Attributivity and Subjectivity in Contemporary Written Czech

Miroslav Kubát[1], Radek Čech[1], Xinying Chen[2]
[1]Department of Czech Language, University of Ostrava
[2]School of Foreign Studies, Xi'an Jiaotong University
miroslav.kubat@gmail.com cechradek@gmail.com cici13306@gmail.com

Abstract

The study focuses on two syntactic indices (attributivity, subjectivity) in various text types and genres in the contemporary written Czech. 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 goal is (a) to find out the utility of the proposed indices in stylometry and (b) to enrich stylistics with new quantitative findings. The research is based on the corpus SYN2020 belonging to the Czech National Corpus. The results show that both indices can distinguish different styles and genres. In general, non-fiction texts tend to have higher values of both indices compared to fiction literature.

1 Introduction

The Czech stylistics is mainly focused on the lexical features of text styles. Phonetic, morphological, and syntactic features are usually rather out of the main interest of scholars (cf. Čechová et al. 2008; Hoffmannová et al. 2016). The exception is Bečka (1992) who paid extraordinary attention to syntax. Czech stylistics is also rather based on qualitative than quantitative approach. Only a few quantitative studies deal with syntactic functions or parts of speech in Czech from a stylistic point of view (e.g. Kubát 2016, Těšitelová 1985, Uhlířová 1974). Since these studies are usually limited to (a) small samples and (b) few analyzed styles or genres, we aim to tackle these issues differently. First, our research is based on a large corpus. Second, we analyze not only the main style groups such as fiction and non-fiction, but we focus also on particular genres such as novels, short stories, etc.

In this study, we focus on two stylometric indices. 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. Proposing these indices is inspired by similar indices successfully applied in stylometry such as nominality, activity, descriptivity (cf. Zörnig 2015).[1] In contrast to these indices based on morphological level (part-of-speech), we focus on writing style in terms of syntactic functions. The goal is to investigate how the resulting values of attributivity and subjectivity vary among different styles and genres in a large corpus. We use data from the Czech National Corpus, namely the corpus SYN2020 consisting of 100 million tokens. We aim to test the utility of these new indices in stylometric research and to enrich Czech stylistics with new findings.

We expect that (a) sentences with higher complexity are generally longer and prefer using more attributes, thus higher attributivity might appear in more formal texts with longer sentences; (b) since Czech has rich morphological features, subjects can be omitted in expressions, therefore, higher subjectivity would also appear in more formal texts which are more syntactically well-formed.

[1] Further research should be done to investigate possible correlations between nominality and descriptivity on the one side and subjectivity and attributivity on the other side. Research of this kind is beyond the scope of this paper.

2 Material

Since we need a big syntactically annotated corpus containing diverse text types and genres of Czech texts, the corpus SYN2020 is used as a dataset for this research. SYN2020 is a synchronous representative and reference corpus of contemporary written Czech containing 100 million tokens (Křen et al. 2020). It is the latest corpus of the representative corpora SYN series (SYN2000, SYN2005, SYN2010, SYN2015, SYN2020), released every five years. Each of the SYN series corpora primarily covers the language of the last five years; thus, SYN2020 consists of the texts published in the 2015–2019 period. Corpus SYN2020 is lemmatized, morphologically tagged, and syntactically annotated.

The syntactic annotation is based on the principles of the annotation used in the Prague Dependency Treebank (cf. Hajič et al. 2020). It marks dependency relations between two words in a sentence and the analytical functions of individual words. The annotation procedure is described in detail on the corpus website[2]. The accuracy rates of SYN2020: UAS = 92.39%, LAS = 88.73%.[3] The error rate is higher for less common syntactic functions and constructions, whereas the most frequent functions in expected contexts have an error rate lower than 5% (cf. corpus website[4]). We have to therefore take into account possible errors when dealing with this dataset. Although the accuracy of syntactic annotation is not perfect, we consider the error rate acceptable for our research.

Since we deal with stylometry in this research, the text style diversity of the corpus is important. SYN2020 consists of texts of various text types and genres. There are three main text type groups (fiction, non-fiction, newspapers and magazines) that consist of several subcategories/genres (see Table 1). The main three groups are equally covered in the corpus.

Fiction (FIC)	
Novels (NOV)	Novels and novellas.
Short stories (COL)	Collections of short stories and other shorter prose texts.
Poetry (VER)	Collections of poetry, marginally song lyrics.
Drama, screenplays (SCR)	Theatre plays, marginally also screenplays for film.
Non-fiction (NFC)	
Scientific (SCI)	Scientific texts, academic publications, university textbooks.
Professional (PRO)	Texts intended for professionals in a given field.
Popular (POP)	Texts intended for a lay audience with an interest in the field.
Memoirs, autobiographies (MEM)	Memoirs, (auto)biographies.
Administrative texts (ADM)	Rules and regulations, meeting minutes, annual reports, etc.
Newspapers and magazines (NMG)	
Newspapers (NEW)	Daily newspapers (current news from home and abroad).
Magazines (LEI)	Special interest magazines focused on thematic groups such as home, garden, hobbies, lifestyle, sports…

Table 1: Text type groups and subcategories/genres in SYN2020.

It is important to mention that SYN2020 is already one of the best quality corpora of such size and style diversity not only for Czech but for all languages. Releasing a large corpus with such a high quality of syntactic annotation and text variety was one of the main motivations for this research.

3 Methodology

We propose two syntactic stylometric indices in this study: attributivity and subjectivity.

[2] https://wiki.korpus.cz/doku.php/cnk:syn2020:automaticka_anotace

[3] UAS (unlabeled attachment score) is the rate of successful parent identification. LAS (labeled attachment score) is the rate of successful identification of both parent and syntactic function.

[4] https://wiki.korpus.cz/doku.php/cnk:syn2020:automaticka_anotace

3.1 Attributivity

Attributivity expresses a magnitude of depicting/describing things in the text. The more detailed description of things, the higher the index of attributivity. The meaning of nouns and pronouns can be modified by attribute (modifier) that provides an extra detail. Attribute is typically realized by adjective (e.g. *nové* auto) [a *new* car] but can be also expressed by pronoun (e.g. *naše* auto) [*our* car], numeral (e.g. *druhé* auto) [a *second* car], noun (úpravy *textu*) [*text* correction]), nonfinite verb (přání *zdokonalit*) [a desire *to improve*]), or adverb (cesta *domů*) [the way *home*]. Attribute can be also realized by a dependent clause. We expect that more formal texts tend to detailed descriptions of things. Thus, an author needs to use more attributes for the description than in less formal texts.

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 formula is as follows.

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

3.2 Subjectivity

Subjectivity indicates a level of expressing subjects. Subject is typically realized by noun or pronoun. In Czech grammar, subject can be omitted whereas a predicate has to be explicitly expressed in every clause. This is caused by the rich morphology of Czech language where the person can be easily identified by the ending morpheme of the predicate, especially for the first and second person. In case of a clause with the predicate in the third person, the subject can be also omitted if it is known from the context. That is why we expect higher subjectivity in more formal texts that tend to explicitly express subjects. On the other hand, less formal texts, especially those close to spoken language, should prefer omitting subjects.

Subjectivity is defined as the ratio of the frequency of subjects to the sum of frequencies of predicates and subjects. The formula is as follows.

$$\text{subjectivity} = \frac{\text{subjects}}{\text{predicates} + \text{subjects}}$$

Both indices are simple ratios expressing style features that can be interpreted straightforwardly. We expect that these features considerably differ in various text groups and genres. It should be noted that the nature of the data (big data, results are not based on average values) prevents us from applying a statistical test.

CQL (corpus query language) queries for searching predicates, attributes, subjects, nouns, and pronouns the corpus SYN2020 used in this research can be found in the Appendix of this paper.

4 Results

4.1 Attributivity

The resulting values show that fiction tends to have much lower attributivity compared to non-fiction and journalism (see Figure 1). This can be explained by the fact that more formal texts need a precise and detailed description of nouns and pronouns. This is also visible in differences between genres inside each text type group. In fiction, drama reaches the lowest attributivity (see Figure 2). Drama is close to spoken language which is generally less formal and has a simpler structure. We can see the same pattern also in the case of non-fiction literature where memoirs and autobiographies are less attributive because of their style close to fiction (see Figure 3). Interestingly, there are no big differences in journalism (see Figure 4).

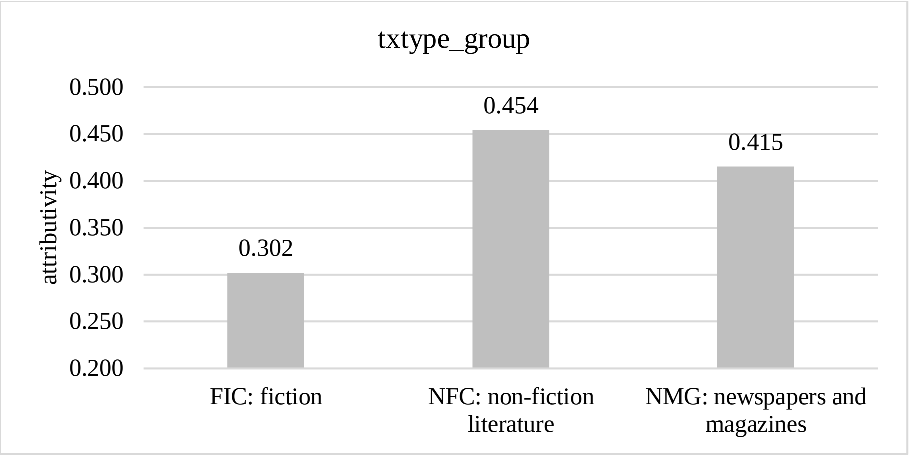

Figure 1: Attributivity in text type groups.

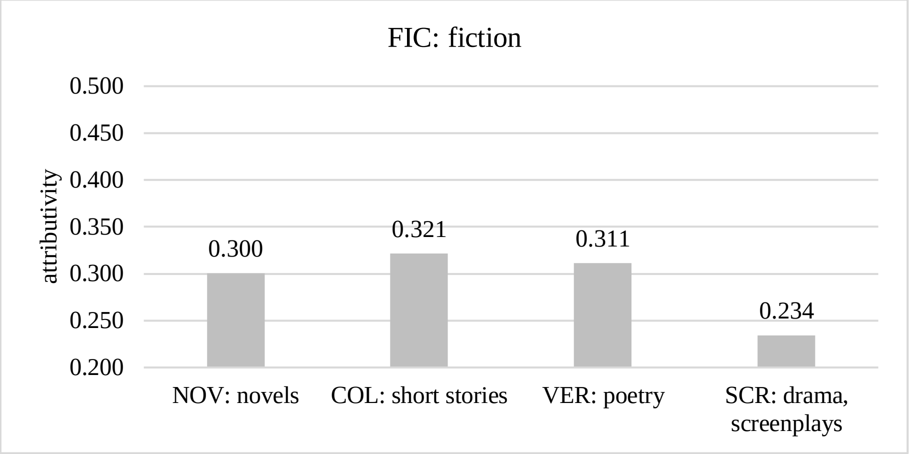

Figure 2: Attributivity in fiction.

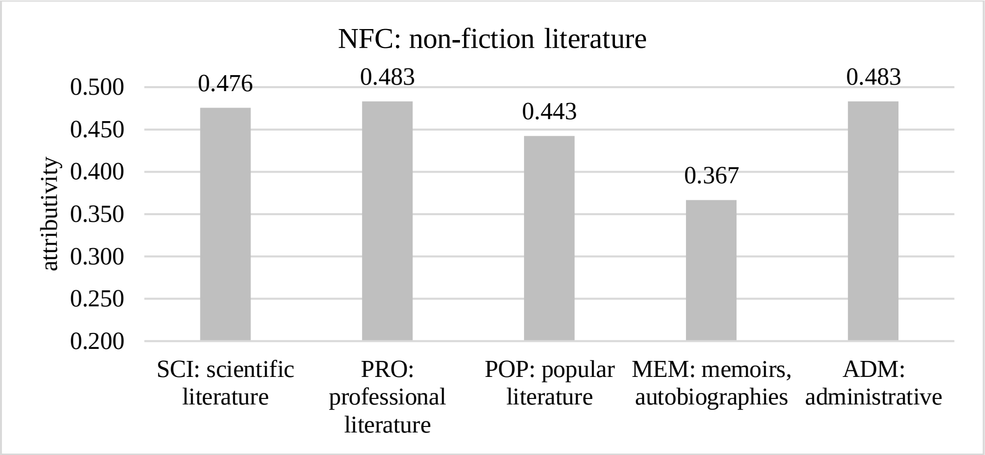

Figure 3: Attributivity in non-fiction literature.

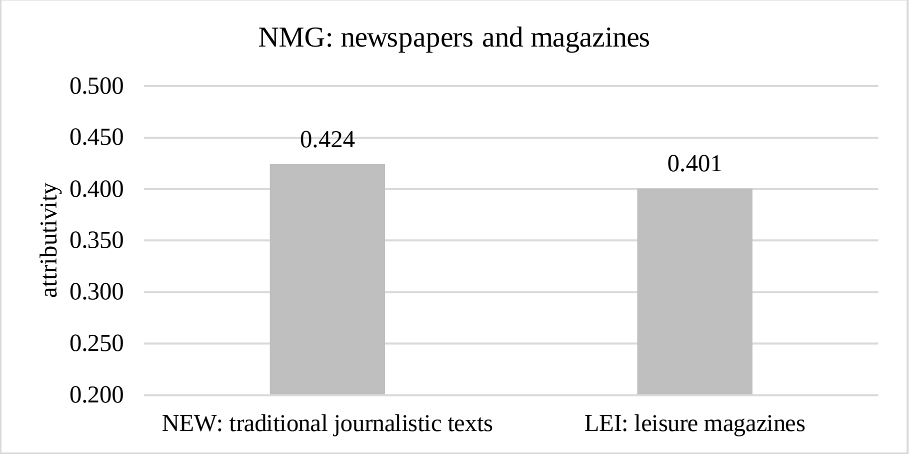

Figure 4: Attributivity in newspapers and magazines.

4.2 Subjectivity

The resulting values of subjectivity in Figure 5 show quite a clear difference between more formal texts (non-fiction, journalism) on the one side and less formal texts (fiction) on the other side. This could be explained by the fact that more formal texts generally tend to have explicit expressions and redundancy, whereas less formal texts prefer simpler forms. Features typical for spontaneous informal spoken language are very common in Czech contemporary fiction literature. The language is simple, sentences are rather short and there are lots of ellipses as well (cf. Hoffmannová et al. 2016). We can also see this tendency in Figure 7 in non-fiction texts where the lowest subjectivity has the genre of memoirs and autobiographies which are close to fiction literature. All the analyzed genres in fiction literature have very similar values of subjectivity (see Figure 6). In journalism (see Figure 8), we can see that magazines have lower subjectivity than daily newspapers. This is in line with our expectations because newspapers have rather formal texts compared to leisure magazines.

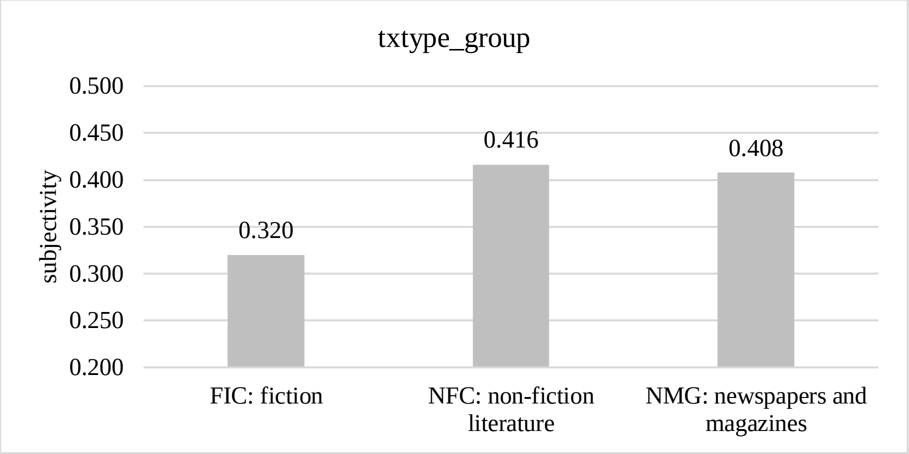

Figure 5: Subjectivity in text type groups.

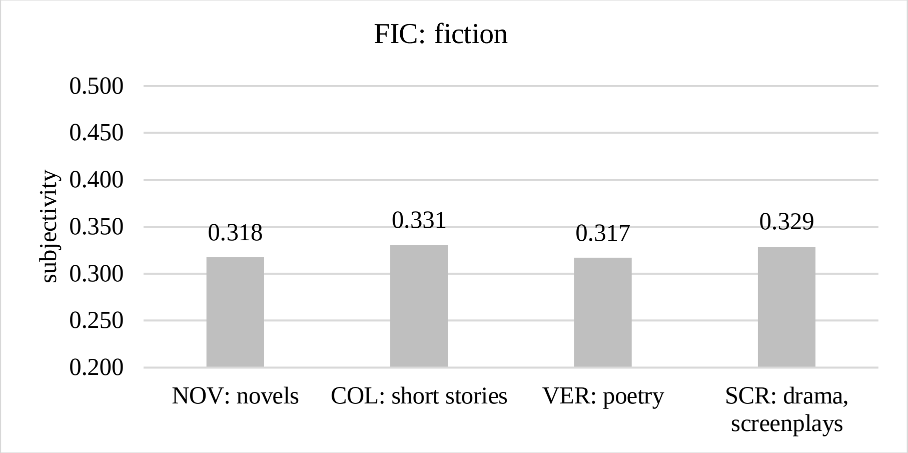

Figure 6: Subjectivity in fiction.

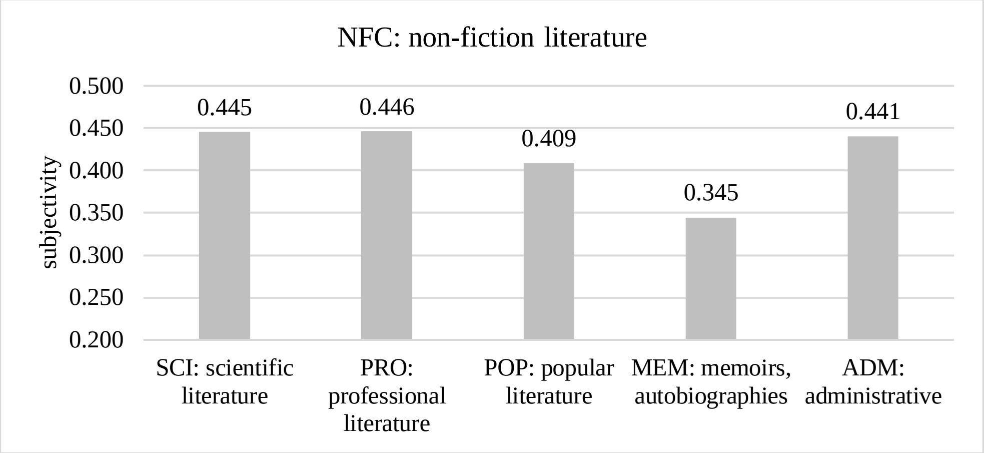

Figure 7: Subjectivity in non-fiction.

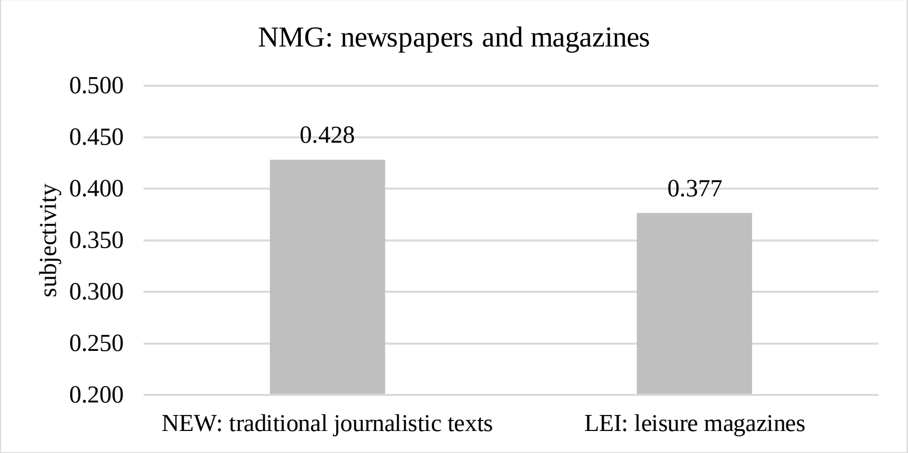

Figure 8: Subjectivity in newspapers and magazines.

5 Conclusion

The proposed syntactic indices (attributivity, subjectivity) seem to be sensitive to different styles and genres. The results show that both indices express stylistic characteristics of texts that can distinguish various text types. We can therefore preliminary conclude that they can be applied in stylometric research as other indices such as nominality, activity, descriptivity, lexical richness.

Fiction literature reached considerably lower values of both indices whereas non-fiction writing reached higher values. Journalism is between them (closer to non-fiction). The more formal text, the higher attributivity and subjectivity. Fiction literature and less formal texts tend to lower subjectivity because of the preference for omitting subjects. Non-fiction literature and more formal texts tend to be more attributive due to the need of precise and detailed description of the nouns and pronouns.

Although the study comes with promising results, we must emphasize that this is just a first attempt to apply indices of attributivity and subjectivity in stylometry. Further research must be done to confirm our preliminary findings. These methods can be also applied in the authorship attribution domain to discover whether attributivity and subjectivity are sensitive to the writing style of different authors. Since our study is limited only to Czech, it is also important to investigate other languages.

Acknowledgements

This work is partly supported by the National Social Science Fund of China (18CYY031).

References

Josef V. Bečka. 1992. *Česká stylistika*. Academia, Praha, Czechia.

Marie Čechová, Marie Krčmová, and Eva Minářová. 2008. *Současná stylistika*. NLN, Praha, Czechia.

Jan Hajič, Eduard Bejček, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague Dependency Treebank – Consolidated 1.0. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020),* pages 5208–5218. Marseille, France

Jana Hoffmannová, Jiří Homoláč, Eliška Chvalovská, Lucie Jílková, Petr Kaderka, Petr Mareš, and Kamila Mrázková. 2016. *Stylistika mluvené a psané češtiny*. Academia, Praha, Czechia.

Michal Křen, Václav Cvrček, Jan Henyš, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Dominika Kováříková, Jan Křivan, Jiří Milička, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Jana Šindlerová, and Michal Škrabal. 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.

Miroslav Kubát. 2016. *Kvantitativní analýza žánrů.* University of Ostrava, Ostrava, Czechia.

Marie Těšitelová. 1985. *Kvantitativní charakteristiky současné češtiny*. Academia, Praha, Czechia.

Ludmila Uhlířová. 1974. O frekvenci větných členů v souvislém textu. *Slovenská reč*: 39(3), 141–146.

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

Appendix

We used the following CQL (corpus query language) queries 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.*"]