Rick Szostak. 2017. Facet Analysis Using Grammar.
NASKO, Vol. 6. pp. 248-256.

Rick Szostak, University of Alberta

Facet Analysis using Grammar

Abstract: Basic grammar can achieve most/all of the goals of facet analysis without requiring the use of facet indicators. Facet analysis is thus rendered far simpler for classificationist, classifier, and user. We compare facet analysis and grammar, and show how various facets can be represented grammatically. We then address potential challenges in employing grammar as subject classification. A detailed review of basic grammar supports the hypothesis that it is feasible to usefully employ grammatical construction in subject classification. A manageable – and programmable – set of adjustments is required as classifiers move fairly directly from sentences in a document (or object or idea) description to formulating a subject classification. The user likewise can move fairly quickly from a query to the identification of relevant works. A review of theories in linguistics indicates that a grammatical approach should reduce ambiguity while encouraging ease of use.

This paper applies the recommended approach to a small sample of recently published books. It finds that the approach is feasible and results in a more precise subject description than the subject headings assigned at present. It then explores PRECIS, an indexing system developed in the 1970s. Though our approach differs from PRECIS in many important ways, the experience of PRECIS supports our conclusions regarding both feasibility and precision.

Facet analysis is widely advocated in the Knowledge Organization literature but proves challenging to apply in practice. This paper suggests a novel approach to facet analysis which relies on basic grammar to structure subject classifications.

Basic grammar can achieve most/all of the goals of facet analysis without requiring the use of facet indicators. Facet analysis is thus rendered far simpler for classificationist, classifier, and user. We compare facet analysis and grammar, and show how various facets can be represented grammatically. We then address potential challenges in employing grammar as subject classification. A detailed review of basic grammar supports the hypothesis that it is feasible to usefully employ grammatical construction in subject classification. A review of theories in linguistics indicates that a grammatical approach should reduce ambiguity while encouraging ease of use. We perform a test with a small sample of books: This suggests both that a grammatical approach to subject classification is quite feasible and that it achieves greater precision that existing subject headings. A review of the experience of the PRECIS indexing system, which employed some grammatical constructions, also suggests that our approach is feasible and will enhance precision.

Facets and Grammar

Facet analysis has two key components:

- A stress on a synthetic or post-coordinated approach to classification, in which simple terms are combined to generate a complex subject heading. [This approach can be contrasted with the enumeration of complex subject headings, as in the Library of Congress or Dewey Decimal systems.]
- The idea that the terms synthesized will represent different "facets" of a subject. We can eschew the challenge of providing an intensional definition of the word "facet" – that is, attempting to identify the essence of the term in a couple of

> sentences – by instead providing *inter alia* in the next section an extensional definition: a list of what are considered to be at least the main facets that need to be addressed.

Sentences are synthetic constructs. We do not enumerate a body of complete sentences that humans may utter but rather allow each human in each utterance to combine terms as they see fit, though (generally) in accordance with some basic grammatical rules. This practice allows humans to communicate novel ideas by creating novel combinations of existing words. This is one of the hoped-for advantages of facet analysis: that new subjects can be signified by new combinations without requiring the classificationist to adjust their schedules.

Sentences, then, are a kind of facet analysis. And though we were once forced to identify nouns and verbs in elementary school, the glory of everyday speech and writing is that we do not have to indicate whether the words we emit are nouns or verbs. Everyday speech is faceted but requires no facet indicators. We perform a kind of facet analysis every time we speak or write a sentence but are not consciously aware of the process by which we construct grammatical utterances.

Facets in Grammar

Szostak (2017) described how each of the 13 facets recognized in Bliss2, and three more posited by the Integrative Levels Classification (www.iskoi.org/ilc), are represented grammatically. Within Bliss2, "Operation" is a verb in which one thing acts on another, and "process" is a verb describing changes within a particular thing. "Property" refers to an adjective or adverb. In a sentence fragment of the form (thing A)(affects)(thing B)(which affects)(thing C), the second and fourth terms are "operations," the third term is a "patient" (defined as a thing that is influenced and influences), and the fifth term is the "product."

Thing A might be a "thing," or "kind of thing" or "part of thing." These three facets can be distinguished if our grammatical constructions draw controlled vocabulary from a hierarchical classification in which "kinds of" are clearly distinguished from "parts of." Such a classification of things could also clearly indicate certain more specific facets: "material," "time," and "place." "Agents" are a type of thing that puposefully affect other things: These will mostly be humans or animals.

It is possible then for all facets to be captured by grammatical constructions that draw controlled vocabulary from separate schedules of nouns, verbs, and adjective/adeverbs, as long as the schedule of nouns follows a logical format that can clearly distinguish different types of thing.

Note though that classifiers and users will generally not need or wish to identify which facets are indicated by which elements of a grammatical subject descrition. Both will let basic grammar perform facet analysis for them – just as we do not bother in general to parse our statements into grammatical elements. Yet just as we can engage in grammatical

analysis if we find that our utterances are less clear than we might like we can examine the grammatical elements of a subject string if we find it somehow problematic..

Challenges in Employing Grammar in Subject Classification

Are grammatical rules straightforward enough that we can employ these to structure a subject classification? Our hope would be that a classifier can move fairly directly from a sentence in a document description to a grammatical subject classification.

Types of Words

There are eight types of words generally recognized in English. The four most prominent have been discussed above: nouns, verbs, adjectives, and adverbs. We can ignore pronouns in subject classification but will need to find a place for conjunctions and prepositions (which are kinds of relators) and some determiners (which act much like adjectives).

Typical Word Order

The most common word order in an English declarative sentence is noun/verb/noun: (dog)(bites)(mail carrier). More specifically, the order is (subject)(verb)(object). This order is important, as we saw above, for the identification of facets. We can also note that subject/verb/object is the standard word order in the vast majority of the world's languages.

When adjectives and adverbs are employed, adjectives precede nouns and adverbs usually follow verbs: (angry)(dog)(bites)(ferociously)(annoying)(mail carrier). Determiners play a role similar to adjectives: (four)(angry)(dogs)(bite)(ferociously)(annoying)(mail carrier). Conjunctions can link nouns or verbs or adjectives or adverbs: (dog)(and)(cat)(bite)(and)(kiss)(ferociously)(but)(charmingly)(annoying)(or)(petrified)(mail carrier). Prepositions generally link nouns: (dog)(from)(junkyard)(bites)(mail carrier).

There is thus a standard word order that we can employ in subject classification of declarative sentences.

Szostak (2017) addressed the basic grammatical rules governing nouns and noun phrases, verbs, adjectives and adjective phrases, adverbs, conjunctions, clauses, and some further complications (such as homophones and inverse verbs). From this analysis we can generate a manageable list of adjustments that need to be made to translate sententes from a document description into a standard subject classification format:

- 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.

Each of these fairly straightforward adjustments could be programmed into a computer – in much the same way that common spelling and grammatical constructions are programmed into my word processing program. As we all know, the advice in our word processing programs is not always perfect (and occasionally annoying). Human oversight is thus likely desirable.

Grammar and Ambiguity

Szostak (2011) drew lessons for Knowledge Organization from different concept theories. Szostak (2017) performed a similar exercise for the most common theories of semantics within the field of linguistics. The basic result of that survey was that all theories of semantics argue 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. That is, the ambiguity surrounding individual terms – a common subject of concern in Knowledge Organization – is reduced by placing these in the context of a sentence. We can thus reduce the ambiguity of subject headings considerably by using standard grammar in constructing these. Humans naturally think in terms of sentences and will better understand subject headings that follow the basic format of sentences or at least sentence fragments.

Would it Work?

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:

- *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**
- *The CANADALAND Guide to Canada.* This is described as "an outrageous exposé of Canada's secrets, scandals, and occasional awkward lapses in proper etiquette." We could use (outrageous)(secrets)(scandals)(and)(norm)(violations)(in)(Canada). In BCC this would be rendered ((outrageous)(and)(secret)(knowledge))(and)(large)(gossip)(and)(disobeyed)(everyday norms)(in)(Canada). Again it should prove fairly straightforward to find the appropriate controlled vocabulary. The BCC notation is ((QE6+QI3)T+QC7→**rtI(**QC3)+CV4)→rs>N1cca The term for disobey comes after the term for norms because "obey" is the inverse of "control."
- *The Underground Railroad*. This is a novel in which slaves seek to escape from the southern United States. The classifier would have to discern this key aspect of the book in order to render (novel)(slaves)(escape)(in)(southern)(United States). In BCC this would be (prose)(slave)(escape [moving from control])(in)(south)(United States). The notation is AN3 SO6→**gm/→rs>**N3sNicus
- *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). The only challenge would involve converting "facing" into "decide about." The BCC notation is TF7b→iqXN8QC2SF→id(→gmI/NB1)
- *Into the Water*. The plot of this mystery novel might be rendered as (two)(dead)(women)(found)(in)(river)(in)(small town)(with)(secrets). Note that the word order is important here: It is the town that has secrets rather than the women. The phrase can be directly translated into BCC, once "found" is replaced with "discovered" and placed at the front of the string: →**ip**XN2HMSG1>NT3r>N1g [Cutter numbers for Milwaukee]

6

- *Exit West*. (Romantic)(couple)(in)(civil war)(pass through)(door)(to)(alternative)(world). In BCC we would use "unusual" rather than "alternate." Civil war is the compound PI2k (→ gxPI1)
- *Fifteen Dogs*. (Dogs)(given)(human)(consciousness)(by)(gods). This can be directly translated into BCC.
- *Beauty and the Beast*. (Romance)(between)(beautiful)(woman)(and)(ugly)(man). This also can be directly translated into BCC.
- *How to be a Bawse*: The last word is defined as "a person who exudes confidence, hustles relentlessly and smiles genuinely" So the classifier might render (how)(to)(exude)(confidence)(and)(hustle)(and)(smile)(genuinely). In BCC we would capture "how to exude" with (achieve)(display). "Hustle" is a word with many meanings; it could be captured in BCC by (offering)(exchange).

For the non-fiction works there was usually a sentence in the book description (or a subtitle) that captured the essence of the book. For *Eviction*, it was necessary to summarize a longer description in one sentence – but this was straightforward. For the works of fiction, it was generally necessary to scan a paragraph or two of description, identifying key elements. Yet this still only took a matter of seconds. It would take a bit longer to translate the subject strings into controlled vocabulary – but not much longer if there was a thesaurus at hand.

The resulting subject strings give a very accurate sense of these books. A user that remembered the description but not the title of any of these works should be able to find it without difficulty. And a user wishing to transcend grief or understand the lives of the poor or read a novel about escaping slaves in the southern United States should likewise move quickly from a search query to a relevant work.

Translating the one-sentence strings garnered from book descriptions into the controlled vocabulary was also quite straightforward. A classifier familiar with the structure of BCC would be able to find most terms expeditiously. The BCC schedules are generally both flat and logically organized. In only a few cases was much thought required as to how to best render a term into BCC. If a detailed thesaurus were developed in conjunction with BCC translation would become even easier.

The subject strings above are certainly far more useful than those provided at present by OCLC. WorldCat provides the following subject headings for *Object B*: grief; bereavement; and loss (psychology). These subject headings completely miss the message of resilience and joy in the book. We can see that the WorldCat subject

entries are missing both "operations" and "products" in omitting "increase strength" and "achieve joy." It is not clear that a standard approach of facet analysis would necessarily capture the essence of a work with multiple operations and products. A classifier might see "strength" as a "patient," perhaps.

For *Eviction*, the subject headings are low-income housing, eviction, poverty, profit, and cities and towns. None of these subjects capture the fact that eight families are described in detail. WorldCat thus misses an "agent," an "operation," a "property," and a "product." Again we have a complex subject with multiple operations and products. The grammatical approach captures these and places them within one subject string. A classifier looking for different facets in turn might – like WorldCat – miss important elements of the book.

For *How to be a Bowse* the subjects are success in business; and anecdotes of the author. These are very vague subjects compared to the string above. Yet again we capture multiple facets missed in WorldCat: operation, product, property. No subjects were provided for the *Guide to Canada*. And of course works of fiction are at best captured by genre with no reference to plot beyond this.

Do the Worldcat subjects catch aspects of works missing in our subject strings? It is notable that the subjects for *Option B* include grief or bereavement rather thsn the more genersl – and thus vague -- adversity. For *Evicted* 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) – a term which would hopefully be linked to landlord in a thesaurus. As for *How to be a Bowse*, note that "hustle" or (offer)(exchange) implies success in business. We could easily add a term that captures genre to subject strings for works of fiction.

Comparing to PRECIS

In Szostak (forthcoming) I draw lessons from PRECIS, the Preserved Context Index System developed by Derek Austin and colleagues for the British National Bibliography in the 1970s, and adopted by several other institutions, including the National Film Board of Canada. Though its purpose was quite different from ours, it nevertheless found it useful to employ grammatical structures within its three key index terms. Notably, this was not the original intent, but it was found that it was the easiest organizing system as indexers moved from sentences in document descriptions to formulating PRECIS entries. And it was recognized that a grammatical approach guided indexers to seek "missing" facets: If they had an

action verb they looked for an object. Though PRECIS was abandoned by the BNB in the 1990s this was for reasons quite distinct from its use of grammar. The experience of PRECIS thus adds further support for both the feasibility and desirability of a grammatical approach to subject classification (Austin 1974, Richmond 1976, Dykstra 1989).

Conclusion

We can achieve synthetic subject classifications that combine nouns, verbs, and adjectives/adverbs in the order these generally appear in sentences. A manageable – and programmable – set of adjustments is required (and was identified above) as classifiers move fairly directly from sentences in a document (or object or idea) description to formulating a subject classification. The user likewise can move fairly quickly from a query to the identification of relevant works (especially if the OPAC contains a thesaural interface, but even if the user had to navigate flat and logical hierarchies). This grammatical approach combines ease-of-use with precision. Since we are all familiar with basic grammar, and since sentences are likely less ambiguous than isolated concepts, the recommended approach acts to reduce ambiguity in subject classification. Since the ideas that documents contain are expressed in sentences, the grammatical approach best captures the essence of works. And the grammatical approach is particularly well-suited to visualization techniques that can guide users to relevant and related works or objects or ideas: The visual interface needs only allow the user to change one term at a time in a subject string. Such links can be drawn across disciplinary or social boundaries. This paper provided evidence for the feasibility of the approach and the precision it achieves both from applying the approach to a small sample of books and reviewing the experience of PRECIS, an indexing system with some important similarities.

References

Austin, Derek. 1977. *PRECIS: A Manual of Concept Analysis and Subject Indexing.* London: Council of the British National Bibliography.

Dykstra, Mary. 1989. "PRECIS in the online catalog." *Cataloging & Classification Quarterly* 10, no.1-2, 81-94.

Richmond, Phyllis A. 1976. "Classification from PRECIS: Some possibilities." *Journal of the American Society for Information Science & Technology* 27, no. 4, 240–247.

Szostak, Rick. 2011. "Complex concepts into basic concepts." *Journal of the American Society for Information Society and Technology* 62, no.11: 2247-65.

Szostak, Rick. 2013. *Basic Concepts Classification.* Available at: https://sites.google.com/a/ualberta.ca/rick-szostak/research/basic-concepts-classification-web-version-2013

9

Szostak, Rick. 2017. "Facet analysis without facet indicators" In *Dimensions of Knowledge: Facets for Knowledge Organization* (Richard Smiraglia and Hur-li Lee, eds.).

Szostak, Rick (fortchcoming) Theory versus Practice in Facet Analysis. In Aida Slavic and Claudio Gnoli, eds., *Proceedings of the 2017 UDC Seminar.*