Miroslav KUBÁT
Radek ČECH
University of Ostrava

DOI: 10.14746/bo.2024.1.1

Attributivity and Subjectivity in Czech Journalism and Scientific Literature[1]

Keywords: attributivity, subjectivity, stylometry, syntax, journalism, scientific literature

Abstract

This study examines two recently proposed stylometric indices, attributivity and subjectivity, in the context of Czech journalism and scientific literature. Index of attributivity is defined as the ratio of the frequency of attributes to the sum of frequencies of nouns, pronouns, and attributes. Index of subjectivity is defined as the ratio of the frequency of subjects to the sum of frequencies of predicates and subjects. The research is based on the corpus of contemporary written Czech language SYN2020. It appears that both indices of attributivity and subjectivity are sensitive to different genres, and could be used in stylometric analyses, especially those that focus on the interpretability of quantitative measures. In general, more formal texts tend to have higher values of both indices across different genres.

1. Introduction

Different styles, genres, and other types of texts can be analyzed qualitatively as well as quantitatively. Empirical quantitative approach refers to one of modern linguistic fields – stylometry. Stylometry deals with text classification using statistics. Although the methodology is primarily focused on authorship attribution, it can also be applied to various text types such as styles and genres. The number of such studies is increasing in contemporary linguistics as a result of several

[1] The research was supported by the Czech Science Foundation (GAČR), project No. 22-20632S.

factors, including easy access to data (language corpora) and tools (software), as well as a general emphasis on empirical research.

Stylometric methods can generally be divided into two groups. The first consists of simple measurements such as average sentence/word lengths, word frequencies, etc., as well as indices of stylistic characteristics such as lexical diversity. Therefore, their advantage lies in their simplicity and straightforward linguistic interpretation. This approach is useful when we want to understand more some text features. The second stylometric direction is primarily concerned with automatic text classification with high accuracy. The research is based on advanced computational methods such as multidimensional analysis (MDA) and neural networks. Moreover, non-linguistic units such as n-grams are commonly used in this approach. The advantage lies in the high accuracy in automatic text classification. On the other hand, the linguistic interpretation is very difficult or impossible because of their black box nature. The analysis we conduct belongs to the more traditional approach based on simple and interpretable methods.

This study deals with syntactic characteristics of Czech journalism and scientific literature using two recently proposed stylometric indices called attributivity and subjectivity (Kubát et al., 2021). Index of attributivity is defined as the ratio of the frequency of attributes to the sum of frequencies of nouns, pronouns, and attributes. Index of subjectivity is defined as the ratio of the frequency of subjects to the sum of frequencies of predicates and subjects. The previous research has shown promising results in a classification of main text type groups such as fiction, administrative texts, scientific literature, etc. (Kubát et al. 2021). In this study, we apply the indices to a finer genre analysis of journalism and scientific literature. The first goal of this study is to discover whether attributivity and subjectivity are suitable for a genre analysis. The second goal is to enrich stylistics with new quantitative syntactic findings.

The material comes from Czech National Corpus – namely the corpus SYN2020 (Křen et al. 2020). We focus on differences inside journalism (NMG) according to different categories defined by the

SYN2020 annotation: a) medium (newspapers and magazines); b) genre (lifestyle, sport, etc.). Scientific books (SCI) will be categorized according to a) genre_group (e.g. social and natural sciences); b) genre (e.g. law, psychology, sociology).

2. Material

In this study, the data was collected from the Czech National Corpus, namely the representative corpus of contemporary written Czech SYN2020 (Křen et al., 2020). It covers texts across different styles and genres published in 2015–2019 period and consists of 120 million tokens. Various text types are proportionally distributed in the dataset. The corpus is divided into three main groups FIC: fiction, NFC: non-fiction, and NMG: journalism where each one covers 33.33% of the corpus. These three main groups are then structured into subgroups. In this study, we work with scientific texts and journalism.

The corpus is lemmatized and both morphologically and syntactically annotated. It is important to mention that SYN2020 is one of the few large representative corpora with syntactic annotation. This annotation is based on the annotation of the Prague Dependency Treebank (Jelínek et al., 2021; Hajič et al. 2020). It marks dependency relations between pairs of words in a sentence and assigns the analytical functions to individual words. The accuracy rates of SYN2020: UAS (unlabeled attachment score) = 92.39%, LAS (labeled attachment score) = 88.73%. According to the authors of the corpus, the error rate is higher for less common syntactic functions, whereas the most frequent functions in expected contexts have an error rate lower than 5% (https://wiki.korpus.cz/doku.php/cnk:syn2020:automaticka_anotace). Despite some errors, this accuracy (especially in such a big corpus) is outstanding in whole contemporary corpus linguistics. As a result, we consider the error rate to be acceptable for our research. Stylometric genre-oriented comparative analysis based on big data has been extremely difficult because of lacking sufficient language material. There are only a few corpora across languages providing comparable accuracy of syntactic annotation. In this context, it is noteworthy to

mention the multilingual project Universal Dependencies developing advanced treebanks across languages (https://universaldependencies.org/). However, these corpora usually consist of a low variety of different genres. Consequently, the release of the corpus SYN2020 was a significant motivation for our pioneering analysis.

Journalism (NMG) is represented by texts from newspapers and magazines covering various topics. In total, the journalism sub-corpus consists of about 40 million tokens. In this study, we will focus on the difference between the published media (NWS: newspapers; J: journal), on the one hand, and among 8 categories representing different topics (NTW: nationwide newspapers; REG: regional newspapers; HOU: home, garden, hobbies; LIF: lifestyle; SCT: social life; SPO: sports; INT: curiosities; MIX: society), on the other.

Scientific texts (SCI) include various academic publications such as books, journals, university textbooks or reference handbooks. We decided to restrict the corpus to scientific books in order to have more consistent data for a fine genre analysis. The size of the dataset we used is more than 8 million tokens. In the analysis, we will measure and compare subjectivity and attributivity in two categories. First, five main academic domains will be analyzed (HUM: humanities; SSC: social sciences; NAT: natural sciences; FTS: formal and technical sciences; ITD: interdisciplinary). Then will focus on even finer level including 22 scientific branches (ICT: information technology; BIO: biology; CHE: chemistry; ART: art, architecture; EDU: education; MUS: music; MED: medicine; ECO: economy, business, logistics; LAW: law; INF: library and information science; POL: politics, military; THE: theatre, film, dance; ITD: interdisciplinary; PHY: physics; REC: sports, recreation, hobbies; SOC: sociology; LAN: philology; HIS: history, biography; ANT: anthropology, ethnography; PSY: psychology; PHI: philosophy, religion; MAT: mathematics).

A detailed description of the corpus SYN2020 can be found on the Czech National Corpus website https://wiki.korpus.cz/doku.php/en:cnk:syn2020.

3. Methodology

The methodology of this research is based on two recently proposed stylometric indices: attributivity and subjectivity. These two indices are simple ratios that express style characteristics that can be interpreted easily. This approach is inspired by similar indices successfully applied in stylometry such as nominality, activity, descrtiptivity (cf. Zörnig 2015). These ratios have proven very useful in stylometry (cf. Chen and Kubát, 2022; Melka and Místecký, 2020). Contrary to the indices based on the morphological level (part of speech), attributivity and subjectivity reflect writing style in terms of syntactic functions.

3.1. Attributivity

The meaning of nouns and pronouns can be modified by an attribute (modifier). Attribute is usually realized by adjective (e.g. *zelené* oči) [*green* eyes]. However other parts-of-speech can be attributes as well: pronoun (e.g. *jeho* kolo) [*his* bike], numeral (e.g. *první* den) [*first* day], noun (úpravy *textu*) [*text* correction]), nonfinite verb (přání *zdokonalit*) [desire *to improve*]), or adverb (cesta *domů*) [way *home*]. Attribute can be also realized by a dependent clause (Vidím ženu, *která zpívá*) [I see a woman *who sings*].

Generally speaking, attributes are the most common way of expressing more detailed description. The index of attributivity is defined as the ratio of the frequency of attributes to the sum of frequencies of nouns, pronouns, and attributes. The more detailed description of things, the higher the index of attributivity.

$$\text{attributivity} = \frac{\text{attributes}}{\text{nouns} + \text{pronouns} + \text{attributes}}$$

3.2. Subjectivity

Subjects are typically realized by noun or pronoun. Since Czech is one of the highly inflected languages, the ending morpheme of the

predicate can be used to identify the person. Therefore, the subject may be omitted, but the predicate must be present in every clause. The index of subjectivity is defined as the ratio of the frequency of subjects to the sum of frequencies of predicates and subjects.

$$\text{subjektivity} = \frac{\text{subjekts}}{\text{predikates} + \text{subjekts}}$$

3.3. CQL queries

Following CQL (corpus query language) queries were used for searching predicates, attributes, subjects, nouns, and pronouns in the corpus SYN2020.

Predicates: [tag="V[B,i,p,q,s,t].*"&afun!="AuxV|AuxT|AuxR"]

Attributes: [afun="Atr" | afun="Atr_Co" | afun="Atr_Ap" | afun="Atr_Pa" | afun="AtrAdv" | afun="AtrAdv_Co" | afun="AtrAdv_Ap" | afun="AtrAdv_Pa" | afun="AtrAtr" | afun="AtrAtr_Co" | afun="AtrAtr_Ap" | afun="AtrAtr_Pa" | afun="AtrObj" | afun="AtrObj_Co" | afun="AtrObj_Ap" | afun="AtrObj_Pa" | afun="AtrAdv" | afun="AtrAdv_Co" | afun="AtrAdv_Ap" | afun="AtrAdv_Pa"]

Subjects: [afun="Sb" | afun="Sb_Co" | afun="Sb_Ap" | afun="Sb_Pa"]

Nouns: [tag="N.*"]

Pronouns: [tag="P.*"]

4. Results

4.1. Attributivity in journalism

According to the resulting values in journalism (see Figs. 1 and 2), the attributivity appears to be sensitive to the different topics of articles rather than to the difference between newspapers and journals (magazines). The lowest value is assigned to sport articles since they provide a summary of the basic facts and results of sports events. Con-

trary to this, texts concerning home, garden, and hobbies have the highest attributivity, which can be explained by their natural need for detailed description. They usually depict some objects connected to the house and garden architecture and design followed by photographs and illustrations. Interestingly, social and lifestyle magazines have lower attributivity than news.

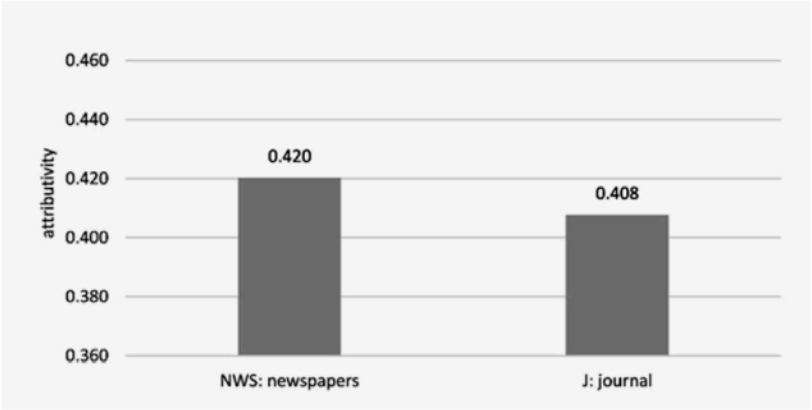

Fig. 1 Attributivity in journalism – media types.

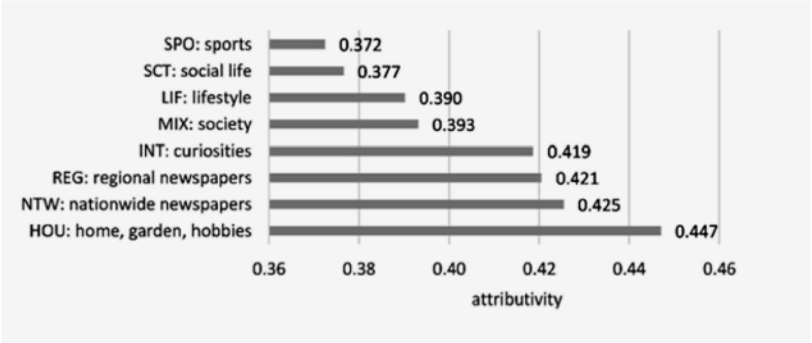

Fig. 2 Attributivity in journalism – individual topics.

4.2. Attributivity in scientific literature

As can be seen in Fig. 3, there is a clear difference between humanities and social sciences on the one side, and natural and technical sciences on the other side in terms of attributivity. This is probably caused by the fact that natural and technical fields tend to be more explicit and precise in the description in general. Interdisciplinary research is just between them. Generally, this observation is confirmed in the finer analysis focused on individual academic fields, with some variations (see Fig. 4). Hard sciences such as information technology, biology or chemistry are on the top while humanities like philosophy, history or philology are less attributive. Some disciplines, however, differ considerably from this general pattern. Mathematics is the most notable outlier, with the lowest value of attributivity. Mathematics is mainly represented by a book that discusses the historical and philosophical background of set theory and its founder Bernard Bolzano. So, rather than being a hard science book, the style of this book belongs to the humanities. In light of this finding, it is evident that attributivity is indeed an important text feature suitable for stylometry.

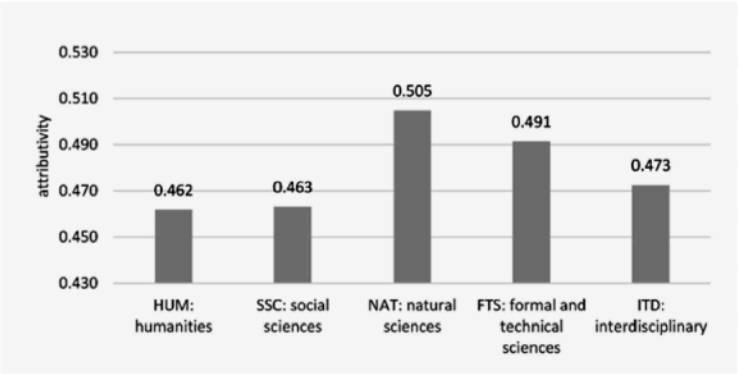

Fig. 3 Attributivity in scientific literature – field groups.

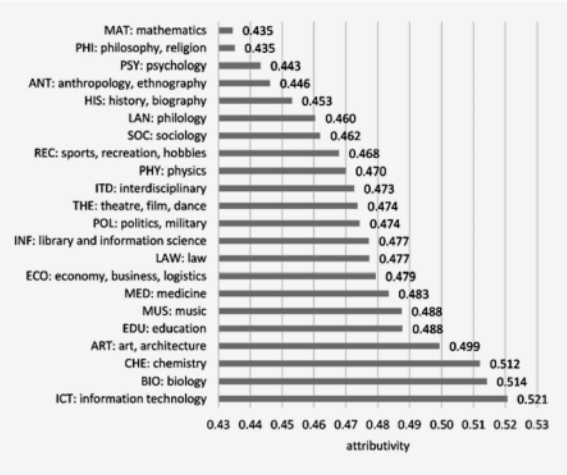

Fig. 4 Attributivity in scientific literature – individual fields.

4.3. Subjectivity in journalism

The resulting values of subjectivity in Fig. 5 show quite a clear difference between newspapers and magazines. That is also confirmed by the data in Fig. 6. Writing style of news is rather formal. The purpose of news is to present facts objectively. Language of lifestyle articles and leisure magazines in general is more informal and creative. Due to the fact that subject is frequently omitted (especially in the first and second person) in informal texts in Czech, it appears natural for news to have higher subjectivity.

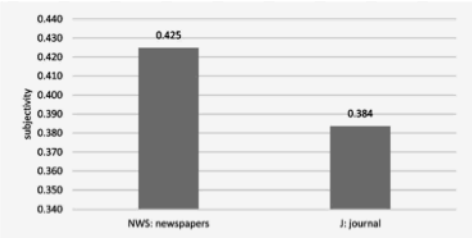

Fig. 5 Subjectivity in journalism – media types.

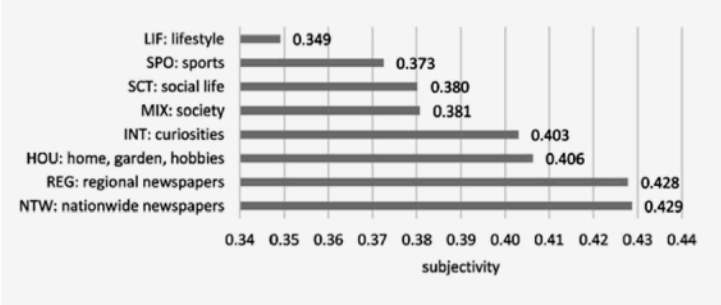

Fig. 6 Subjectivity in journalism – individual topics.

4.4. Subjectivity in scientific literature

According to data in Figs. 7 and 8, hard sciences tend to have higher subjectivity which can be explained by their more formal nature compared to the humanities and social sciences. However, the results vary quite a lot.

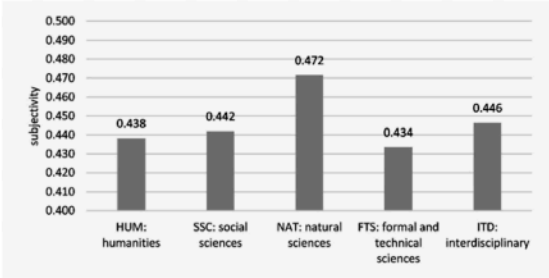

Fig. 7 Subjectivity in scientific literature – fields groups.

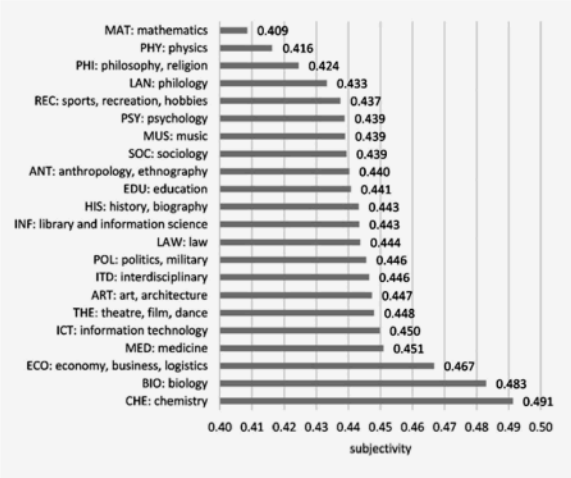

Fig. 8 Subjectivity in scientific literature – individual fields.

5. Conclusion

The obtained data bring several findings. Both attributivity and subjectivity seem to be sensitive to different genres. The simplicity and ease of interpretation of these indices make them suitable for a wide range of stylistic studies.

We can observe in both journalism and scientific literature that texts which require more explicit and precise descriptions (news, natural science and social science) tend to be more attributive. In terms of subjectivity, more formal texts like news have higher subjectivity than less formal texts like magazine articles dealing with leisure topics. It is also worth noting that these findings correspond to the overall comparison in which scientific literature has higher values (attributivity = 0.468, subjectivity = 0.442) than journalism (attributivity = 0.415, subjectivity = 0.408). In general, the results of the study are consistent with previous research showing higher values of attributivity and subjectivity in more formal texts (Kubát et al., 2021).

Since our study shows promising results of application attributivity and subjectivity in stylometry, it would be interesting to see future research based on these methods in different languages and different text types such as different authorships. This approach seems to have great potential in research where a straightforward interpretation is needed (e.g. literary studies).

References

Chen X., Kubát M., 2022, Rural versus urban fiction in contemporary Chinese literature-Quantitative approach case study. *Digital Scholarship in the Humanities*, 37(3): 681–692.

Hajič J., Bejček E., Hlaváčová J., Mikulová M., Straka M., Štěpánek J., Štěpánková B., 2020, *Prague Dependency Treebank – Consolidated 1.0.* In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*. Marseille, pp. 5208–5218.

Jelínek T., Křivan J., Petkevič V., Skoumalová H., Šindlerová J., 2021, SYN2020: A new corpus of Czech with an innovated annotation. In K. Ekštein, F. Pártl, M. Konopík (eds.), *Text, Speech, and Dialogue. TSD 2021.* Cham: Springer, pp. 48–59.

Křen M., Cvrček M., Henyš J., Hnátková M., Jelínek T., Kocek J., Kováříková D., Křivan J., Milička J., Petkevič V., Procházka P., Skoumalová H., Šindlerová J., Škrabal M., 2020, *SYN2020: reprezentativní korpus psané češtiny*. Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague, Czechia. Available at http://www.korpus.cz.

Kubát M, Čech R., Chen X., 2021, Attributivity and Subjectivity in Contemporary Written Czech. In X. Chen, R. Čech (eds.), *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021).* Sofia: Association for Computational Linguistics, pp. 58–64.

Melka T. S., Místecký M., 2020, On stylometric features of H. Beam Piper's Omnilingual. *Journal of Quantitative Linguistics*, 27(3): 204–243.

Zörnig P., 2015, *Descriptiveness, activity and nominality in formalized text sequences*. Lüdenscheid: RAM-Verlag.