THEORY VERSUS PRACTICE IN FACET ANALYSIS

Rick Szostak
University of Alberta
UDC Seminar, London, September 2017

N.B.

- This presentation builds on some ideas in
- “The Simplest Approach to Subject Classification,” *Proceedings of the IFLA satellite conference*, Columbus OH, Aug., 2016.
- “Facet analysis without facet indicators” for *Dimensions of Knowledge: Facets for Knowledge Organization* (Richard Smiraglia and Hur-li Lee eds.).
- “Facet analysis using grammar,” *Proceedings of the NASKO conference*, Champaign IL, June, 2017.

Approaches to Facet Analysis

1) Developing a classification system in which a variety of potential subjects are “unpacked” in terms of all/most appropriate facets, and then hoping that most documents can be captured by one entry in this multi-faceted classification or perhaps by linking small numbers of entries.
2) Developing instead a classification system which contains elaborate schedules relevant to each facet, and then using a synthetic approach to develop subject classifications of any particular document. [Note: UDC employs this strategy with its auxiliary schedules, but also follows strategy 1) in developing subject-specific facets (Gnoli, 2011).]
3) Developing some even simpler classification system that allows each facet to be represented by different schedules of controlled vocabulary, and developing some rules of synthesis that guide classifiers to engage with key facets without actually having to consciously perform facet analysis.

Main argument

- The facets identified in Bliss2 and ILC can each be understood as either nouns, relators (mostly verbs), and adjectives/adverbs
- They can be distinguished by a combination of their place within a concept string that follows grammatical rules and their place within logical hierarchies of nouns.
- We can thus perform facet analysis without facet indicators.

Sentence structure

- In a string such as:

(Thing A)(influences)(Thing B)(influences)(Thing C)

Grammatical structure dictates that Thing B is the “patient” facet and Thing C the “product,” while the two influences would be “operations.” The influences would be drawn from schedules of relators.

We could also easily identify a “property” if some qualifier such as (big) were placed in front of any of our things.

Thing A might be a thing or type or part of a thing.

Hierarchy of Things

- We can distinguish “thing” from “kind of thing” within a logical hierarchy of things where subdivision generally occurs by “kind of”
- We can clearly indicate when we subdivide by “parts of.” [or perhaps have a separate table of “parts”]
- ……If we place all things in only one place in the hierarchical schedules

Hierarchy 2

- We can also make sure that “materials,” “space,” and “time” are clearly indicated in our schedules and have distinct notation. [But this would almost inevitably occur in a logical hierarchy of things.]
- “Agents” – another kind of thing – are a bit trickier, though it is certainly easy to capture types of human agency.

Note

- It is generally not necessary to identify which facets are captured by a particular term.
- The contention here is that we can achieve the goals of facet analysis by employing basic grammar in conjunction with logical schedules of things, relators, and properties

Indeed

- Classifiers should move directly from a sentence (perhaps more) in a document description to a subject string that captures the essence of the work. (They will capture key facets – things, operations, properties etc. – as they do).
- Users likewise can move directly from a query in the form of a (declarative) sentence to a search string. Note that we could easily develop search algorithms that valued the order of search terms.
- Since descriptions and queries are sentences, the subject heading that mediates between these should be sentence-like

Addresses Various Challenges

- There is debate regarding how many facets there are.
- There is debate regarding how to define these facets.
- There is debate regarding whether and how the order of facets should vary by subject
- There is debate regarding how many facets to include in particular subject classifications

Theories of Semantics

- Though there are a variety of theories of semantics they all suggest that humans understand grammar (either genetically or through learning) and that this understanding is important (along with understandings of the meanings attached to terms) in comprehending utterances.
- We in KO worry a lot about ambiguity of terms. We should appreciate that grammar reduces the ambiguity of utterances.

In Other Words

- Humans naturally think in sentences that follow a certain grammatical order
- A subject heading that is grammatical will thus be easier to comprehend and less ambiguous, especially for users lacking KO training

Are there problems?

- Though we all have a basic grammatical understanding, linguists struggle to identify the precise rules of grammar – and we may have struggled to learn just some of these in grade 5. In particular, linguists try to identify rules for long and complex sentences, and cases in which the standard order in which nouns, verbs, and adjective/adverbs appear is altered.
- We will want to standardize the word order in subject strings, and this will require occasional adjustments to the word order in document descriptions.
- [Note: We can prioritize word order in subject queries.]

Here's the list

- Translating interrogative, imperative, and exclamatory sentences or clauses into declarative format.
- Ignoring pronouns and most determiners.
- Using only the most specific form when nouns are repetitive.
- Translating verbs into the infinitive.
- Using combinations with auxiliary verbs to capture verb tenses.
- Translating phrasal verbs and idioms into synonyms (a task for a thesaurus).
- Placing simple adjectives before nouns, but post-adjectival phrases after.
- Using compound adjectival forms to capture gradation.
- Translating adjectival phrases with "that" (or similar words) into adjectival phrases using prepositions or infinitives.
- Ignoring or translating the rare adverb that does not appear after a verb or before an adjective or adverb.
- Using an extra set of parentheses if necessary (or some other notational device) to clarify whether a modifier is an adjective or adverb.
- Distinguishing adverbs from prepositions when the same word can be used for each.
- Ignoring the first component of a correlative conjunction.
- Addressing inverse verbs, ideally by preferring one form over its inverse.

Is this a lot?

- A classifier could surely learn these with practice
- A computer could also perform these adjustments …. perhaps with some human review

- And of course we will occasionally encounter sentences that defy these simple operations.

A Small Exercise

- To test the feasibility of this approach I looked at the descriptions of the first nine books on the Indigo Books (Canada) list of New and Hot Books, April 14, 2017.
- In each case it was straightforward to identify a defining sentence fragment.

First Book: *Option B*

- The subtitle of the book is this: “Facing adversity, building resilience, and finding joy” This captures the essence of the book in which two noted authors explore how to overcome adversity in life.
- It may be better to employ “overcome.” We would thus seek controlled vocabulary for (overcome)(adversity)(and)(build)(resilience)(and)(find)(joy).
- In BCC terminology this would be rendered (overcome)(grief)(and)(increase)(strength [under personality])(and)(achieve)(joy); it would thus not be too difficult to find controlled vocabulary. [BCC may want to introduce a more general term for adversity.]
- The BCC notation is →**ioGE9c+**↑**ID3+**→**ivGE8**

Worldcat provides for *Option B*

- grief; bereavement; and loss (psychology).
- These subject headings completely miss the message of resilience and joy in the book.
- It is notable that the subjects for *Option B* include grief or bereavement rather than the more general – and thus vague -- adversity.

Evicted: Poverty and Profit in the American City.

- This book by a sociologist describes eight poor families in Milwaukee facing eviction. (sociological)(description)(eight)(poor)(families)(facing)(eviction)(in)(Milwaukee).
- In BCC this would be (sociology)(describe)(eight)(poor)(families)(deciding about)(evict = (move)(someone)(from)(home or office))(in)(Milwaukee).
- The only challenge would involve converting “facing” into “decide about.”
- BCC notation is TF7b→iqXN8QC2SF→id(→gmI/NB1)

- Note: Richard and I have found that BCC notation is similar in length to UDC notation but encompasses more separate terms.

Worldcat again

- low-income housing; eviction; poverty; profit; and cities and towns.
- None of these subjects capture the fact that eight families are described in detail.
- Milwaukee already signals urban areas. Eviction would itself signal housing. Note that the book is not about dedicated low-income housing as provided by governments but about poor people with private landlords: the subject heading of low-income housing is thus less accurate than the combination of (poor) and (eviction). As for ‘profit,’ it is not clear that this is the sort or work that someone searching by that term would seek. Someone interested in the behavior of landlords is more likely to search for (eviction).

Comparing to PRECIS

- Purpose was to produce multiple index entries (of form Lead, Qualifier, Display) automatically:
- Canada

  Paper industries. Management

- Paper industries. Canada

  Management

- Management. Paper industries. Canada

- Had 26 “role indicators” plus several conventional marks for data entry

PRECIS and Grammar

- Did employ grammar within elements of the index (e.g. Paper industries)
- Had tried other organizing principles such as significance before deciding on grammar
- “The fact that general rules of this kind can be deduced and applied in practical indexing would seem to indicate that natural language is endowed with a greater measure of underlying logic than many classificationists would allow” (Austin 1974, 82).

PRECIS and Precision

- A separate indexer would assign LCC and DDC numbers after the PRECIS string had been assigned to a document.
- In 85% of cases the PRECIS string was more precise than the DDC tables.

Other PRECIS Insights

- Assume indexers start from sentence
- Translated into French
- Applied to books, films, etc.
- High indexer consistency found
- Developed a thesaurus without controlled vocabulary (advantage of focusing on terms rather than complex headings)

Revisiting the CRG

- “The searcher may be compared to an observer who has analysed a witnessed event into its correct elements – Tom, Bill, Alice, saw, and beat – but does not know how to combine the words so as to make clear that it was Alice who saw Tom beat Bill. The searcher is more handicapped than the observer, since the order of citation in a subject heading is dictated not only by meaning, but also by cataloguing conventions which may not be obvious to the uninitiated.”
- They suggest that the classificationist can either insist on one word order that the user needs to divine or must provide cross-references from other possible lead terms. The use of grammar as recommended here provides a third possible solution: It employs a word order with which the user is already intimately familiar.

More on the CRG

- Emphasized the need for logical hierarchy
- But urged the first type of approach to facet analysis above, despite their stress on synthesis
- Much easier with computers than card catalogs to pursue the second or third approaches

- Also stressed that users need both an approach to synthesis that makes sense, and recourse to logical hierarchies

Tentative Conclusions

- We get subject classifications that are more precise and capture key facets.
- They are easier to develop.
- Since they follow standard grammar they reduce ambiguity and enhance comprehension.
- These strings can be used for shelving (the classifier can identify a key term for this purpose).
- They facilitate search for related documents.
- Schedules are compact, flat, and logical.
- Fulfils CRG goals in manner better suited to computers