

Using Natural Language and SBVR to Author Unambiguous Business Governance Documents

Donald Chapin

Business Semantics Ltd
Flat 21, Dovedale Cottages
240a Battersea Park Road
London SW 11 4LN
United Kingdom

Donald.Chapin@BusinessSemantics.com

John Hall

Model Systems
17 Melcombe Court
Dorset Square
London NW1 5EP
United Kingdom

john.hall@modelsystems.co.uk

Abstract

The Object Management Group standard “Semantics of Business Vocabulary and Business Rules” (SBVR) was designed to enable natural language sentences to be written so they can be read unambiguously by business people, and interpreted unambiguously in formal logic by computers. This paper discusses the key factors that need to be present in addition to SBVR to realize this SBVR design goal.

1 Introduction

Ambiguity in business communication, especially in business governance documents, introduces avoidable business risks. Sometimes these business risks are very costly, even catastrophic, to the organization involved.

The key challenge is to remove ambiguity in governance documents without business authors having to learn new grammar rules.

2 Business Audience – Not IT Audience

SBVR Terminological Dictionaries and Rulebooks “document the meaning of terms and other representations that business authors intend when they use them in their business communications, as evidenced in their written documentation, such as contracts, product/service specifications, and governance and regulatory compliance documents.”¹

SBVR “is conceptualized optimally for business people rather than automated processing. It is designed to be used for business purposes, independent of information systems designs.”²

A key aspect of natural language simplification is to choose one of the synonyms as the preferred

term and use it consistently within a document for each distinct audience.

An organization (*a semantic community*) has speech communities (*audience that shares terms*) that each use a given natural language and, typically, at least three speech communities that use the same natural language, each with a distinct vocabulary with its own preferred terms for the same concepts:

- Employees: jargon, abbreviations, transaction codes, form numbers, etc. But much of the vocabulary would be in understandable business language. It would usually be the most comprehensive vocabulary, providing default terms for the others.
- Legal, for contracts, product and service specifications, compliance reporting, etc. The vocabulary would be formal, include standard legal and industry terminology, and be strictly policed.
- Public, for advertisements, public-facing web sites, scripts for helpdesks, etc. The vocabulary would be everyday language - and probably also be strictly policed.

There would probably also be smaller, specialized speech communities, such as accountancy and finance. Their vocabularies would usually be drawn from the employees’ and legal vocabularies, supplemented by terms adopted from their practices.

2.1 Different Terminological Dictionaries for Different Audiences

Each speech community within the business would have a terminological dictionary, in its own language, that would be a view of the terminological database – a structured subset delivered as a

¹ SBVR Clause 1.4 “Terminological Dictionaries and Rulebooks”

² SBVR Clause 1.2 “Applicability”

report, or the output from a canned query, or a live view via a custom interface.

In a given language within a given business:

- Each concept must have a preferred designation
- Preferred designations may have synonyms:
 - A synonym for a given concept in one terminological dictionary may be a preferred term for that concept in another terminological dictionary.
 - A synonym might not be a preferred term in any terminological dictionary – but may be a synonym in more than one terminological dictionary.

2.2 Terminological Dictionaries are for People; Data Models are for IT Systems

Terminological dictionaries document the meanings intended by business authors for words and phrases they use in their business documents. These documents are used by business people to operate the business.

For example, ISO 1087-1_2000 *Terminology work - Vocabulary - Part 1: Theory and application* defines the meaning of the terms used in ISO 704:2009 *Terminology work – Principles and methods*.

Data Models and their data definitions document the data maintained in IT systems. These models are used by IT professionals for design of IT systems.

For example, ISO 30042 *Systems to manage terminology, knowledge, and content – TermBase eXchange (TBX)* is a data model that documents XML data structures for exchanging terminological database content.

Terminological dictionaries and data models are both important but serve very different audiences and purposes. Neither is an adequate substitute for the other.

2.3 Importance of Context

Dealing with homonyms is essential to removing ambiguity. At the heart of terminology science is the principle that there is a one-to-one relation, in a given context, between a given word or phrase and the concept that designates it.

ISO TC 37 Terminology standards and SBVR together can support several kinds of context for disambiguating part of speech words and phrases:

- Subject field
- Part of speech
- Speech community
- Context concept (disambiguation context)

- Subject concept
- Period of time

In terminological dictionaries the context within which the preferred term and its synonyms have exactly one meaning is explicitly stated in the terminological entry.

When authoring business documents, there are a number of techniques to make the context explicit, thus minimizing the likelihood of linguistic analysis engines getting it wrong:

- Including the intended audience (speech community), the subject field(s), and document applicability dates as document properties
- Including a subject field and/or context concept as metadata in the document's outline headings
- Noting the subject field and/or context concept in a (xxxx, yyyy) notation after the word or phrase.

3 Subset of Natural Language Grammar – Not Artificial Grammar

The approach this paper advocates is to use a selected subset of natural language grammar structures & terms defined in a terminological dictionary. It is *not* to define a new artificial language or artificial extensions to natural language.

This means not changing the natural language syntax of sentences in any way that requires business users to learn new syntax or different interpretations of syntax from what they already know or could know from natural language grammar.

3.1 Keeping Natural Language Grammar Natural

While there is a continuum from:

1. “sloppily, even wrongly, used natural language grammar” through
2. “good quality simple, plain natural language” through (*and across the boundary to*)
3. “additional artificial grammar having to be learned and remembered” through
4. a “fully formal language that looks as much like natural language as possible” *to (but is deceiving like COBOL was)*
5. “formal logic programming languages” (*like Prolog and Datalog*),

the transition from stage 2 to stage 3 is a clearly identifiable boundary that, when crossed, moves from pure natural language to some form of artificial language.

Business people need to be able to use natural language with the help of tools to express business definitions and sentences unambiguously – without being required to learn something that is not part of natural language grammar.

Of course, natural language grammar can be supplemented with good practices for which subsets of natural language grammar structures and patterns are least ambiguous. The “Plain English” requirement of the US Government is an example of this.³

Once one crosses this boundary, the whole approach is on a slippery slope from making it easy for business people to communicate unambiguously to making it easy for IT developers.

If business people have to learn artificial grammar / syntax / notation, there is a shift of responsibility – and effort – for unambiguous communication. Rather than business people working with good semantic authoring tools in their own natural language, they have to speak the language of IT professionals. The more that happens, the greater the risk to the clarity of the documents that people in the business use.

3.2 “Plain Language” as Basis for Least Ambiguous Subset of Natural Language

There is a large international “Plain Language” community that is rich with good practice materials, training, tools and practical experience that support writing using plain language.^{4 5 6 7}

The knowledge, know-how and involvement in unambiguous business document authoring of the Plain Language community fits exactly with the business audience of this approach. This is in sharp contrast to an audience of logicians and IT professionals.

As such, the Plain Language community becomes both the starting point for, and context of use of, the work envisioned in this paper.

3.3 Using Interactive Software Tools to Remove Ambiguity

Clearly, the best of linguistic analysis know-how is needed to optimize resolving indirect references via pronouns, etc.; for determining context for disambiguating parts of speech to a single meaning in the terminological dictionary; and similar sources of ambiguity.

The document author is always the final authority for intended meaning, when linguistic analysis can’t do the job correctly alone.

An example of a sentence where a software tool should ask for clarification is London Underground’s rule:

“Dogs must be carried on escalators”.

This could be interpreted either as:

“A person who is accompanied by a dog must carry the dog when riding an escalator”

or as

“A person may ride an escalator only if the person is carrying a dog”

Compare this with “A hard hat must be worn when visiting a construction site”.

4 Unambiguous Words/Phrases

Ambiguous words and phrases are one of the two major sources of ambiguity in business documentation. Removing ambiguity from part-of-speech words and phrases is the focus of the discipline of terminology science.

Terminology work is standardized in the ISO TC 37 terminology standards with ISO 704:2000 and ISO 1087-1 being the core standards. SBVR builds on the foundation of these standards and adds:

- Semantic features to terminological dictionaries so that the definitions of concepts can be grounded in formal logic.
- The ability to define the skeleton of a sentence clause; i.e. sentence clauses without their quantifications -- typically “subject verb object [preposition object]”. These skeleton clauses are known as “verb concept wordings” in SBVR. In formal logic they are propositions with at least one variable (subject or object) unquantified.

4.1 Importance of Defining Verbs / Verb Phrases and Prepositions

An important application of natural language is in formulating policies, rules and advice to guide the behaviour of organizations and the people in them.

Verbs are the key, but they are often the poor relations in terminology. Governance documents all too often reveal definitions – almost all of nouns – and rules, with nothing connecting them but the assumption that use of the nouns around

³ <http://www.plainlanguage.gov/whatisPL/>

⁴ <http://www.plainlanguagenetwork.org/>

⁵ <http://centerforplainlanguage.org/>

⁶ <http://plain2015.ie/>

⁷ <http://www.intelligentcontentconference.com/>

the verbs in the rules will be commonly understood.

Verbs provide the infrastructure - they connect the nouns with the rules. Nouns have subject and object semantic roles with respect to verbs in sentence clauses. Nouns in these semantic roles denote the roles played in the real world behaviour, represented by the verb, by the real world things in the extensions of the concepts represented by the nouns.

Verbs are modified - with 'must', 'should', 'may' and their negations - to create rules and advice. For example, a car rental business's terminology might include the SBVR verb concept 'credit card guarantees open rental' (where an open rental is one for which the customer has possession of the car).

A 'must' modifier and quantifications added to a single verb concept would create a behavioral rule: 'a credit card must guarantee each open rental'. This is too general. Which credit card has to guarantee which open rental? Other clauses can qualify the nouns to develop the practicable rule the business needs: 'An open rental must be guaranteed by a credit card that is in the name of the customer who is responsible for the rental.'

Structuring verbs into skeleton clauses (*SBVR verb concept wordings*) allows software tools to report on coherence and completeness of bodies of guidance - identifying rules that use undefined verb concepts and verb concept wordings that use undefined nouns. It also enables checking consistency of use of verb concepts across guidance propositions.

Another aspect of formalizing use of verbs is managing different meanings of a verb phrase in different contexts. For example, the rental car company of the example above sells its cars at the end of their useful rental life. In a rental context, 'car is handed over to customer' means 'the car is given to the customer for use for an agreed time and return to an agreed drop-off location'. In a sales context it means 'ownership of the car is transferred to the customer'.

This could be handled by defining narrower categories of the concepts represented by the nouns: 'the rental car is handed over to the rental customer' and 'the sold car is handed over to the purchasing customer'. But the people in the business do not talk or write this way and should not be forced to change their vocabulary. They know what they mean within their context.

Prepositions also have objects and are also part of skeleton clauses (*verb concept wordings*). There are a limited number of prepositions (*only*

around 100 in English) but many prepositions have several meanings. A single vocabulary for prepositions could be adopted into all terminological dictionaries for a given natural language.

4.2 Importance of Defining Adjectives / Adjectival Phrases

Characteristics play a very important role in both ISO TC 37 and SBVR in removing ambiguity. They are the meanings, in natural language grammar, of adjectives and adjectival phrases - just as general concepts are the meanings of common nouns.

Each concept is made up of a set of characteristics: its intension. Characteristics are qualifiers or conditions that narrow the scope of the extension of the concept.

Each intensional definition is composed of a superordinate concept and one or more delimiting characteristics. A set of essential (*necessary and sufficient*) characteristics determines the concept. All other characteristics in the concept's intension are implied from the set of essential characteristics and the other concepts in the terminological dictionary. The set of essential characteristics is the set of all the delimiting characteristics in definitions all the way to the top of the superordinate concept hierarchy.

Sets of essential characteristics have an interpretation in formal logic and map directly to "necessary and sufficient conditions" in OWL. As such they create a direct bridge from normal natural language intensional definitions to reasoning engine models.

Sets of essential characteristics are key to removing ambiguity, not between words/phrases and meanings, but from the meaning of each concept. They also provide a means to determine objectively whether two intensional definitions are semantic equivalents for the same concept, or definitions of two different concepts. This capability comes from the semantic formulation of the definitions of the characteristics, which gives each characteristic an unambiguous meaning (*see next section*).

Characteristics are also powerful as they define conditions which can be used in governance documentation. The adjectives and adjectival phrases serve as condition names for the (usually) longer definitions of the characteristic.

5 Unambiguous Definitions & Sentences

The ability to write sentences in business documents and definitions in terminological dictionaries that are unambiguous both to business people and in formal logic is the key overall capability that SBVR adds to the ISO TC 37 Terminology standards.

SBVR defines a very abstract syntax for specifying the logic structure (*semantic formulation*)⁸ of sentences and definitions that is semantically equivalent to the natural language sentence or definition.

SBVR Semantic Formulations were designed to be easily mappable to natural language grammar.

There are a number of cross-language natural language grammar metamodel standards or de-facto standards that could be mapped to the SBVR Semantic Formulation metamodel, such as:

- ISO TC 37/SC 4 Linguistic Annotation standards⁹
- Penn TreeBank¹⁰ and PropBank¹¹
- NOOJ Text Annotation Structure (TAS)¹²

Software tools that support these natural language metamodels are increasingly being made available as low cost Cloud Services.

Serializations of these models for data interchange are usually specific to a given linguistic analysis tool, but that is a concern for implementers – not of the standard proposed in this paper.

The SBVR approach to writing unambiguous natural language sentences and definitions includes these components, in addition to existing SBVR terminological dictionary and rulebook tools:

- A standard that specifies a cross-language approach to documenting a subset of natural language grammar. This standard should:
 - Select a cross-language linguistic annotation metamodel,
 - Identify the subset of its cross-language grammar structures that can be mapped to SBVR semantic formulations in a way that leaves the fewest opportunities for ambiguity.
 - Provide a mapping of the chosen natural language grammar structures to SBVR semantic formulation constructs.

- Add no new syntax rules that would need to be learned by business people.
- A “Simplified Natural Language” version for the natural language(s) to be used in business documents, preferably US English as the first one. The simplified natural languages should be of the kind, and documented according to the approach, specified in standard.
- An authoring software tool that:
 - Takes text from business documents, preferably as it is being written,
 - Uses the proposed standard and the semantic/logical layer of linguistic analysis to:
 - Identify ambiguous grammar situations,
 - Ask the author for clarification from suggested options,
 - Record all author decisions and/or computer decisions.
 - Uses the above proposed standard to generate SBVR semantic formulations from the definitions and sentences.
- Adoption and/or creation of terminological dictionaries whose concepts cover the content of the documents to be authored.

5.1 SBVR Semantic Formulations: Unambiguous Formal Logic Sentence Equivalents

SBVR does not provide a logic language for restating business rules in some artificial language that business people don’t use. Instead, its semantic formulations provide a means for describing the structure of the meaning of sentences as expressed in the natural language that business people do use.

SBVR Semantic formulations are not representations or expressions of meaning. They represent the logical composition of meaning. They are used to specify the formal semantic structures underlying business communications that comprise concepts, propositions and questions.

There are two kinds of semantic formulation:

- Logical formulations: they structure propositions, both simple and complex. There are specializations for various logical operations, quantifications, atomic formulations based on verb concepts and other formulations for

⁸ SBVR Clause 21 “Logical Formulation of Semantics”

⁹http://www.iso.org/iso/home/store/catalogue_tc/catalogue_tc_browse.htm?commid=297592&published=on&development=on

¹⁰ <http://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf>

¹¹http://clear.colorado.edu/compsem/documents/propbank_guidelines.pdf

¹²https://hal.inria.fr/file/index/docid/498045/filename/Budapest_2008_Disambiguation_Tools_for_NooJ.pdf

special purposes such as objectifications and nominalizations.

- Projections: they structure intensions as sets of things that satisfy constraints. Projections formulate definitions, aggregations, and questions.

Semantic formulations are recursive. Several kinds of semantic formulation embed other semantic formulations. Logic variables are introduced by quantifications and projections so that embedded formulations can refer to instances of concepts.

The following is a simple business rule - one rule, with one meaning, stated in different ways. Other statements are also possible

- It is obligatory that each rental has at most three additional drivers.
- A rental must not have more than three additional drivers.
- No rental may have more than three additional drivers

Figure 1 is a representation of a semantic formulation of the rule, as sentences that convey the rule's full structure.

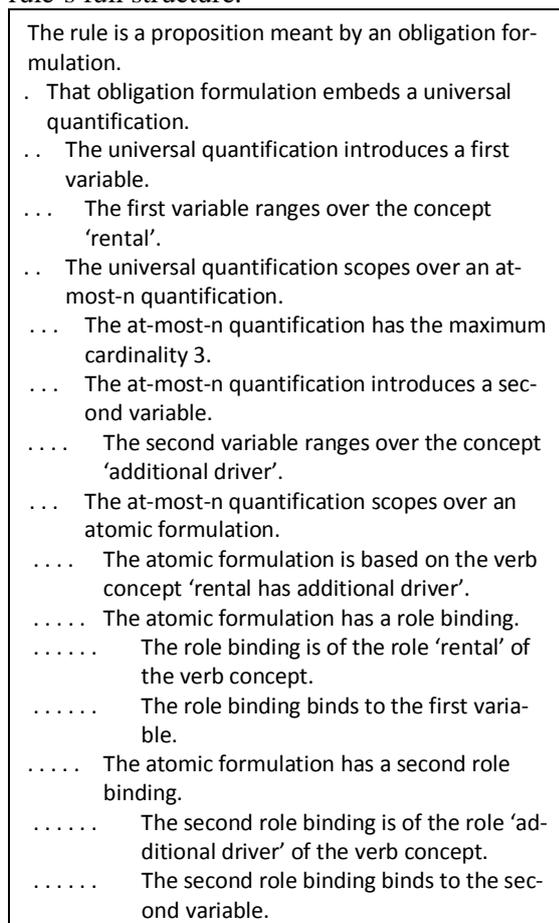


Figure 1: Structure of a Semantic Formulation

The indentation illustrates the composition of a semantic formulation: a hierarchical structure in

which a semantic formulation at one level operates on, applies a modality to, or quantifies over one or more semantic formulations at the next lower level.

Each kind of logical formulation, including modal formulations, quantifications, and logical operations, can be embedded in other semantic formulations to any depth and in almost any combination.

Different semantic formulations are possible for the same meaning. Two semantic formulations can be determined to have the same meaning either by logical analysis or by assertion (as a matter of definition).

Designations like 'rental' and 'additional driver' represent concepts. The semantic formulations involve the concepts themselves, so identifying the concept 'rental' by another designation (such as one from another language) does not change the formulation.

Semantic formulations are structures, identified structurally as finite directed graphs. The reference schemes for semantic formulations and their parts take into account their entire structure. In some cases, a transitive closure of a reference scheme shows partial loops (partial in the sense that only a part of a reference scheme loops back, never all of it).

The main categories of semantic structure of SBVR semantic formulations are:

- Variables and Bindings
- Quantifications
- Logical Operations
- Atomic Formulations
- Instantiation Formulations
- Model Formulations
- Projecting Formulations
- Objectifications
- Nominalizations of Propositions and Questions
- Projections

5.2 Mapping to SBVR Semantic Formulations to remove Remaining Ambiguity

The ISO TC 37 Linguistic Annotation standards fall into these four main categories of annotation:

- Morpho-syntactic annotation
- Linguistic annotation
- Syntactic annotation
- Semantic (logical) annotation

Other linguistic annotation standards and de-facto standards approximate these same categories.

Table 1 provides some examples of mappings from Linguistic Annotation structures to SBVR semantic formulation structures:

Natural Language Annotation Feature	Mapped to SBVR Metamodel Construct
common noun <i>part of speech</i>	signifier of a general concept
common noun <i>in object/subject semantic role</i>	placeholder in verb concept wording representing the general noun concept playing a verb concept role in the verb concept represented by the verb concept wording
proper noun <i>part of speech</i>	signifier of an individual noun concept
proper noun <i>in object/subject semantic role</i>	quantifier for the general noun concept playing a verb concept role in the verb concept being used in a sentence clause
verb or verb phrase (<i>part of speech</i>)	(ISO TBX term of partOfSpeech: verb)
verb phrase <i>in relation to object/subject semantic roles</i>	(<i>part of</i>) verb symbol that is part of verb concept wording that represents a verb concept
preposition <i>in relation to object/subject semantic roles</i>	(<i>part of</i>) verb symbol that is part of verb concept wording that represents a verb concept
"each", "at least one", "a given"	universal quantification
"some"	existential quantification
"a", "an"	universal quantification <i>or</i> existential quantification <i>depending on use</i>
"at least ...n..."	at-least-n quantification
"must"	obligation formulation
"always"	necessity formulation
"that"	<ol style="list-style-type: none"> when preceding a designation for a noun concept, this is a binding to a variable (as with 'the'). when after a designation for a noun concept and before a designation for a verb concept, this is used to introduce a restriction on things denoted by the previous designation based on facts about them. when followed by a propositional statement, this is used to introduce a nominalization of the proposition <i>or</i> an objectification, depending on whether the expected result is a proposition or a state of affairs.

Table 1: Mappings to SBVR Semantic Formulation Structures

Since SBVR semantic formulations are recursive and provide features like objectification and nominalization, they are able to support the most complex sentences.

Figure 2 illustrates how the meaning of a rule can be structured into an SBVR semantic formulation using two stages of linguistic analysis: syntactic and semantic annotations.

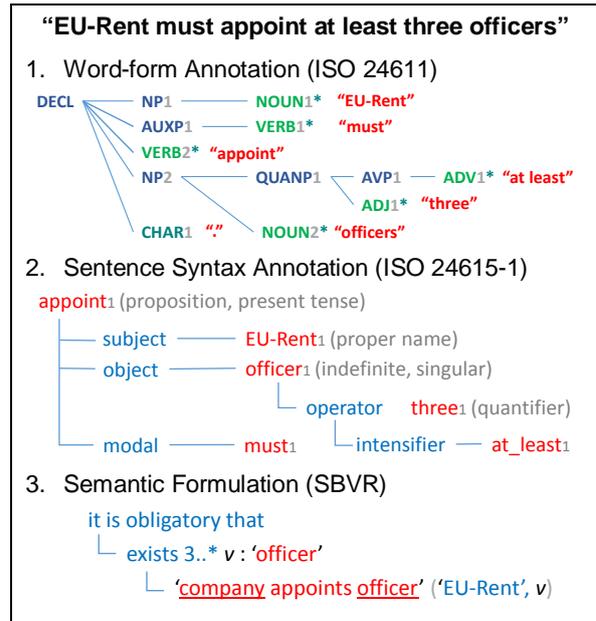


Figure 2: Linguistic Analysis leading to SBVR Semantic Formulations

Lévy and Nazarenko describe a software-supported approach for building sets of business rules from regulatory documents, developed by the Laboratoire d’Informatique de Paris-Nord (LIPN). It uses a three-step process, in which SBVR Structured English stands in an intermediate position between the natural language of the regulatory documents and the formal language of the rules.

This development originated in OntoRule, a large-scale integrating project partially funded by the European Union's 7th Framework Programme. LIPN and Audi AG were two of the partners, and developed the first version of the approach using the EU regulations for car safety systems (brakes, seat belts, air bags). Audi required the business rules as part of its compliance with the regulations.

5.3 The Power of HTML 5 to Bring the Author’s Meaning to Readers

HTML 5 enables a whole new level of semantic markup of the text in business documents, enabling readers to know exactly what meanings the author intended.

HTML 5 semantic markup of part of speech words/phrases, in conjunction with the Unicode character set, can support the following software features:

- An HTML 5 (...) based mark-up for SBVR Structured English text styles for common nouns, proper nouns, adjectives, verbs, prepositions, and SBVR keywords, both as single words and as phrases. This mark-up includes the definition, the

meaning identifier, and all of the contexts required for a unique connection between the word/phrase and the meaning.

- A corresponding MS Word style sheet.
- An AutoComplete feature that inserts the semantic markup behind the word/phrase from the terminological dictionary into the document.
- Mouse-over tooltips in a document that show the concept definition for terms/names or the statement for rule names, along with the context in which the term/name is uniquely connected to this concept or rule.
- Ability to choose which natural language the definitions are displayed in.
- On-the-fly replacement, on simple refresh, in any screen, report or document of all uses of a semantically marked-up term or name when the preferred term/name for that concept/rule is changed.
- On-the-fly redisplay of semantically marked-up terms/names in any screen, report or document with the preferred term/name in the language of the new Speech Community, when the Current Speech Community is changed, whether in the same or a different natural language.
- Optional display of visual font styling for words/phrases that have semantic markup.
- Validation of each definition or rule statement with semantic markup and its SBVR Semantic Formulation against each other.

These capabilities can support multilingual speech communities. Most of them have already been implemented at least once.

6 Conclusion

This paper proposes an approach to using potentially any natural language grammar as a notation for SBVR formally understood definitions and sentences.

The semantic formulation capability of SBVR provides the metamodel of storing and exchanging the semantic structure of the meaning of sentences and definitions in a way that has an interpretation in formal logic.

Linguistic analysis software has now matured to the point where the “semantic/logical” layer of linguistic analysis and annotation enables software tools to generate SBVR semantic formulations from sentences in business documents. The increasing availability of linguistic analysis as low-cost cloud API services makes creating such software tools increasingly feasible.

As part of the generation of SBVR semantic formulations, these software tools can work interactively with business authors while they are writing their documents. They can ask the authors for clarifications, record them, and use them in the generation process. This enables the authors to work with natural language grammar, as is, and frees them from having to learn a new, artificial syntax to ensure unambiguous business governance documents.

The key components that still need to be developed to make this scenario a reality are:

1. A standard that specifies a cross-language approach to documenting a simplified natural language that adds no new syntax rules to be learned and that is as good as practical to be the basis for generating SBVR semantic formulations.
2. A simplified natural language, preferably English as the first one, of the kind specified in item 1, documented according to the approach in item 1.
3. A software tool that takes text from business documents, preferably as it is being written, and uses the semantic/logical layer of linguistic analysis to clarify the meaning of ambiguous sentences and generate SBVR semantic formulations.

The driver is to reduce business risk by providing the people in a business with unambiguous governance documentation, in natural language, using familiar terminology. Requiring them to use an artificial language designed to enable processing by software will work against this.

The language needs an underlying formality to enable bridging to software, but this ought to be ‘under the covers’, not visible to the business users.

The technology, knowledge and know-how to meet both requirements – natural language for the users, underlying formality for the software – already exists, and practice is beginning to grow.

Reference

- [Lévy and Nazarenko] Formalization of Natural Language Regulations through SBVR Structured English (http://link.springer.com/chapter/10.1007/978-3-642-39617-5_5)
- [Ontorule] OntoRule Project (<http://www.ontorule-project.eu>)
- [SBVR] Semantics of Business Vocabulary and Business Rules v1.3, Object Management Group (OMG), <http://www.omg.org/spec/SBVR/1.3/>