Logical Semantics Approach for Data Modeling in XBRL Taxonomies

Olga Danilevitch [0000-0002-7444-0637]

Belarusian State Economic University, Scientific and Research Laboratory for Tax Studies and Tax Policies, Partizanski Ave, 26, Minsk, 220070, Republic of Belarus
Nonprofit information and research organization «Digital Standards of Data Transformation "XBRL BY", Belskogo str, 15, Minsk, 220092, Republic of Belarus
odanilevitch@xbrl.by

Abstract. The contemporary world of human beings is as connected as ever providing for the integration, interdependence and interoperability of various industries and areas of expertise. The World Wide Web enabled by the Internet became a systematic means of communication by use of conventional symbols that represent a language tool. As any language tool, it has two major constructs - syntax and semantics. The syntax of a language defines its surface form. The well-developed, standardized and agreed upon Syntactic Web allows humans to seamlessly send and receive information in a digital form from virtually everywhere in the world. The role of the Semantic Web is to make this information unambiguously understood by both humans and machines. It represents a challenge. Despite the fact that the scientists are equipped with a plethora of methods the Semantic Web remains quite untamed. In this PhD proposal, we suggest the logical semantics approach to the Semantic Data modeling that would allow both analytic and synthetic native language users to build successful Semantic Data Models for an XBRL taxonomy. The perfect datasets to test this approach are generated at the cross-section of multidisciplinary areas of expertise: financial reporting, applied linguistics, natural language processing and computer science.

Keywords: logical semantics, natural language processing, formal language, structural semantics, Semantic Data Model, XBRL taxonomy, analytic and synthetic languages

1 Problem Statement.

1.1 Industry domain description.

The XBRL (eXtensible Business Reporting Language) technology was developed in 1998 with the idea of the machine-based analysis of the financial, regulatory and business reporting. XBRL is standardized outside of World Wide Web Consortium by an independent organization XBRL International. [1] It makes a heavy use of such XML technologies as XML Schema, Namespaces, XPointer, XLink, etc. allowing for the automated exchange of metadata tagged according to XBRL taxonomies.

XBRL taxonomies are electronic directories of XML Schema elements (tags) nested in a hierarchical manner that provide the description and classification system for the content of financial statements and other business reporting documents through a set of linkbases enabled by XLink specification. Each of the XBRL Schema elements are linked to a real-world object representing an accounting, economic or business concept unambiguously defined in the authoritative regulatory document. Essentially, the XBRL taxonomy is the mechanism of the digitization of the regulatory framework and financial reporting supply chain. XBRL standard facilitates the agreement on semantics through creating standard names for business reporting concepts, linking them to their standard definitions in the authoritative literature and providing standard business rules to test and validate the relations between these concepts.

XBRL hierarchical taxonomy closely resembles a star schema in a computer technology where the concepts of the XML Schema are the hub and the linkbases are the nodes projecting from the hub. There can be unlimited number of projections from a single XBRL element through the introduction of dimensions which are descriptive attributes related to the fact data. XBRL binds each reported fact to a concept in a reporting taxonomy. Further utilization of dimensions allows the users of the XBRL-tagged instance documents to "slice and dice the metadata" the way it is done in the data warehouse.

In the past twenty years, XBRL International has grown into a global consortium, which is now comprised of 27 participating country-specific organizations. Yet, there have been only 145 known implementations of the XBRL standard for the financial, regulatory and business reporting in all known jurisdictions. In contrast, in the past fifteen years the number of public API protocols amounted to 21,281 as reported by ProgrammableWeb [2]. Of course, one cannot compare XBRL to API as it would be comparing "apple to oranges", still the difference in the adoption speed and volume is striking. What are the roadblocks on the way of the mass adoption by the worldwide businesses and regulators and why the adoption is slow and challenging?

1.2 The challenge of the successful design of the Semantic Data Model for the XBRL taxonomy is of a dual nature.

It has been often sited in XBRL community that the main challenge of the XBRL taxonomy design is twofold:

1) To design an XRBL taxonomy one has to be a subject matter expert in his/her specific business domain. For example, to be able to design and build an XBRL taxonomy for financial reporting the individual has to be either an accountant or a financial analyst by trade;

2) This particular subject matter expert has to possess the extensive knowledge of XML and XBRL specifications which requires steep learning curve. In other words, the accountant has to learn XML or at some level acquire an assistance of an IT professional who knows XML. It rarely works the other way around, i.e. an IT professional who knows XML would not be able to design Semantic Data Model for the specific industry without a full involvement of the subject matter expert.

XBRL US Style Guide normative document emphasizes that the structure and style of the resulting taxonomy affects developers, preparers of data, and analysts, as well as systems that generate and receive data. This document also specifies that the process is twofold and "...should not be seen as a second part of the taxonomy development process that occurs after the Semantic Data Model has been completed. Neither should the development of the Semantic Data Model occur after developing the Taxonomy.... Rather, creating a Compliant Taxonomy will be an iterative process that involves making changes to the Semantic Data Model during the development of the Taxonomy...." [3]

The complexity and rigor of a technical expertise required from the individual to create a proper Semantic Data Model for an XBRL taxonomy often expands well beyond the primary area of his/her original professional background.

1.3 A natural language component as a third factor added to the challenge of the successful Semantic Data modeling.

In this paper we suggest that the process of the successful Semantic Data modeling is not twofold but threefold with the third component being the **natural language processing** aspect. In fact, this component is so important that it dwarfs the first two while often remains unnoticed as a proverbial "elephant in the room".

Regardless of the natural language spoken by the taxonomy designer and the natural language environment where the work it done; the designer who works on the taxonomy has to take a number of critical steps. These steps require converting the content of the reporting environment from its natural human language to a formalized (formal language) that can further be modeled and understood by a computer.

Human-based approach to the initial data modeling in XBRL taxonomies involves studying of authoritative literature and performing subsequent analysis of the regulatory reports written in the form of text-based documents. While the key semantic meaning behind the same financial notion is predominantly identical across different natural languages, the form of expressing this meaning varies significantly from language to language. For example, "fixed assets" (in English), "основные средства" (in Russian), 固定资产 (in Chinese), "ფიქსირებული აქტივები" (in Georgian), "Tài sản cố định" (in Vietnamese), "Anlagevermögen" (in German), etc. These nouns and noun-phrases have the same meaning of "assets which are purchased for long-term use, such as land, buildings, and equipment, and are not likely to be converted quickly into cash". In other words, their semantics is identical. At the same time common semantics is conveyed using different phonological, morphological, graphical and syntactical means existing in each individual language. In addition to this kind of variability due to different natural languages financial data is often presented in certain contexts, that answer the questions "Who?", "When?", "Where?", "Why?", "During what time?", "Under what circumstance?" etc. These contexts add a great amount of additional variations to the reported data. The contexts also strictly follow the morphological rules of a given language that should not be ignored in the process of the initial modeling. In the proposed research we are going to focus on the differences in the morphology and semantics of the primarily analytic versus primarily synthetic natural languages and the way these differences could influence the data modeling in the XBRL taxonomies.

Our research, therefore would include: 1) collecting linguistic metrics from the quantifiable number of business reporting forms in synthetic and analytic languages; 2) analyzing morphological elements and constructs that convey semantic meanings though grammatical means; 3) defining algorithms that would allow to extract data with common semantics expressed by different syntax and morphology; 4) other types of linguistic analysis required for natural language processing of textual data in order to retrieve unambiguous meanings for future elements of the XBRL taxonomy. The result of the proposed research is expected to be formalized and documented as "how to" protocol for modeling successful XBRL taxonomies. This kind of protocol could be helpful for transition of data semantics from a human language to a formal computer language.

Computers require structure to accomplish given tasks, therefore computer languages are designed to be unambiguous to provide an anticipated result. Structure that is built on top of the highly expressive taxonomies and ontologies allows to generate data automatically from its source with minimal further augmentation and preparation. This structured data can be further analyzed by machine-readable AI mechanisms and make this data comparable and in the most complex contextual settings. To properly design highly expressive taxonomies the data modelers should take into consideration the linguistic aspects of the natural languages, the technical rules of the particular professional domain and the applicable means of computer technologies. We believe that the research of all of these aspects would generate useful results that would further advance the field of the Semantic Web in general.

2 Relevancy.

Interest in Artificial Intelligence (AI), Machine Learning (ML) and automated data analytics has surged to new highs across industries. Corporate finance, Financial reporting, Audit and financial data analysis domains are not exceptions. The Finance professionals cannot help observing and being fascinated by the way this field keeps evolving along with the evolution of information and internet technologies. Accountants, financial analysts, corporate controllers and many other financial professionals are no longer merely exercising accounting judgments but increasingly finding themselves in positions where they have to design, set up, configure and operate complex informational systems. The accounting department can no longer rely on the information technology department to get the financial reports out. In the US this is especially true for public companies that for over a decade have to report their financial statements and notes to the Securities and Exchange Commission in the XBRL-based machine-readable format. The users of the financial information are increasingly relying on the reports that have been digitized and virtualized. Which in its turn puts the pressure on the financial professionals to generate a high-quality output of the data using the most advance technological tools. In the contemporary world of financial reporting XBRL is a mechanism of creating a high-quality financial and business data. XBRL-based business reports are both human-readable and machine readable. The benefits of the machine-readable nature of the structured XBRL reports are such that this data can be further augmented and leveraged by application of the machine learning tools. By creating sophisticated machine-based models, auditors can significantly improve fraud detection, speed up the sampling of the financial records during routine audits or analyze a large number of contracts, such as leases, in much less time than it is possible with a human effort. Data analytics augmented with Artificial Intelligence mechanisms works best in the contexts where AI can make most sense of the data. For complex knowledge domains directly tied to the regulatory pronouncements, rules or laws, the classification has to be designed by humans who understand the technical nature of these business rules. Data modeling in XBRL taxonomies represents an example of a human-based approach to classification of human-readable information into a machine-readable format. We believe that the proposed research would advance the field of the Semantic Web in relation to means of transformation on unstructured data into structured reports.

3 Research questions.

In linguistic topology, an **analytic natural language** is a language that conveys grammatical relationships without using inflectional morphemes. A morpheme is the smallest grammatical, and therefore meaningful, unit in a language. Every morpheme can be classified either as unbound (free) or bound. Unbound morphemes can work in the sentence independently as words (e.g. taxonomy, design, dog, house) or can appear with other basic meaning of units (lexemes) and could be written together or separately (e.g. taxonomy design, doghouse). A **synthetic natural language** is a language with a high morpheme-per-word ratio as opposed to an analytic language. Many synthetic languages evolved from the Proto-Indo-European group of languages that had complex grammatical conjugation, grammatical genders, singular and plural morphemes, inflection of six to eight cases in its nouns, pronouns, adjectives, numbers, participles, demonstrative and possessive identifiers, prepositions and verbal voice.

The terms "analytic" and "synthetic" are used in this paper in a relative rather than in an absolute sense. The difference between synthetic and analytic languages are not always distinct, but rather should be understood as a spectrum. For example, the most widely used contemporary analytic language English that evolved from its Proto-Germanic, Old Saxon and Old English ancestral languages has lost much of their initial inflectional morphology. Another example of a rather analytic language is Dutch that is spoken in the Netherlands. It has morphological features as compounding which makes it more synthetic than English. However, Greenlandic is even more synthetic than Dutch thus pushing Dutch further to the analytic side of the language spectrum. Most of the contemporary European languages that were originally synthetic are currently skewed toward the analytical side of the language spectrum, e.g. new Greek, Italian, Spanish, Portuguese, French, Danish, Dutch and other. On the extreme analytic side of this spectrum there is a type of language that not only has a very low morpheme per word ratio but also no inflectional morphology whatsoever. It is called an **isolating language**. A most common example of the widely used isolating language is Mandarin

Chinese. It has many compound words each representing a separate morpheme with its own semantics, thus giving it a moderately high ratio of morphemes per word. Yet, since it has almost no inflectional affixes to convey grammatical relationships, it is a very analytic language.

3.1 RQ1: Are synthetic natural languages more challenging for the XBRL taxonomy design then the analytic natural languages? What is the difference and how does it play out for the native language speakers?

3.2 RQ2: Is there a universal approach in Semantic Data modeling for the XBRL taxonomy? What is the difference the approach would make for the native language speakers of both analytic and synthetic natural languages?

4 Hypotheses.

Synthetic natural languages are more challenging for the proper initial Semantic Data Model design than analytic natural languages.

One of the most widely used synthetic languages is Russian. It belongs to the Indo-European family of languages, Slavic or Slavonic group, East Slavonic branch along with its closest relatives, Belarusian and Ukrainian languages. Other languages classified into the same Slavonic group are Serbo-Croatian, Macedonian, Bulgarian, Slovene, Polish, Czech, Slovak, and Sorbian. Other examples of synthetic languages are Finno-Ugrian, Turkish, Arabic, Semitic and Native American languages.

Russian is a highly synthetic language, and being a synthetic language, it is **flective,** which means it uses many prefixes, suffixes, it can express in one word what analytic language like English has to use several words for. However, the same flections might express many different grammatical categories while different flections might express the same grammatical category. In Russian the meaning can be conveyed through the means of more than twenty different classes of grammatical tools.

To illustrate what challenge the Russian grammar could present to anyone who attempts to build a Semantic Data Model in a synthetic language environment, we would like to briefly list **the basics of Russian Grammar**:

1) there are three genders: masculine, feminine and neutral; **2)** there are three persons, two numbers (singular and plural) along with the use of an archaic use of dual number from the Old Russian; **3)** nouns, adjectives, pronouns, participles decline; **4)** there are 6 noun cases: Nominative, Genitive, Dative, Accusative, Instrumental and Prepositional. Russian does not have a formal Vocative case that is present in Ukrainian, Polish and many other Slavic languages, but some Russian words retain a Vocative case in archaic spoken forms; **5)** there are 3 classes of noun declension; **6)** adjectives decline according to case, gender and number and agree with nouns in case, gender and number; **7)** there are short adjectives that do not decline; **8)** verbs conjugate according to person, number, tense, voice and mood; **9)** there are two classes of conjugation, 3 tenses (Past, Present and Future) and 3 moods (Indicative, Subjunctive and Imperative); **10)** verbs have two forms: Imperfective and Perfective, similar to English Present and Perfect infinitives, but these two forms in Russian both consist of one word; **11)** participles exist in 4 forms: Present Active, Past Active, Present Passive and Past Passive; **12)** there are short participles corresponding to two Passive forms of regular participles that like short adjectives do not decline; **13)** there are adverbial participles that do not decline and exist in Present and Past forms; **14)** numbers also have several classes: cardinal, ordinal, collective and fractional constructions; **15)** there are no articles; **16)** word order is free, moreover, by changing the word order any word in a sentence can be emphasized.

This fabulously complex grammar makes Russian one of the most expressive, rich and semantically diverse languages in the world. Russian is also a highly synthetic language on the spectrum scale. We can argue that its richness in unbound morpheme applications and grammatical complexity makes is harder for a Russian language native speaker to create a normative Semantic Data Model for the XBRL taxonomy than for the native speaker of the English language.

There is the following hypothesis: when applied to financial and business reporting as to an industrial domain a synthetic language offers the same meaning to a comparable financial report as the analytic language, but it uses a set of far more complex grammatical tools to convey the same semantics. It is possible that if equipped with **logical semantics** mechanism a synthetic natural language speaker could successfully build a Semantic Data Model for the XBRL taxonomy with the comparable ease and technical accuracy as the analytic natural language speaker.

In logical semantics the fundamental relation between an language symbol and its meaning is believed to be not a two- but a three-dimensional: 1) a relation between a language symbol and its meaning; 2) a relation between a language symbol and an object it expresses: 3) a relation of a meaning expressed by a language symbol to the object it expresses. It is a triangle, formed out of two and not one vector. This construct is often called "a Frege triangle" [4] by the name of the author of the one for the first fundamental studies on logical semantics. The terms of this "semantic triangle" were first defined for natural languages but then they have been transferred over to formalized (formal) languages.

Formalized or **formal language** is an artificial language that consists of words (language symbols) whose letters are taken from the alphabet and well-formed according to a specific set of rules. A formal language theory is primarily focused on the syntactical aspects of such languages, i.e. their internal structure patterns. A formal language is often utilized in mathematics, computer science and linguistics. A great example of the way the formal language works is provided by Lewis Carroll in his famous poem "Jabberworky" [5] that illustrates the critical role that function words play in the language (analytic English language in this particular case):

"'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe."

The comparable example that illustrates the exact same formal language theory also exists in the synthetic Russian language: **"Глокая куздра штеко будланула бокра и кудрячит бокренка".** [6] Each and every word in this phase is meaningless, as they do not exist in Russian language. Nevertheless, the use of unbound morphemes tells the reader unambiguously that some female being (not necessarily a human, but most likely an animal) has performed a harsh and single action toward a male being of a different species and is doing something to the male offspring of that male being.

In computer science, formal languages are used for defining the grammar of programming languages. In this paper we suggest that based on the logical semantics approach to the Semantic Data modeling in a given industrial domain we could create a mapping protocol that would allow the users of both analytic and synthetic natural languages to build a successful Semantic Data Model for an XBRL taxonomy. We could also attest that the native speakers of the synthetic language could benefit the most from the application of such logical semantics tool.

5 Approach and preliminary results.

From December 2017 to April 2018 the Nonprofit information and research organization «Digital Standards of Data Transformation "XBRL BY", [8] a Belarusian national jurisdiction of the XBRL International consortium performed a research that was commissioned by the National Bank of the Republic of Belarus. The research studied the possibility of rolling out the XBRL standard of financial reporting for collecting financial information from a small subset of non-banking financial institutions in Belarus, the companies who provide their customers with access to the financial arbitrage (primarily FOREX trading companies). During this research a high number of standard financial regulatory forms were analyzed from the perspective of creating a working Semantic Data Model for the XBRL taxonomy for the FOREX trading companies. While common face financial statement did not represent any significant challenge, there were certain forms required for reporting of the detailed complex financial information that involved derivatives trading. To accommodate for this complex reporting the **logical semantics** approach was implemented to be able to convey the most accurate information most fully compliant with the regulatory requirements and the rules of XBRL taxonomy modeling. The following example represents logical semantics approach to the analysis of the Table 1, Attachment 7 of the SOP (Standard Operating Procedure) of the National Bank of the Republic of Belarus N72 dated February 12, 2016 and is available at XBRL.BY website via **http://xbrl.by/lsa-sl/.** [7]

The presented Semantic Data Model has been designed based on logical semantics approach applied to the reporting framework of the synthetic natural language environment. We argue that with the application of the logical semantics we have successfully transferred the semantics of the initial paper-based Table 1 from its natural language into a formal language further allowing to build an XBRL taxonomy package that the XBRL parser can automatically process.

6 Related work.

This paper focuses on the Semantic Data modeling of the datasets found in the cross-section of the three areas of expertise: linguistics, computer science and financial and business reporting. To address the aspects of the taxonomy modeling from the perspective of all involved areas of expertise, we looked into the related work in distinctly different industries.

In "A Theory of Structural Semantics" Abraham and Kiefer [8] made an attempt of formalization of semantic aspects of the natural languages providing a mechanism of a **deductive semantics**. They observe that the deductive method has to work with simplified concepts and self-imposed limitations, therefore suggesting that the fundamental semantic concept cannot be used in their entire generality in the deductive semantics. Their theory is based on the transformational grammar whose central assumptions are explained by the authors. They restrict their semantic analysis mainly to a sentence as a basic unit of transformational grammar. Their semantic theory includes two primitive terms and three basic definitions. The primitive notions are "morpheme" and "category" and the basic definitions are: 1) a rule is a formalized linguistic relation; 2) a word is a "sentence of morphemes" that is derived by applying rules of morphemes; 3) a sentence is a distinguished category of the auxiliary vocabulary that is one of the subsets of the broader vocabulary of the natural language. In the authors' theory the semantic characterization of a word is provided by a **labeled tree graph** beginning with the word followed by a grammatical category designator or designators indicating the class or classes to which the word belongs, further followed by the semantic designators indicating the meaning class or classes to which the word belongs.

If we apply the deductive semantics mechanism to the dada modeling in XBRL taxonomies we find interesting similarities. Semantic relationship inside labeled tree graph closely resemble relationships within XBRL taxonomy hypercubes. In XBRL hypercube is a fundamental building block of the multidimensional model which can be described as an "n-dimensional" matrix or array with an infinite number of dimensions. As we mentioned in the Introductory section XBRL hierarchical taxonomy closely resembles a star schema in a computer technology where the concepts of the XML Schema are the hub and the linkbases are the nodes projecting from the hub.

In order for the business data analysis to be most effective the taxonomy used to create this data has to be highly expressive. The highly expressive taxonomy needs to be modeled keeping in mind that typical business information is multidimensional. Dimensions built into a taxonomy create a model for expressing characteristics of information in infinite variability of contexts. Hypercubes can be hard to describe in two-dimensions, i.e. in a paper-based document. Computer software, on the contrary, can process dimensions and express information in hypercubes very well. The following example of dimensional relationships in a hypercube is created as a part of the research commissioned by the National Bank of Republic of Belarus and is available at XBRL.BY website via **http://xbrl.by/lsa-al/.** [9] It illustrates business information converted from textual data in the **analytic natural language** (American English) into a structured data in the XBRL instance document.

Russian linguist Ju. S. Martem’janov in his paper "Valency-Junction-Emphasis Relations as a Language for Text Descriptions" [10] suggests that grammar in a synthetic language consists of at least three systems. The first system, called **"valency" grammar** represents a type of case-grammar where predicative are described in terms of "valencies" (roughly: cases). This grammar is conceived in order to express the abstract relationship between lexical elements of a simple sentence. The second system, called **“valency-junctive grammar”** is constructed in order to describe relations between word groups including sentences. The third system, called **“valency-junctive-emphasis grammar”**, is meant to provide a means for the description of logical emphasis and topicalization in general. What is meant by here by grammar can be termed "logical syntax" since Martem’janov takes into consideration only abstract semantic relations. This type of the classification can also be described as three-level generative semantics. We find the Martem’janov three-level valency grammar theory useful for the analysis of the complex sentences in the synthetic language environment.

Another interesting example of the use of a linguistic model is described by Greek architect Chris I. Yessios in his paper "A Linguistic Model for 3-D Constructions"[11]. Yessios is known by his input into development of innovative techniques for use in environmental design, architectural modeling and computer-aided design. His paper presents a linguistic model for generation of three-dimensional structures used in physical constructions of building. In his paper he suggests that a 3-D structures-oriented model should contain (1) means for the definition of the primitive elements (to be called "objects" and (2) means for composing them to derive composites. These basic functions should also include the capabilities of deriving variable copies for a single element and the capability of operating upon (sculpting) the original form of an element to derive a new more complex form. For the definition of primitives, the existence of regularity and standardization suggest that frequently used shapes should be parameterized. The model also accepts the definitions of arbitrary irregular objects, so the generality requirements are also satisfied. Each primitive (and composite) object is defined with respect to a local system of orthogonal axes. For each object the origin of its local system of axes is designated as its reference point. Since all other points of an object are defined with respect to its reference point, in a way, the reference point can be viewed as representing the whole object. We find it fascinating how closely the Yessios' 3-D model for real-world objects resembles the multidimensional model of XBRL taxonomy. In multidimensional model a set of context oriented orthogonal axes and domain members nested within a hypercube in a parent-child hierarchy allows to "slice and dice" of the information with respect to every specific context.

XBRL US jurisdiction of the XBRL International Consortium has done significant amount of work to facilitate the building of high-quality taxonomy through Technical Guidance, Certification and Governance. One of the most comprehensive documents that provide the detailed guidance on the modeling of the taxonomy concepts is "XBRL US Style Guide" [3]. Section 3 of this document specifically addresses the language guidelines, specifying what SHOULD and SHOULD NOT be used in the process of concept modeling. For example, the use of nouns is required; the use of articles is restricted as well as the use of the pronouns; adverbs are restricted with the exception when they may represent a recurring subject, etc. In section 4.4.2. of XBRL US Style

Guide there is a guidance on the order of adjectives immediately preceding a noun in the concept naming, i.e. 1) Quantity; 2) Opinion, 3) Size; 4) Physical Quality; 5) Shape; 6) Age; 7) Color; 8) Origin; 9) Material; 10) Type; 11) Purpose. This specific guidance corresponds to the adjective order in English language grammar. It allows the users of the analytic natural language to rely on natural language grammar rules for taxonomy modeling. In contrast with the analytic language, there is no comparable strict order of the adjective in the noun phrase in the synthetic natural language. In their paper "The order of attributive adjectives in the history in Russian and the position of adjectives in the noun phrase" P. Grashchenkov and O. Kurianova [12] researched the order of different semantic classes of attributive adjectives and possible implications for the semantic hierarchy based of the (non)-observed linearization. Two corpus-driven studies are presented in this paper. The first study is focused on contemporary Russian, the second deals with the complex corpus, consisting of Old Church Slavonic and Old Russian texts from XI to XVII centuries. Although quantitative analysis shows the tendency towards ordering of attributive adjective in a noun phrase, this tendency is not strong and regular enough. The paper concludes that attribute adjectives cannot be viewed as representing syntactically ordered functional structure the way it is normally defined in analytic natural language. This is a good example of the type of challenges that users of the synthetic natural language would incur in semantic modeling of the multidimensional XBRL taxonomies.

7 Evaluation plan.

The logical semantics approach to the Semantic Data modeling suggested in this PhD proposal would enable its users to design XBRL taxonomies that could be successfully mapped at the data source where the data is originated. Almost every contemporary corporate accounting system contains a dormant XBRL module that is not normally utilized. It is primarily due to that fact that additional mapping to external standard XBRL taxonomies is required in order to pull the transaction level data and further aggregate it for the regulatory reporting. We believe that with the proposed logical semantics approach it is possible to design an algorithm compatible with an accounting system enabling accurate extraction of a structured data either for the corporate internal use or for the required regulatory reporting. We plan to test this algorithm on the subset companies of the national financial market where XBRL BY is currently operating.

8 Reflections.

XBRL as a Semantic Data standard has significantly matured since its inception in 1998. In the past twenty years, a lot of guidance has been accumulated through the initiatives of the XBRL International community and its national jurisdictions. We look forward to utilize this wealth of knowledge and push it forward with a practical implementation. We believe that the detailed analysis of the complex grammar of the synthetic natural language could bring interesting results that would enable algorithmization of the logical semantics approach that we are proposing. The other main key to our success is the wholesome multidisciplinary approach at the cross-section of the financial and regulatory reporting, structural and applied linguistics and XBRL technical standard.

Acknowledgments: I would like to express my special thanks of gratitude to my PhD advisor, the Head of the Scientific and Research Laboratory for Tax Studies and Tax Policies at the Belarusian State Economic University Irina Loukianova for the valuable subject matter expertise; Nikita Smirnov, a graduate student at the Belarusian State Economic University for volunteering at the XBRL.BY national jurisdiction; the National Bank and the Ministry of Finance of the Republic of Belarus for facilitating the national XBRL adoption. This PhD research is entirely funded by its author.

References

1. XBRL International homepage, https://www.xbrl.org/, last accessed 2019/07/02
2. ProgrammableWeb: https://www.programmableweb.com/api-research, last accessed 2019/04/15
3. XBRL US Style Guide. A Language Guide for Creating Concepts and Labels. https://xbrl.us/wp-content/uploads/2017/09/style-guide-20170907.pdf, last accessed 2019/06/14
4. Gottlob Frege: ‘Über Sinn und Bedeutung’, in *Zeitschrift für Philosophie und philosophische Kritik*, 100: 25–50. Translated as ‘On Sense and Reference’ by M. Black in *Translations from the Philosophical Writings of Gottlob Frege*, P. Geach and M. Black (eds. and trans.), Oxford: Blackwell, third edition, 1980.
5. Lewis Carroll: Jabberwocky. The Random House Book of Poetry for Children (1983).
6. This phrase is considered to be authored by the famous Russian linguist Lev Scherba who frequently used it as an example during his lectures on linguistics and lexicology in 1930s.
7. XBRL Belarus national jurisdiction homepage, http:// http://xbrl.by/lsa-sl/, last accessed 2019/07/02
8. Samuel Abraham, Ferenc Kiefer: A Theory of Structural Semantics. Mouton & Co, The Haugue – Paris (1966).
9. XBRL Belarus national jurisdiction homepage, http:// http://xbrl.by/lsa-al/, last accessed 2019/07/02
10. JU. S. Martem’janov: Valency-Junction-Emphasis Relation as a Language for Text Description. In: Trends in Soviet Theoretical Linguistics, pp. 335–388. D. Reidel Publishing Company, Dordrecht - Holland / Boston - USA (1973).
11. Chris I. Yessios: A Linguistic Model for 3-D Constructions. In: Quantitative Planning and Control edited by Yuji Ijiri and Andrew Whinston, pp. 37–57. Academic Press, New York, San Francisco, London (1979).
12. P. Grashchenkov, O. Kurianova: The order of attributive adjectives in the history in Russian and the position of adjectives in the noun phrase. Rhema. 2018. № 4. ISSN 2500-2953.