance of the consumer linguistic fingerprint. If the indexing is not accurate the consumer will possibly miss how to connect the classification with the sought for information.

In most disciplines, natural language processing (NLP) is used for machine learning only. However, it is a practice that stems from how humans process languages. NLP is used to bridge the gap between human language and machine cognition—or Human-Machine Interface (HMI). When indexing, linguistics and the natural lexicon of consumers, or NLP, are used to translate information into formats consumers can understand.

Information specialists serve as curators of the knowledge interface between the creators of content and the consumers of content. Breaking down the barriers between information and consumers rests on how well an indexer can achieve cohesion of natural language and formalized language in content. There are natural language words that are hard to quantify because they are infused with contextual value measures.

The study was conducted with copper piping.

The copper piping conducted in the study.

Figure 5 : Example of Structural Meaning Change

- Example: A book on the subject of “cooking well.” What value is the author placing on the word well? Does this refer to how the food tastes, how it was prepared, or the nutritional value?
- Example: “What is the best hotel in town?” How will a value-driven term be understood by a consumer? How will the information professional convey this sentiment through classification?

Again, linguistics can assist in determining how to link the language used in context with the natural language used by information consumers. NLP is a complex study better left to another discussion; nevertheless, by understanding word placement, context and meaning can be extracted from natural forms of language.

Linguistics assists in breaking down the barrier between what the information should be classified as and the probability of what consumers will actually use to find the information they are looking for. Again, it is helpful to have an accuracy probability for your indexing to show your classifications will be accurate enough to lead consumers to the information.

Context classification is most often organized in a uniform classification schema. Information professionals strive for uniformity of classification to achieve order. Taxonomy and other classification schemas enable continuity of information in a repeatable and dependable application that consumers can use to search for information.

Validating that classification on content adheres to classification schemas accurately is another form of indexing ROI.

ROI As A Measure Of Performance

Deconstructing language enables mapping of information to consumer’s linguistic fingerprint. The consumer’s linguistic fingerprint is just as unique as a biological fingerprint and the level of cognizance is dependent on the strength of the indexing knowledge bridge. Linguistic components create the foundation for knowledge bridges to be made as well as a way to validate if the classification is accurate enough for successful consumer information retrieval.

Information and human understanding is messy and will not always conform to orderly classification methods—nor should it have to. That is our job. Proving that your indexing has the highest probability of accuracy will bring a level of continuity to your indexing that would otherwise be difficult to muster. Having a high probability of accuracy, based in solid linguistic analysis, gives indexers an empirically founded ROI statement to support their work.

References

For more information on linguistics:
Glossary of linguistic terms: http://www-01.sil.org/linguistics/GlossaryOfLinguisticTerms/contents.htm
Lexicon of Linguistics: http://www2.let.uu.nl/uit-ots/lexicon/

We advocate, educate, and provide the central resource for indexing.

Contact Information:
American Society for Indexing
10200 West 44th Avenue, Suite 304,
Wheat Ridge, CO 80033
Phone: 303-463-2887   Fax: 303-422-8894