Academic Style Marker Ontology Design

Viacheslav Lanin and Sofia Philipson
National Research University Higher School of Economics, Perm, Russian Federation
vlanin@hse.ru, lyubov.filipson@inbox.ru

Keywords: Style Marker, Scientific Paper, Ontology, DSL.

Abstract: As any other genre, academic paper can be characterized by its own specific rules and fitches. The authors assume that academic style fitches called in this research "style markers" can be modelled by means of ontology engineering. The article is aimed at describing the academic style markers ontology design and its practical using. The designed ontology is divided into two levels. The first level provides information about linguistic terms and the second level consists of style markers, which were suggested by experts in linguistic. It is assumed that two tasks will be solved on the basis of developed ontology. The first task is generating lexical-semantic templates, which is used to identify the list of markers in a text. Due to ontology approach and Domain Specific Language (DSL) technologies applying users can be able to extend and modify marker templates. The second task is developing an expert system rules for text style enhancement.

1 INTRODUCTION

The contribution of research results in scientific publications is the most significant academic performance indicator of scholars and research co-workers. Papers written in English extend the audience match but non-native scholars usually face some difficulties connected with the conventions of formal academic writing style when write in English. A huge number of researchers have investigated academic writing features in general (Wallwork, 2016; Biber & Gray, 2016; Strongman, 2014; Castelló & Donahue, 2012; Hyland, 2009), in terms of grammar (Wallwork, 2013), structure (Sawaki, 2016), genre (Bruce, 2008) and other crucial features of academic writing. Much of the previous research projects were focused on identifying and evaluating the features of academic writing implementing cross-cultural and cross-linguistic approaches (Lakić et al., 2015). However, not enough attention has been paid to building a systematic approach to academic writing as a segment of academic discourse both oral and written as a functional style, a segment of the language in which coherence and formality of the discourse being its most prominent features determine the choice of other elements.

Literature review has shown that recommendations given in guides and handbooks for both competent and novice academic writers in English (Belcher, 2009; Bailey, 2011) are not systematized and sometimes even have obvious internal contradictions.

Typical webpages of university writing centers define academic writing as having several distinctive features: "*Academic writing is to some extent: complex, formal, objective, explicit, hedged, and responsible. It uses language precisely and accurately. It is also well organised and planned*". However, such definitions do not help understand importance of particular of academic writing features and their hierarchy. Therefore, the purpose of our research is developing the ontological system of academic writing features. In this project (Lanin, 2015) we attempt to extract style markers, define interrelations between them and design the style model of academic English. Investigation of hierarchical relations between style elements are also crucial as it helps to determine their frequency occurrence in English scientific texts and describe usage pattern of these elements on the texts pieces of different levels.

Traditionally, functional styles have been explored by analyzing few texts representative for particular style while at the moment more valid data might be received by analysis of large corpora using computer technologies. This enables the processing of huge corpora and get empirical material which can help create language patterns, study language consistency, and describe linguistic phenomena typical of a particular language area, i.e. derivation of

Lanin, V. and Philipson, S.
Academic Style Marker Ontology Design.
In *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2017) - Volume 2: KEOD*, pages 189-194
ISBN: 978-989-758-272-1

style markers. In the paper, we adopt the term ‘style marker’ instead of a broader ‘style feature’ due to several reasons. 'Style feature’ represents a more vague conception of a characteristic while ‘style marker’ being ‘a sign’. Therefore, style markers in this paper are considered as main features of academic English in its linguistics meaning.

The data, collected from corpus annotated in accordance with the style markers identified by experts, give the information about the frequency of occurrence of the elements and their leading or minor role in building academic functional style.

2 RELATED WORK

One of the actively developing branches of theoretical and applied stylistics is a complex analysis of English scientific papers published in most leading peer-reviewed academic journals conducted through processing and comparative stylistics study and critical discourse analysis of the large text corpora. The analysis of English academic text produced by non-native author offers the greatest challenge of corpus linguistics research and the field of software development for corpus analysis (Jeffries, 2009). It is worth saying that there has been little quantitative analysis of large corpora of written academic speech. What is not yet clear is the hierarchy of linguistic elements of academic writing what is a serious obstacle to describing English of the particular subject area with certainty, identify key features and study usage pattern. The usage of computer technologies simplifies statistical processing of large corpora in linguistic research.

At the moment, a great variety of tools for corpus processing exist. The most widespread of them are AntConc, WordSmith Tools, Gate Developer, Sketch Engine and CQPweb. There are specialized solutions for academic papers style analysis, for example project Fapas (Full Automatic Paper Analysis System) (Scholz & Conrad, 2011).

It is also possible to find projects connected with the creation of ontologies, which describe linguistic domain (Zagorulko et al., 2010). One of them is GOLD ontology which is General Ontology for Linguistic Description (Farrar, 2003). It gives the description of linguistic basis including most foundational categories and relation between them. The ontology is connected with SUMO (Standard Upper Merged Ontology) (Pease et al. 2002) which is based on four main domains: expressions, grammar, data constructs, and metaconcepts.

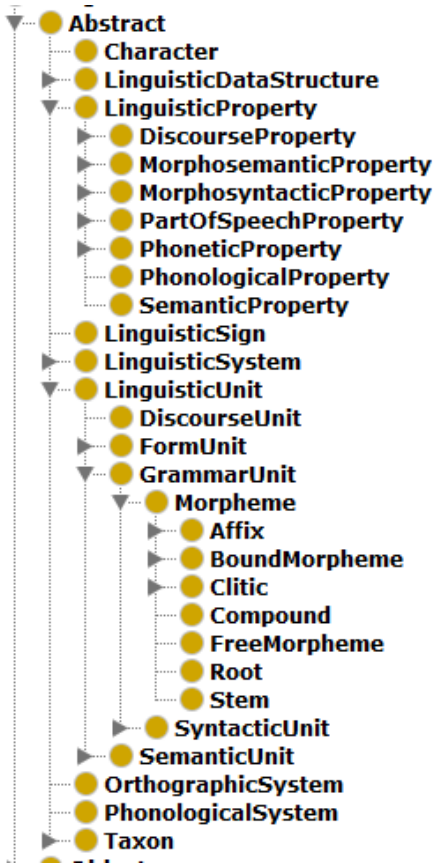

Figure 1: Visualization fragment of ontology concepts.

The category expressions mean the physically accessible aspects of language. The base for this aspect was taken from SUMO and to the concept LinguisticExpression were added new ones like WrittenLinguisticExpressions and SpokenLinguisticExpressions.

Grammar category includes the abstract properties and relations of language, the domain that is primary interest to linguists. It means that anything expressed by a grammatical system is represented by the concept GrammaticalCategory.

Data constructs are used by linguists to analyze language data, such as paradigms, lexicons and feature structures. Metaconcepts are the most basic concepts of linguistic analysis, including language itself. There are many ways in which language can be viewed and without a working concept of the language, an ontology cannot be used to describe and compare data of different languages. Language was defined as the set of data associated with a common grammatical pattern. Generally speaking, the ontology tries to describe all the aspect of the natural language which can be applied to all languages.

Another example of ontology is also from linguistic field but it is concentrated on computational linguistics. Developed ontology is built on the basis of scholarly knowledge ontology and is divided into five hierarchies "whole-part" which are connected to each other with associative relations. The subject of investigation of computational linguistics is the properties and the systems of linguistics units, operations, connected with their functioning in the process of communication, and application processes replied to defined request.

3 THEORETICAL BACKGROUND

The theoretical foundation of the system described in this paper consists of a list of style markers that were selected from reference and study materials, Internet resources about academic writing as well as scientific papers devoted to this topic. All markers from this list can be divided into three main groups: lexical markers, grammar markers, syntactic markers.

Lexical markers include three types of features:

- specific words and terminology (high frequency of terminology; usage of abstract semantic verbs, desemantisized verbs, intensifying adverbs; low frequency of personal pronouns *you*, *he*, *she*);
- words corresponding to specific word-formation constructions (nouns with -or suffix, commonly used in terminology; abstract nouns derived by suffixes -ment, -ness, -tion, etc.);
- words in specific part of speech (high frequency of nouns, low frequency of pronouns).

Two types of features that belong to grammar markers category are:

- wide usage of verbs in the Passive Voice;
- presumable prevalence of verbs in Present Tense.

Syntactic markers can also be classified into two types:

- features described by syntactic structures (simple, complex or compound sentence structure; prepositive and postpositive attributes by most of the nouns; possible prevalence of prepositive attributes in technical texts);
- specific conjunctions, linking expressions, etc. (subordinating and correlative conjunctions; archaisms thereby, therewith, hereby; prepositional phrases; means of logical expressions).

Most of these features can be automatically annotated using lexical-syntactic patterns, although absolute accuracy cannot be guaranteed. That's is why expert control and means of manual annotation correction are highly desirable for system implementation. Flexibility of the system components is also important for development and further testing and debugging due to specificity of academic style feature tagging and natural language processing in general.

Currently our system annotates texts based on all of the described style markers except terminology and sentence structure. Although some components are still being tested, recent resulting annotation sets provide enough information to analyze academic writing and deepen the studies about some of the features.

For the present style markers are represented as desperate data set. Language is dynamic and always developing system, so besides markers systematization the method should give the opportunity of enlargement and adaptation.

4 ACADEMIC STYLE MARKER ONTOLOGY

In the paper the way of regulation and systematization of disparate data set called ontology is described. The ontology reveals dependences between entities in the form of style markers, and any existing interconnections are indicated. Thus, a huge variety of different style markers turn into a controlled system which then can be used as a part of larger project focused on improving the quality of text annotating.

The combined approach to class hierarchy development, which is based on dealing with the most significant terms where the developer then categorizes or typifies them, was chosen to design the ontology.

After the data analysis of GOLD ontology was done, it was decided to design the ontology which will be based on the part of GOLD ontology, describing lexical and syntax terms, i.e. the most relevant terms of written English.

Figure 2: Main parts of ontology.

All aspects which were derived by experts earlier are the part of subclasses called *LinguisticProperty* and *LinguisticUnit*. It was decided to create the main classes which are: *LinguisticPropery* and *LinguisticUnits* – linguistic terms, *Aspect* – aspects derived by experts and *StyleMarker* – markers which were also given by experts. Class *StyleMarker* represents particular style markers, which are the features of written academic English. This class is made of terms that were described earlier. Subclasses of *StyleMarker* include individuals, which are style markers of written academic English. Markers have not any relations between each other but they are connected with particular aspect, which them expresses. Experts derive six main aspects of academic English. They are not connected with each other like markers, however they are connected with terms of *StlyleMarker* subclasses as well as Linguistic subclasses. At the current moment, there is only hierarchic dependence between classes and subclasses. Individuals of ontology are particular examples of style markers of written academic English.

It was decided to link the terms of classes named *Aspect* and *StyleMarker* with properties express and it's inverse property *isExpressedBy*. Every class of aspects is expressed by subclass/subclasses of particular style marker of written academic English. Now we will talk about properties which are used for linking of subclasses of *Aspect* with subclasses of *StyleMarker* and *LinguisticPropetry*. For example, the graph, which is represented on the figure 4, shows the relations of aspect named *Adverb*.

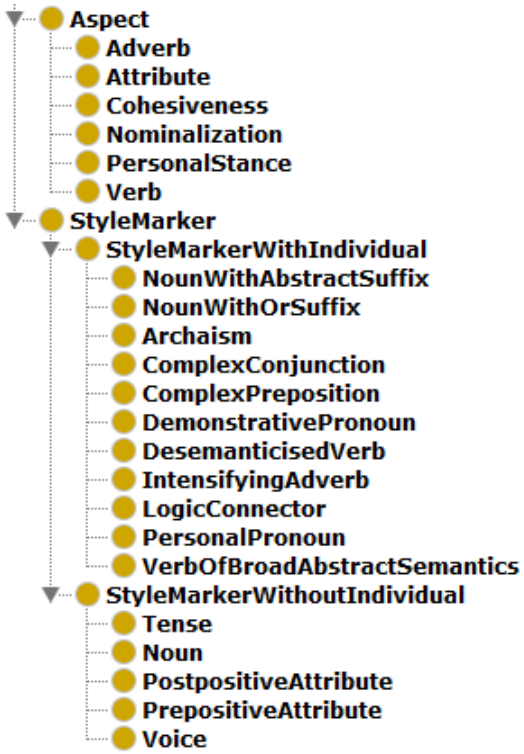

Figure 3: Ontology concepts hierarchy.

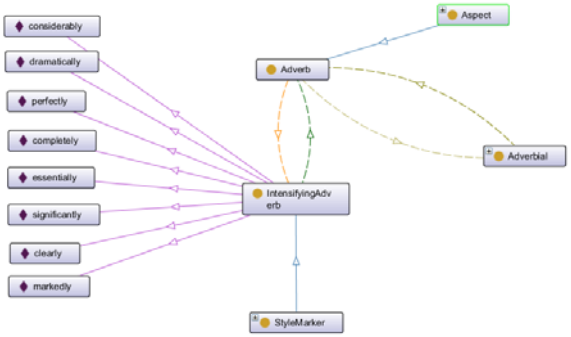

Figure 4: Relations and instances of «Adverb» aspect.

It can be seen that this class is a subclass of class named *Aspect* and class *Adverb* is connected with a subclass *IntensifyingAdverb* of class *StyleMarker*. The following properties were created: *Adverb isExpressedBy IntensifyingAdverb*, which means that aspect *Adverb* is expressed by style marker *Intensifying adverb*. Besides there is one more property *hasIndividual* and particular style marker's individuals. If we talk about connection between linguistic terms and aspects there are properties called include and *isPartOf*: *Adverb* isPartOf *Adverbial* and it's inverse sentence with property include.

Class *Verb* is subclass of class named *Aspect*, it is connected with subclasses *Voice*, *Tense*, *DesimanticisedVerb* of class *StyleMarker*. The properties are: *Verb isExpressedBy Voice*, *Tense*, *DesimanticisedVerb* and *Voice*, *Tense*, *DesimanticisedVerb express Verb*. The relations between linguistics and aspect are following: *Verb isPartOf Verbal* and *Verbal include Verb*.

Graph of aspect called *Cohesiveness* shows relations between style markers and linguistic terms. This aspect is expressed by subclasses *ComplexConjunction*, *ComplexPreposition*, *LogicConnector* and *Archaism*. This aspect is part of *Functor linguistic* term.

Aspect *Nominalization* is expressed by *Noun* and NounWithAbstractSuffix. *NounWithAbstractSuffix* has the property *hasSuffix*, which refers to abstract suffixes, which are represented as individuals of class *AbstractSuffix*. Aspect *Nominalization* is a part of linguistic term *Noun*.

Aspect *PersonalStance* is expressed by style markers *PersonalNoun* and *DemonstrativePronoun*. *Aspect* is connected with linguistic term *PersonalPronoun*. *Style markers* has property *hasIndividual*, that is shown on the figure 5.

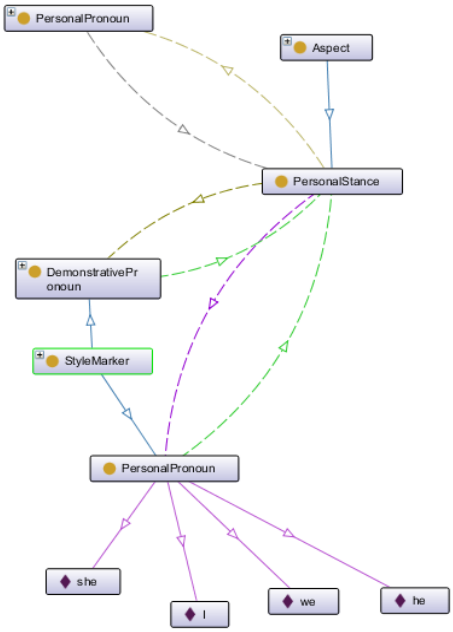

Figure 5: Relations and instances of *PersonalStance* aspect.

5 PATTERN GENERATION BASED ON DSL-TECHNOLOGIES

Ontology is necessary not only for style markers systematization but also as the foundation of lexical-semantic patterns (IJntema, W. et al., 2012) generation. Currently, JAPE-templates (Java Annotation Patterns Engine) of GATE (Cunningham et al., 2011) text processing system are used. Rule generation architecture is demonstrated on figure 6.

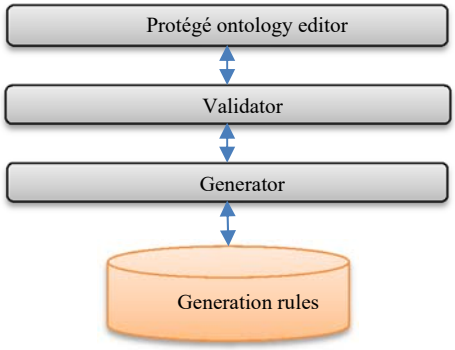

Figure 6: The architecture of the lexical-semantic pattern generator.

Protégé ontology editor is used for ontology describing and its representation in the OWL format. Validator is the component which is meant for accuracy check of user's models. While designing a model, the user can make some mistakes or make models which are not satisfy the ontology limits constraints. Generator is the component responsible for code generation on target language. Generator is used for transformation of user's models into textual representation on the description language of lexical-semantic patterns as well as file generation into the formats of the computer linguistic systems for example JAPE. To extend the interoperability the system gives users the opportunity of manual determining the transformation rules. It is crucial on this level of metamodel to make text pattern for every language element in accordance to which code generation would be implemented. Text pattern includes the statistic part which is not depended on certain model and the dynamic part, which makes possible the reference to attributes values of different DSL-constructions.

General algorithm of search pattern generating on the basis of ontology consists of several steps:

1. Get all the individuals of the subclasses which main class is StyleMarker.

2. Get data property Template, which contains JAPE - expression to search particular marker.
3. Get JAPE –expression or name of file which is aimed on searching of marker.
4. Combine expressions with the help of OR-operator in one total JAPE - expression, get to file name, consisted style marker search implementation.

To perform these steps SPARQL-queries should be written.

6 CONCLUSIONS

The standard tools and software applications are used to design the ontology which simplifies the process of development and decision maintenance process. The described approach has an expanding property, i.e. in order to add a new marker the user needs to add its description and the identification rule will be generated automatically. Moreover, using the linguistics level, described in the ontology, makes the description of related domains possible.

ACKNOWLEDGEMENTS

The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2017 (grant № 17-05-0020) and by the Russian Academic Excellence Project "5-100".

REFERENCES

Bailey, S., 2011. *Academic Writing: A Handbook for International Students*, Taylor & Francis.

Belcher, W.L., 2009. *Writing Your Journal Article in Twelve Weeks: A Guide to Academic Publishing Success*, SAGE Publications.

Biber, D. & Gray, B., 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*, Cambridge University Press.

Bruce, I., 2008. *Academic Writing and Genre: A Systematic Analysis*, Bloomsbury Publishing.

Castelló, M. & Donahue, C., 2012. *University Writing: Selves and Texts in Academic Societies*, Emerald.

Cunningham, H., Maynard, D. & Bontcheva, K., 2011. *Text Processing with GATE (Version 6)*, University of Sheffield.

Farrar, S. & Langendoen, T., 2003. A linguistic ontology for the semantic web. *Glot International*, 7, pp.97–100.

Hartley, J., 2008. *Academic writing and publishing: A practical handbook*, Routledge.

Hyland, K., 2009. *Academic Discourse: English In A Global Context*, Bloomsbury Publishing.

IJntema, W. et al., 2012. A lexico-semantic pattern language for learning ontology instances from text. *Web Semantics: Science, Services and Agents on the World Wide Web*, 15, pp.37–50.

Jeffries, L., 2009. *Critical Stylistics: The Power of English*, Palgrave Macmillan.

Lakić, I., Živković, B. & Vuković, M., 2015. *Academic Discourse across Cultures*, Cambridge Scholars Publishing.

Lanin, V., Strinyuk, S. & Shuchalova, Y., 2015. Academic papers evaluation software. In *2015 9th International Conference on Application of Information and Communication Technologies (AICT)*. pp. 506–510.

Pease, A., Niles, I. & Li, J., 2002. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. *Imagine*, 28, pp.7–10.

Sawaki, T., 2016. *Analysing Structure in Academic Writing*, Palgrave Macmillan UK.

Scholz, T. & Conrad, S., 2011. Style Analysis of Academic Writing, pp.246–249.

Strongman, L., 2014. *Academic Writing*, Cambridge Scholars Publisher.

Wallwork, A., 2013. *English for Academic Research: Grammar, Usage and Style*, Springer, Boston, MA.

Wallwork, A., 2016. *English for Writing Research Papers*, Springer, Cham.

Zagorulko, Y., Borovikova, O. & Zagorulko, G., 2010. Knowledge Portal on Computational Linguistics : Content-Based Multilingual Access to Linguistic Information Resources. In *Selected topics in Applied Computer Science. Proceedings of the 10th WSEAS International Conference on Applied Computer Science (ACS'10)*. pp. 255–262.