https://doi.org/10.26552/com.C.2017.1.86-90

Marek Debnar*

THE SEMANTIC FIELDS OF SELECTED ETHICAL TERMS IN THE WRITTEN AND WEB SUBCORPUS OF THE SLOVAK NATIONAL CORPUS

The paper investigates the issue of semantic field of selected ethical terminology (e.g. morality, conscience, moral action, etc.) in the Slovak National Corpus (SNC). The examination covers two different subcorpuses (written and web) containing a collection of about thousand text units. The basic framework of written subcorpus consists of scientific articles from the field of humanities, religion and art. The web subcorpus was generated from website www.salon.eu.sk offering essays and feuilletons published by the European and world press, with special focus on Central European countries. Firstly, the paper describes the way of designing and building these subcorpuses based on critical interpretation of text metadata, and, secondly, it focuses on the rate of occurrence for the subject areas − Domains and Subdomains − that refer to ethical terminology included in the SNC. The paper then outlines the ways of interpreting the statistical results of the quantitative occurrence of the selected word forms, and it also defines the ways of using the subcorpus not only for linguistic and literary purposes, but also in interdisciplinary research areas.

Keywords: *Ethics, Slovak National Corpus, Style and Genre annotation, Semantic fields.*

1. Introduction

Methods used in the humanities are undergoing a period of transformation caused by the application of digital technologies in all areas of scientific research. The aim of this paper is to present the procedure used when constructing a genre subcorpus and to represent the results of this study – an approximation of the semantic fields of texts from the essay genre. Another goal is to present the results of related empirical research and to inform experts about certain online tools that are available and ready for further use – the corpus of religious texts and the corpus of essayistic texts.

2. The Usage of Corpora in Interdisciplinary Research

Extensive text corpora (such as prim-7.0-public-all [1]) are compiled using various types and genres in order to make them representative, i.e. they should present the linguistic system in a form that is actually used in practice. Each of the three main types currently distinguished in prim-7.0-public-all (literary texts, journalistic texts, professional texts) encompasses further subtypes and genres. In addition to standard journalistic texts (primarily articles from newspapers and magazines), the category of journalistic texts in the SNC (Slovak National Corpus) database also includes administrative texts and texts from certain websites and blogs. Professional literature includes scientific monographs, textbooks, course literature, articles from popular-science magazines and specialized magazines as well as encyclopaedia entries. In addition to poems, short stories and novels, literary texts can also include other works – for instance a book of essays or interviews with significant figures from cultural and social life.

The extent of the prim-7.0-public-all corpus is currently at approximately 1.25 billion tokens, which is why the focus now lies on the thematic diversity of the linguistic material and the database is being completed with texts from previously underrepresented areas of social sciences as well as natural sciences. This is due to the fact that in certain cases even a very small or specific area, whether a scientific discipline or a genre of literature, can be noteworthy.

The way specialized corpora are built is based on the unique needs of specific research, most often the needs of linguistics and terminology, but even other scientific areas as well. Their extent is significantly smaller than that of general corpora. An example of such a corpus is the Corpus of Religious Texts blf-2.0 [2], which

* **Marek Debnar**
Department of Journalism, Faculty of Arts, Constantine the Philosopher University, Nitra, Slovakia
E-mail: mdebnar@ukf.sk

contains 16 million tokens, or the Essay Corpus ess-1.0-all, with more than 5 million tokens.

3. The Issue of Genre and Sources

The vastness of the essay genre and its penetration into several areas of social science make any attempts at a clear definition of the essay significantly more difficult. Efforts to define the genre more accurately are also complicated by the confusion which is caused by the fact that the term "essay" is used in different meanings based on the tradition that each instance of the word refers to. This is why the first step of this research is excluding those texts that are typically referred to as *essays* but can be classified as schoolwork. Afterwards, the perspective of the European tradition is assumed, which means that the term *essay* is only used to refer to texts with certain literary qualities.

As early as during the preparatory stages of this paper (which consisted of necessary technical procedures), it became clear that simply excluding schoolwork and theses from the set of essays would not suffice. Despite the fact that the essayistic texts from the Slovak National Corpus (SNC) never included schoolwork, the list of texts had several type variations. Although in most cases essays were categorized as literary texts, e.g. *Etela Farkasova: Uvidiet hudbu* a *ine eseje – "Seeing Music and Other Essays"*), we have also found cases where texts classified as essays were assigned to the category of either journalistic texts (e.g. *Lubomir Durovic: Europeizacia Balkanu, nie balkanizacia Europy – "The Europeanization of the Balkans, not the Balkanization of Europe"*) or professional texts (*Bruno Latour: Nikdy sme neboli moderni – esej* o *symetrickej antropologii – "We Have Never Been Modern – an Essay on Symmetrical Anthropology")*. Shorter professional texts published in journals or collective monographs dealing with the essay genre that simultaneously tried to bring about a feeling of essayistic writing were no exception. Examples of these are *Milos Horvath: Spory* o *esej* a *jej miesto* v *slovenskej kulture* a *kulturnej historii* (*"Disputes over the Essay and Its Place in Slovak Culture and Cultural History"*) and others.

These texts had been collected in the corpus since 2005 and annotated by different annotators. Since the aim of the study was to determine the set of essayistic texts as accurately as possible, a decision was made to go through all the annotations again and check them. It turned out that the aforementioned text by Lubomir Durovic, which was published in the *Pravda* daily, was in fact not an essay (as its original classification indicated) but rather a feuilleton and the book by Bruno Latour (although its title explicitly states it is an essay) is not an essay but rather a scientific monograph. The same applied to the text by Milos Horvath – although it deals with the topic of the essay, the text itself is a study.

After a thorough inspection of all the texts, which was the first step towards the creation of an essay subcorpus (Genre: ess), only half of the original two hundred entries in SNC that were originally categorized as essays remained. These approximately one hundred texts, most of which are books (as can be seen from the Medium statistical analysis), were then used to create the first model of the essay subcorpus.

A change which meant a significant expansion of the corpus came with the acquisition of rights to the essays published by the Forum project on the website www.salon.eu.sk. Out of the originally more than 700 texts, over 400 were included in the corpus. The largest part of the excluded texts consisted of texts published in Czech. Problems were also caused by occasional errors in encoding, which caused issues during text conversion. In the final stage of the first version of the ess-1.0-all essay subcorpus, which was created in July 2016, the subcorpus now contains approximately 500 texts (books and texts published online) from the essay genre. The percentage is as follows:

The **Medium** key [3] for the essay genre:

lib (book)	88.16%
net (texts found online)	10.83%
ebk (e-book)	0.62%
jou (magazine)	0.36%
npu (unpublished texts, manuscripts)	0.03%

In addition to texts written in Slovak, we have also included Slovak translations of foreign-language essays, e.g. Virginia Woolfe: *Vlastna izba* (*A Room of One's Own*), some of T.S. Eliot's *Essays* as well as other authors (the proportion of foreign-language texts to Slovak ones is indicated by the Origlang statistical analysis).

The **Origlang** key [4] for the essay genre:

slk (Slovak)	58.62%
eng (English)	21.05%
hun (Hungarian)	4.33%
cze (Czech)	3.42%
rus (Russian)	3.25%
nld (Dutch)	1.91%
ger (German)	1.47%
fra (French)	1.36%

In regard to the gender of the authors whose texts were included in the essay subcorpus, 90.44% were men, 7.79% women and the rest were mixed collectives of authors.

4. The Scoring Method of the ess-1.0-all Subcorpus

To enable the use of essayistic texts for further analysis, a set of unique type-genre characteristics has been assigned to texts classified as essays in the SNC, similarly to the way novels or

monographs are tagged. Before proceeding to a closer description of the type-genre specification of the essay according to the system of structural tags used in the SNC, it is necessary to outline the character and method of recording this information.

Recording the basic bibliographic and type-genre metadata is part of primary text processing, which is collectively referred to as type-genre annotation (TGA) and consists of 35 entries that characterize a text. These entries are used to categorize each text into a specific factual, thematic and semantic field, which makes it possible to study the current form of written language and indirectly study linguistic meaning in a constantly growing database of texts.

TGA is performed by assigning fixed values, which form the core of type-genre annotations of texts, to the subkeys of three basic keys: the *domain* (non-fiction or social science), *genre* and *type* of text. During annotation, the annotator has the understanding that the text itself carries a certain intention that it is created with. A dissertation thesis is written with a different intention, a different type of language and for different readers than a newspaper report, a statute book or a collection of poems. The same applies to essays.

The interconnection between the *genre* and the *type/subtype* of a text also begs the question of genre classification and transitional types in individual sets of texts. In other words, there are different variations of the *type* or *subtype* values for each value of *genre* and vice versa. But even the *genre* itself is divided into *subgenres* – e.g. the novel or the short story are divided into the following subgenres: prose, children's literature, detective fiction, travelogue, sci-fi, fantasy, non-fiction, women's novel, adventure novel, etc. These variations are determined for each text individually and are the result of a thorough research of the text by an annotator. While variations of the *subtype* and *subgenre* within one genre are rather common, variations of the *type* within one *genre* are much rarer. It is this transitive character that has until recently been most characteristic of the set of texts labelled as essays in the corpus.

When re-checking the texts labelled as essays, the goal was accurately determining the individual sets that had transitional values of *type* for this genre. Just like in the cases given above, other cases also revealed that this set can be unified if we apply stricter rules for genre classification on these texts. The result was a certain unification within the essay genre, which no longer has variations in *type* (which is now stabilized as *literary text*) but rather in its *domain* and *subdomain* (the areas of professional literature and non-fiction).

5. An Outline of Semantic Fields in ess-1.0-all

In order to improve the thematic structure and categorization of linguistic devices, text annotations in the SNC database include keys which delimit the semantic field of each annotated text by indicating the non-fiction and social-science *domain* and *subdomain* that the text is related to. The *domain* is semantically superordinate to the text, which manifests itself in the *type* and the choice of characteristic linguistic devices, i.e. *genre* of the text. From the perspective of the thematic categorization of all texts labelled as essays, the following statistical analysis has arisen in the final version of ess-1.0-all:

The **Domain** key [5] for the essay genre:

ars (art)	52.21%
hum (the humanities)	17.37%
blf (faith, the supernatural)	16.72%
ins (interdisciplinary sciences)	3.90%
MIX (mixed value)	3.47%
nat (natural sciences)	3.38%
lif (lifestyle)	2.37%
ecn (economics, management)	0.44%
tec (technology)	0.10%

Individual *domain* areas are then further categorized in the *subdomain* key. The *subdomain* contains 53 required values which further represent specific thematic areas of the superordinate set of *domain*, e.g. *art* is further divided in the subdomain key into: literature (literary science and criticism), theatre (theatre studies and criticism), architecture, film, music (opera and ballet) and visual arts (which also include photography and sculpture). Other areas of *domain* have similar subdivisions, which is especially true for the areas of the humanities, natural sciences, law, technology, economics, industry, faith and lifestyle.

The following statistical analysis indicates the percentage of these areas in the essay subcorpus:

The **Subdomain** key [6] for the essay genre:

lit (literature, literary science and criticism)	50.95%
rel (religion, faith, cults)	16.72%
pol (political science)	8.72%
YYY (undefinable value)	7.40%
phi (philosophy, ethics, aesthetics)	4.32%
eth (ethnology, ethnography)	2.54%
env (environmental science, ecology)	2.53%
sct (social life)	1.86%
mus (music, opera, ballet)	0.88%
his (history)	0.82%
bio (biology)	0.70%
eco (economics, banking)	0.41%

Other subdomains only have trace values or mixed values. The undefinable value YYY always appears if the superordinate set of keys contains no more further specified subsets, such as the domain ins (interdisciplinary sciences).

6. The blf-2.0 and ess-1.0.-all Subcorpora

A specialized corpus of texts from the thematic area of religious texts, faith and the supernatural labeled blf-2.0 was made available in 2014 with the extent of almost 66 million tokens (i.e. it is ten times more extensive than ess-1.0-all). The subcorpus was created in the SNC primarily for the purposes of research in the area of religious terminology, which had been pushed aside before 1989 due to political and ideological reasons [7]. The subject of religious terminology is closely connected to the subject of interpreting the concepts of certain representatives of religious philosophy [8], social philosophy [9] and religious studies [10]. However, this exceptionally interesting connection is yet to receive academic treatment.

The treatment of the blf-2.0 and ess-1.0-all subcorpora differs in one significant aspect – as has been demonstrated, the ess-1.0-all subcorpus is built on the *genre* key, which enables a rather rich semantic diversity that manifests itself in the variations of the *domain* and *subdomain* values. The blf-2.0 subcorpus is built on the *domain* key and thus it only defines the semantic field of the texts for three *subdomains*: rel (religion, faith, cults), teo (theology), exc (the supernatural, occultism, magic, astrology). However, this limitation enables a great variability of types and genres – from religious poetry, hymnbooks and religious children's books through sermons to scientific papers on these topics.

Therefore, a comparison of the semantic fields of these two subcorpora using *domain* and *subdomain* would not yield any relevant results, with the exception of the aforementioned penetration of essays into the area of religious texts, which is a 16.72% correspondence at the level of *domain.* Thus a decision was made to compare the quantitative occurrence of the most frequently occurring lexical variants.

Let us present the ten most frequently occurring words in the blf-2.0 subcorpus (excluding conjunctions and prepositions):

- *clovek* (man) 238,808 (occurrences)
- *boh* (god) 213,174
- *zivot* (life) 198,093
- rok (year) 183,392
- cirkev (the church) 144,736
- Jezis (Jesus) 133,717
- den (day) 99,755
- laska (love) 97,376
- otec (father) 93,531
- slovo (word) 93,267

The ten most frequently occurring words in the ess-1-0-all subcorpus (excluding conjunctions and prepositions):

- clovek (man) 17,222 (occurrences)
- zivot (life) 10,440
- svet (world) 8,998
- boh (god) 8,408
- cas (time) 6,621
- povedat (say/tell) 6,065
- vediet (know) 6,056
- slovo (word) 5,703
- sam (alone) 5,569
- kniha (book) 4,504

As can be seen from the top positions of the frequencies of occurrence, both subcorpora are mutually interconnected and indicate the thematic character of essayistic and religious writing in that both of them try to grasp humans in their entirety. Of course, a deeper interpretation would require a more detailed statistical analysis, such as the frequencies of occurrence for the most common multi-word phrases, etc. However, the aforementioned data is sufficient in order to demonstrate the basic possibilities that computer processing of a certain set of texts makes possible.

Both specialized subcorpora are publically available on the website htttps://bonito.korpus.sk and can be searched using the NoSketch Engine [11] statistical and search tool. Further processing of selected sets of texts using other corpus tools is also possible.

7. Conclusion

Having illustrated the way texts are annotated using the model example – a type-genre description of texts labeled as essays in the SNC – the next goal is documenting further ways how general and specialized corpora can be used to analyze larger texts. Eventual similarities and differences between these sets could shed even more light onto the subject of using language and terminology in individual areas of the humanities. However, research that would take into account all possibilities of using this potential and would also be able to present a high-quality quantitative and semantic analysis of text genres is yet to be conducted.

Acknowledgement

This submission was created as part of a research assignment for the VEGA 2/0107/14 grant "Hypermedia Artefacts in a Post-digital Era".

References

[1] Searches in the main corpus as well as the subcorpora of the Slovak National Corpus can be performed via https://bonito.korpus.sk.

[2] http://korpus.sk/blf.html

[3] **Medium** is a key used to determine the so-called textual medium or source. Values such as *book*, *magazine* or *newspaper* are supplemented by further values such as *course literature*, *journals*, *internet* and *unpublished* or *one-off texts*. In 2014, the *e-book* value was added.

[4] The **Origlang** key indicates the language in which the original text was written.

[5] **Domain** is a key which determines the domain (thematic area of a certain activity or knowledge) and which is connected to the largest fixed set of values. By correctly determining the domain, i.e. by correctly assigning the text to a specific thematic area, the semantic field of the expression used in a search is efficiently narrowed down.

[6] **Subdomain** is a key used to further categorize the superordinate thematic area of a certain activity or knowledge (domain). It consists of a fixed set of 53 values.

[7] More on this topic can be found in the following article: SIMKOVA, M.: *Religious Texts and Terminology in the Slovak National Corpus* (in Slovak). In: Jan Durica (ed.): *Brief Dictionary of Catholic Theology* (in Slovak), Proc. of intern. conference, Trnava: Dobra kniha, 2015, 77-98. ISBN 978-80-7141-955-6

[8] KRALIK, R., PAVLIKOVA, M.: *The Reception of Kierkegaard's Work in Slovakia* (in Slovak), *Filozofia*, vol. 68, No. 1, 2013, 82-88. ISSN 0046-385X; KRALIK, R.: *The Reception of Soren Kierkegaard in Czech Language* (in Czech), *Filozoficky casopis*, vol. 61, No. 3, 2013, 443-351. ISSN 0015-1831.

[9] JUROVA, J.: On Etzioni's Concept of a Responsive Community. *European J. of Science and Theology,* vol. 12, No. 3, 2016, 101-111. ISSN 1841-0464.

[10] KONDRLA, M., PAVLIKOVA, M.: From Formal Ethics to Existential Ethics. *European J. of Science and Theology,* vol. 12, No. 3, 2016, 101-111. ISSN 1841-0464; KONDRLA, P., KRALIK, R.: The Specifics of Mission of the Thessalonians Brothers and the Potential for they Actualization (in Slovak), *Konstantinove listy*, vol. 9, No. 2, 93-99. ISSN 1337-8740; VALCOVA, K., PAVLIKOVA, M., ROUBALOVA, M.: Religious Existentialism as a Countermeasure to Moralistic Therapeutic Deism. *Communications - Scientific Letters of the University of Zilina*, vol. No. 3, 2016, 98-104. ISSN 1335-4205.

[11] *https://nlp.fi.muni.cz/trac/noske.*