Modality of verbs as stylometric feature in Czech genres

Miroslav Kubát[1], Xinying Chen[1*]

[1] University of Ostrava
[*] Corresponding author's email: cici13306@gmail.com

DOI: https://doi.org/10.53482/2023_55_413

ABSTRACT

The goal of the study is to analyze modality of verbs from the perspective of its usage in different types of Czech texts. The analysis is based on data from the Czech National Corpus, specifically the balanced corpus of contemporary written Czech SYN2020, which contains 100 million words. The proportion of modal verbs to all verbs is used to measure modality. Furthermore, different types of modality are considered: necessity and possibility. The findings reveal distinct patterns in the use of modal verbs in different genres. Thus, index of modality seems to be a promising stylometric feature. Non-fiction literature, especially administrative texts, exhibits the highest modality. In contrast fiction texts, namely poetry, has the lowest modality.

Keywords: syntax, modality, stylometry, Czech.

1 Introduction

Stylometry is a branch of linguistics that focuses on the quantitative analysis of various styles and the identification of distinctive features within texts. Stylometry has numerous applications, including authorship attribution, genre classification, and text clustering, and has become an increasingly popular tool for literary scholars, linguists, and forensic investigators (cf. Holmes 1998; Juola 2007; Savoy 2020). By providing insights into the stylistic characteristics of texts, stylometry can offer a deeper understanding of language use.

There are two main branches of stylometry. The first one is based on simple quantitative indicators such as word frequencies, mean sentence length, or text features like lexical diversity. This approach is more traditional and allows straightforward interpretation of the obtained data, which is important when one wants to understand different styles of writing. In the second approach, the methods usually belong to machine learning algorithms, most recently neural networks (see e.g. Matthews and Merriam 2020; Savoy 2020). Thus, this approach belongs to the black box method category, in which linguistic interpretation is rather difficult or even impossible at this stage.

Our study belongs to the traditional stylometric approach based on simple and straightforward indices. The study deals particularly with one feature of verbs - modality. The methodology is inspired by similar stylometric indicators such as subjectivity, objectivity, descriptivity, activity or nominality (cf. Kubát et al. 2021, Zörnig at al. 2014). These indicators are based on simple ratios expressing text features. Despite their simplicity, they have shown to be useful for distinguishing different genres or authors (see e.g. Chen & Kubát 2022; Kubát et al. 2021; Místecký 2018; Zhou et al. 2022).

In this study, we propose a new index of modality. Modality of text is defined as the ratio of the number modal verbs to the number of all verbs in a text. Furthermore, we also focus on a ratio between modal verbs experessing possibility and necessity. Our goal is to discover how modality varies across different styles and genres in a big balanced corpus of contemporary written Czech SYN2020.

Modal verbs in Czech, as in many languages, exhibit a high degree of variability in their usage patterns. This variability is not random but is closely tied to genre-specific conventions and the communicative purposes of texts. For example, academic writing might favor certain modal verbs to express certainty or probability, while fiction may use them differently to depict character intentions or hypothetical scenarios. Analyzing the frequency and context of these modal verbs can thus provide valuable insights into the stylistic fingerprints of different text types. (cf. Chong et al. 2023; Huschová 2015)

2 Material

The language material comes from a large balanced corpus of contemporary written Czech SYN2020 (Křen et al. 2020) belonging to the series of synchronous corpora developed by Czech National Corpus. SYN2020 covers texts mainly from 2015–2019. The size of the corpus is 100 million words. SYN2020 is divided into three equally sized parts: FIC: fiction, NFC: non-fiction, NMG: newspapers and magazines. These three text-type groups are then divided into subcategories such as novel, poetry, humanities, etc. The text-type structure of SYN2020 can be seen in Table 1. A detailed description can be found on the website of SYN2020 https://wiki.korpus.cz/doku.php/cnk:syn2020.

Table 1: Text-type structure of SYN2020.

Text-Type Group	Text-Type
FIC: fiction	NOV: novels
	COL: short stories
	VER: poetry
	SCR: drama, screenplays
NFC: non-fiction	SCI: scientific literature
	PRO: professional literature
	POP: popular literature
	MEM: memoirs and autobiographies
	ADM: administrative
NMG: newspapers and magazines	NEW: news
	LEI: leisure magazines

SYN2020 is a syntactically annotated corpus, using a parser from the NeuroNLP2 toolkit trained on data from the Prague Dependency Treebank (Bejček et al. 2012) and the FicTree corpus (Jelínek 2017). It marks dependency relations between words and assigns syntactic functions. The corpus achieves high accuracy rates of 92.39% for UAS (unlabeled attachment score) and 88.73% for LAS (labeled attachment score). While errors are more common in less frequent syntactic functions, the most frequent functions have an error rate of less than 5% (https://wiki.korpus.cz/doku.php/cnk:syn2020). Despite some errors, SYN2020 is an outstanding syntactically annotated corpus. This is especially due to its size and balanced structure of various types of text.

3 Methodology

Verb modality expresses the speaker's attitude towards the action or state described by a verb. There are three verbs that are considered to be primary modal verbs in Czech language: *muset* [must], *moci* [can], *smět* [may]. These verbs express necessity or possibility (see Table 2). These primary modal verbs are also defined by several syntactic characteristics (see. Karlík & Šimík 2017)[1]. The verb *mít* [to have] can also be used as modal verb in Czech. However, *mít* is mainly used as non-modal verb.[2] That is why *mít* is not counted as basic modal verb in this study.

[1] Karlík and Šimík (2017) define following several syntactic characteristics traditionally attributed to modal verbs:
(a) They only go with infinitives, not with subordinate clauses.
(b) They cannot be expanded with a noun phrase.
(c) They do not form imperatives (commands).
(d) They do not form passive voice.
(e) They do not have aspectual counterparts (different forms showing the completion or duration of the action).
(f) They do not form action nouns or verbal nouns.
(g) It is possible to separately expand the modal verb and the full verb in a sentence.
(h) Both the modal verb and the infinitive can be negated separately.

[2] Based on our small analysis of 300 random occurrences in SYN2020, 67 had a modal meaning. Since *mít* it is mainly used as a non-modal verb and the exisiting SYN2020 annotation framework does not support automatical distinguision between the two, we decided to exclude the verb *mít* from our observation.

Although other verbs expressing desire, intention, or ability (e.g., *chtít* [to want], *potřebovat* [to need], *doufat* [to hope]) can be also considered as modal verbs in a broader sense, we work only with three basic aforementioned modal verbs in this research. The list of analyzed modal verbs can be seen in Table 2.

Table 2: Analyzed modal verbs.

Possibility	Necessity
moci 'can'	*muset* 'must'
smět 'may'	*nemoci* 'cannot'
nemuset 'need not'	*nesmět* 'must not'

We measure the level of modality by a ratio of modal verbs to all verbs[3]:

$$modality = \frac{number\ of\ modal\ verbs}{number\ of\ all\ verbs}$$

4 Results

The results in Figure 1 and Figure 2 show that non-fiction literature (NFC) has a higher modality than fiction (FIC). The newspapers and magazines (NMG) are then positioned in the middle. As for text-types (see Figure 2), administrative texts (ADM) have the highest modality (0.054) among all the analyzed text-types. Poetry (VER) has the lowest value (0.020). Although memoirs and autobiographies (MEM) are in SYN2020 structure assigned to non-fiction literature, these texts have a rather fiction-like modality. That can be explained by the fact that the writing style of memoirs and autobiographies is very close to fiction literature (cf. Soukupová 2015). This genre is therefore naturally somewhere between fiction and non-fiction literature.

This observed phenomenon can be attributed to the intrinsic purposes and contexts inherent in each genre. Non-fiction texts, particularly administrative documents, are largely concerned with providing directives, guidelines, and regulations. These texts need to articulate rules and expectations clearly, often dictating what actions are required, permitted, or prohibited in specific situations. See following examples from the corpus:

- "Souhrn finančních potřeb se vždy musí rovnat souhrnu finančních zdrojů." [The sum of financial needs must always equal the sum of financial resources.]

[3] In SYN2020, for searching all verbs, we use CQL query [tag="V.*"]; for searching modal verbs [lemma = "moci"], [lemma = "muset"], [lemma = "smět"].

- "Hygienické zařízení vyčleněné pro personál smí využívat jen personál." [Sanitary facilities reserved for staff may only be used by staff.]
- "Žadatel nesmí požádat v průběhu roku, ve kterém mu byla poskytnuta dotace z Programu rozvoje venkova na stejný předmět dotace." [The applicant may not apply during a year in which a subsidy from the Rural Development Programme was granted for the same object of subsidy.]

Conversely, fiction texts, which primarily encompass narratives, tend to focus on storytelling rather than instructing or informing. The narrative style of fiction is more about weaving events into a storyline, and less about prescribing behaviors or outcomes. Thus, the relative scarcity of modal verbs in fiction is indicative of a stylistic choice that aligns with the genre's focus on creative expression and character development.

Journalistic texts exhibit a modality that strikes a balance between the definitive nature of non-fiction and the narrative freedom of fiction. This intermediate modality reflects journalism's objective to inform with factual precision while also crafting compelling stories. Modal verbs in journalism navigate between asserting facts and suggesting possibilities, providing a versatile approach to engaging readers with news and narratives.

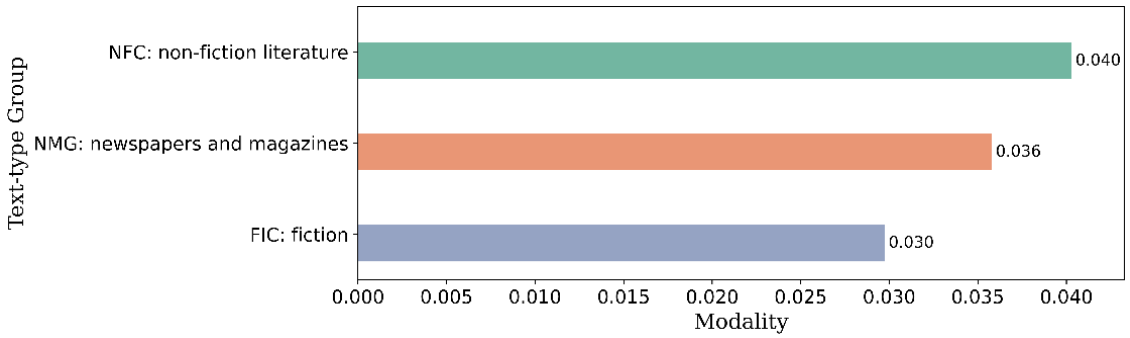

Figure 1: Modality in text-type groups.

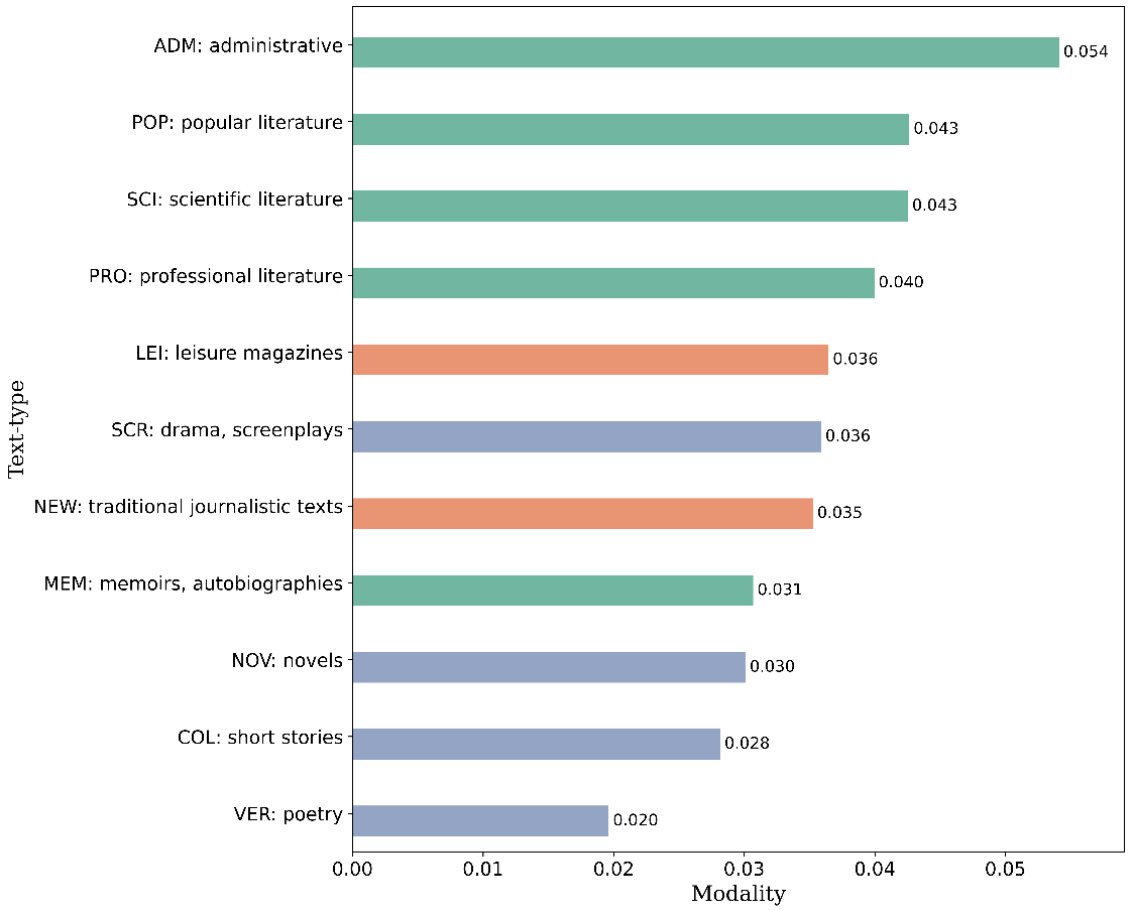

Figure 2: Modality in text-types.

Since the analyzed modal verbs belong to two groups of modality (possibility and necessity), we also focus on their more detailed usage. A percentage representation of modal verbs of possibility and modal verbs of necessity is calculated. In Figure 3, we can see that newspapers and magazines (NMG) and non-fiction texts (NFC) tend to use modal verbs that express possibility rather than necessity. Conversely, fiction (FIC) has more modal verbs of necessity. As shown in Figure 4, the ratio between possibility and necessity in particular text-types is fairly consistent without any clear outliers inside text-type groups. The only exceptions are memoirs and autobiographies (MEM) which again tend towards fiction (FIC) rather than non-fiction literature (NFC). It is interesting to note how consistent the values of fiction (FIC) are. Even so different genres such as fiction, drama, and poetry use modal verbs in similar proportions.

The results suggests that in fiction, there is often a stronger emphasis on situations where characters are compelled to act or are restricted from doing so, reflecting the conflicts and constraints that drive narrative tension. On the other hand, non-fiction texts, particularly scientific literature, show a preference for possibility modal verbs. This could be because scientific writing frequently explores hypotheses, suggests potential explanations, and discusses findings that are not absolute but rather indicative of a

probability. The high use of possibility modal verbs aligns with the tentative and exploratory nature of scientific inquiry.

It is also interesting to note that contemporary poetry, which is more intimate and personal form of fiction, demonstrates the lowest possibility and the highest necessity. This could be due to its focus on the human condition and personal experiences, which are often framed by necessity and constraints.

In general, the preference for necessity modal verbs in fiction can be seen as a reflection of the genre's focus on dramatizing human experiences, while the prevalence of possibility modal verbs in non-fiction, especially scientific literature, underscores a stylistic approach that accommodates the uncertainty and openness inherent in scientific exploration. Tables 3 and 4 present exact frequencies of specific modal verbs, shedding light on their individual roles in expressing possibility and necessity.

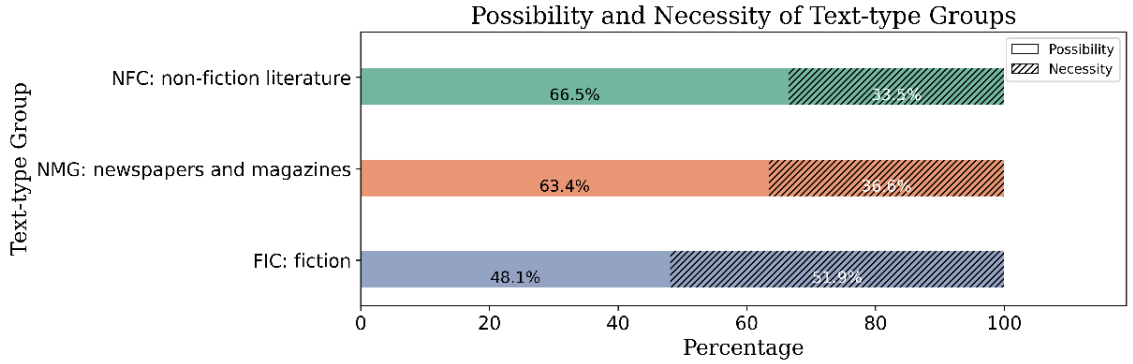

Figure 3: Percent representation of possibility and necessity modal verbs in text-type groups.

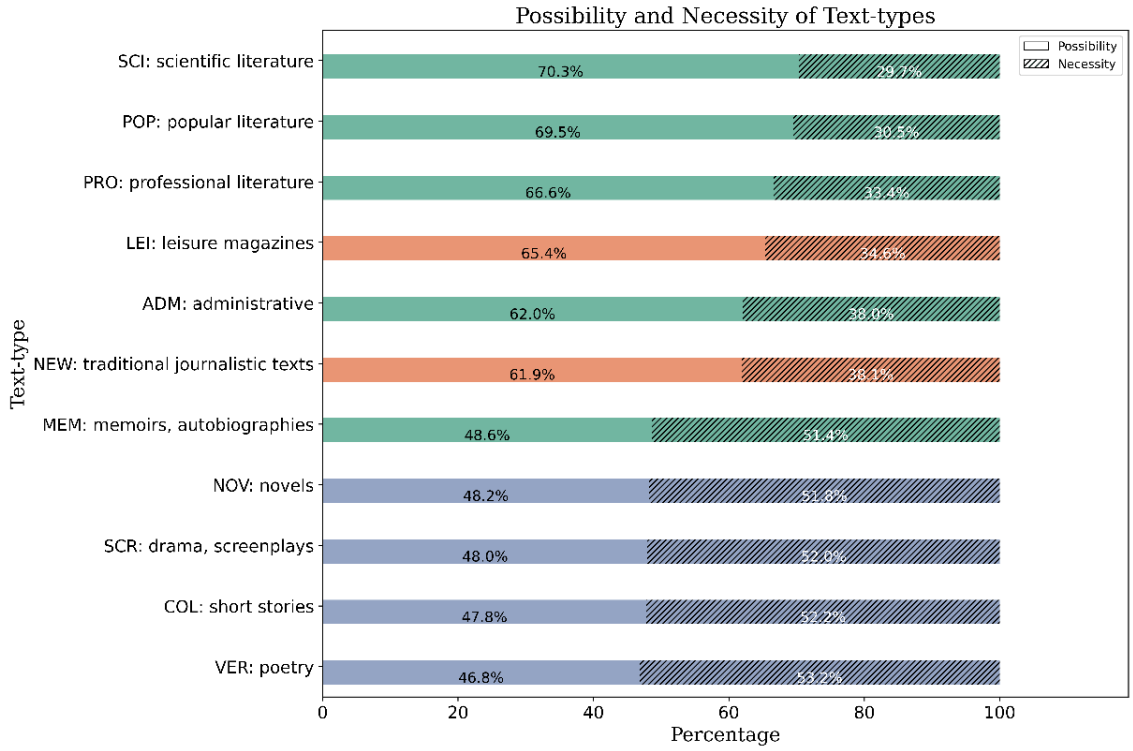

Figure 4: Percent representation of possibility and necessity modal verbs in text-types.

Table 3: Relative frequecies per milion (ipm) of modal verbs in text-type groups.

Text-type group	Possibility			Necessity		
	moci	**smět**	**nemuset**	**muset**	**nemoci**	**nesmět**
FIC: fiction	2886.335	53.65681	262.3288	2125.354	1155.582	174.0779
NFC: non-fiction	3703.843	27.41155	246.4694	1362.105	519.4416	123.6891
NMG: newspapers and magazines	3400.447	20.67973	281.7208	1533.359	482.8113	125.3138

From Table 3, it is evident that journalistic texts use *moci* less than non-fiction but more than fiction, suggesting a balance between expressing capabilities and acknowledging limitations in practical discourse. Fiction, while less frequently employing *moci*, shows a greater use of *muset*, reflecting a narrative emphasis on necessity and inevitability. Journalistic texts demonstrate a moderate usage of *nesmět*, possibly indicating a focus on the boundaries of societal norms and regulations within reported events. These variations in modal verb frequencies underscore the different linguistic strategies employed across genres, offering insightful data for stylometric studies.

Table 4: Relative frequecies pre milion (ipm) of modal verbs in text-types.

Text-type	Possibility			Necessity		
	moci	**smět**	**nemuset**	**muset**	**nemoci**	**nesmět**
NOV: novels	2976.97	50.38	265.08	2193.69	1175.91	170.53
COL: short stories	2598.65	43.17	238.09	1896.15	1072.73	175.11
VER: poetry	1424.51	156.62	172.58	1118.26	663.38	214.47
SCR: drama, screenplays	3527.30	91.72	409.76	2579.18	1561.26	220.33
SCI: scientific literature	3644.69	22.51	181.80	1076.71	447.94	102.86
PRO: professional literature	3273.35	14.12	230.84	1353.94	300.13	112.35
POP: popular literature	4254.46	28.37	307.56	1355.69	543.72	119.27
MEM: memoirs, autobiographies	2767.64	52.35	227.00	2029.83	1003.50	184.08
ADM: administrative	3550.96	100.31	154.76	1653.67	303.79	378.31
NEW: traditional journalistic texts	3208.11	21.09	231.30	1546.24	477.51	110.04
LEI: leisure magazines	3688.83	20.07	357.32	1514.04	490.76	148.22

From Table 4, we can see that in general, *moci* largely shapes the concept of possibility, while *muset* and *nemoci* are key in defining the necessity. This distinction in modal verb usage between non-fiction and fiction genres suggests different stylistic approaches to expressing potentiality and obligation, which is a valuable observation for stylometric analysis. Furthmore, the results indicate that a) fiction texts use *muset* and *nemoci* more heavily, underscoring the themes of obligation and constraint in storytelling. b) Non-fiction texts, particularly scientific literature, rely more on *moci*, highlighting the prevalence of capability and theoretical possibility in academic discourse. c) Administrative texts use *nesmět* extensively, which is aligned with the genre's regulatory nature.

5 Conclusion

The study shows that non-fiction literature, especially administrative texts, reach the highest modality. In contrast, fiction, especially poetry, shows the lowest possibility. Journalistic texts are between them. In terms of usage modal verbs expressing possibility and necessity, fiction tends to use more modal verbs of necessity compared to non-fiction and journalism.

In conclusion, the observed consistency of modality values within text-type groups suggests that modality may offer a reliable and insightful tool for understanding different styles and genres. Index of modality seems to be therefore a useful measure for analyzing a wide variety of texts. It could thus be used in stylometry alongside other similar measures, such as subjectivity, attributivity, nominality, descriptivity, etc.

It is important to note that this is only an initial attempt to employ modality index in stylometry. Our preliminary conclusions need to be validated by further research. Furthermore, this technique may also be used for authorship attribution to determine whether modality is indicative of distinct writing styles among various authors. Since this study is limited to the Czech language, it would be interesting to examine this feature in other languages as well.

Acknowledgements

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

References

Bejček, E., Panevová, J., Popelka, J., Straňák, P., Ševčíková, M., Štěpánek, J., Žabokrtský, Z. (2012). Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In: *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012).* Mumbai, pp. 231-246.

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

Chong, S. T., Ng, Y. J., Karthikeyan, J., Lee, S. Y. (2023). The use of modals in academic discourse: A comparative analysis. In AIP Conference Proceedings (Vol. 2685, No. 1). AIP Publishing.

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3), pp. 111-117.

Huschová, P. (2015). Exploring modal verbs conveying possibility in academic discourse. *Discourse and Interaction*, 8(2), pp. 35-47.

Jelínek, T. (2017). FicTree: a Manually Annotated Treebank of Czech Fiction. In: Hlaváčová, J. (Ed.). *Proceedings of the 17th Conference on Information Technologies - Applications and Theory (ITAT 2017)*, pp. 181-185. Accessible at http://ceur-ws.org/Vol-1885/181.pdf.

Juola, P. (2007). Authorship Attribution. *Foundations and Trends in Information Retrieval*, 1(3), pp. 233-334.

Karlík, P., Šimík, R. (2017). Modální sloveso. In: Karlík, P., Nekula, M., Pleskalová, J. Eeds.). *CzechEncy – Nový encyklopedický slovník češtiny*. Accessible at https://www.czechency.org/slovnik/MOD%C3%81LN%C3%8D%20SLOVESO.

Křen, M., Cvrček, V., 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*. Ústav Českého národního korpusu FF UK, Praha 2020. Accessible at: http://www.korpus.cz.

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

Kubát, M., Mačutek, J., Čech, R. (2021). Communists spoke differently: An analysis of Czechoslovak and Czech annual presidential speeches. *Digital Scholarship in the Humanities*, 36(1), pp. 138-152.

Matthews, R. A., Merriam, T. V. (2020). Distinguishing literary styles using neural networks. In Handbook of neural computation (pp. G8-1). CRC Press.

Místecký, M. (2018). Counting Stylometric Properties of Sonnets: A Case Study of Machar's Letní sonety. *Glottometrics*, 41, pp. 1-12.

Savoy, J. (2020). *Machine learning methods for stylometry. Authorship Attribution and Author Profiling*. Cham: Springer.

Soukupová, K. (2015). Autobiografie: žánr a jeho hranice. *Česká literatura*, 63(1), pp. 49-72.

Zhou, H., Jiang, Y., Wang, L. (2022). Are Daojing and Dejing stylistically independent of each other: A stylometric analysis with activity and descriptivity. *Digital Scholarship in the Humanities*, 38(1), pp. 434-450.

Zörnig, P. (2014). *Descriptiveness, activity and nominality in formalized text sequences*. Lüdenscheid: RAM-Verlag.