sciendo
DOI 10.2478/jazcas-2021-0050

THE NEW VALUE OF THE STRUCTURAL ATTRIBUTE *SECTION* IN THE SYN v8 CORPUS AND ITS POSSIBLE APPLICATION IN LINGUISTIC RESEARCH

ZUZANA LAUBEOVÁ – MICHAL ŠKRABAL
Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic

LAUBEOVÁ, Zuzana – ŠKRABAL, Michal: The new value of the structural attribute *section* in the SYN v8 corpus and its possible application in linguistic research. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 545 – 555.

Abstract: The paper introduces a new section separated from journalistic texts in Czech corpora, namely interviews. This genre is highly specific; from among the texts that can be found in newspapers and magazines, it is probably the closest to spoken language. In two case studies, we present the possible application of the interviews subcorpus in linguistic research. The first one deals with the role of paralinguistic behaviour, especially laughter in written interviews vs. spoken dialogues. The second one investigates the specifics of the demonstrative *ten* in the function of a nominal attribute, again in both written and spoken data.

Keywords: Czech spoken corpora, interviews, paralinguistic behaviour, determiner *ten*

1 INTRODUCTION

Language corpora contain many metadata which help to analyse the actual data. In general, information about written texts is divided according to the whole document, the text itself, the paragraph, and the sentence, and is generally referred to as containing structural attributes. Each of them is filled with different values. This paper aims to introduce the structural attribute of a *section,* covering the content of a newspaper or a magazine (hereinafter: NMG) split into the sections, e.g., news, sport, crime, etc. The attribute value is based on the original newspaper; therefore, it can vary from one title to another or stay empty, unfilled.

The attribute of *section* has been introduced in the corpus of written Czech SYN2015 [1]. Besides 13 original values of this attribute, another one – *rozhovory* (‘interviews’) – was added in the SYN v8 corpus [2]. This genre is highly specific; of most of the texts in NMG, it is probably the closest to spoken language – albeit admitting that interviews are edited and “smoothed” towards the easy-to-read form. Moreover, it is easily detectable too (most usually titled *Rozhovor s…* ‘Interview with’), and as such, it served us well as a starting dataset for our research. Among other things, it turned out that the information about the interviewee’s behaviour is usually added in the parentheses, and these paralinguistic comments (e.g., *směje se* 'laughs', *krčí rameny* 'shrugs his/her shoulders', etc.) can specify the interpretation of the speaker's verbal message. We present the results and compare them with the transcribers' comments within the Czech spoken corpora in the first case study of our paper. The second case study deals with the demonstrative *ten* ('the') in the function of a nominal attribute in both written and spoken data.

2 CASE STUDY 1: PARALINGUISTIC COMMENTS IN NMG INTERVIEWS

2.1 Introduction

The first case study deals with the role of paralinguistic behaviour, especially laughter, in written interviews vs. spoken dialogues. Especially recently, the interviewee's nonverbal behaviour is sometimes captured in NMG interviews (hereinafter: NMGi) to specify and/or complete the meaning of the utterance. It is similar to the author's comments about the behaviour of characters in a script or a play.

What the comment is comprised of and what is essential to mention depends on an interviewer (or an editor). There are no strictly given rules or requirements. We can only assume that there is a common practice of processing the spoken transcript of an interview into a written form that is different in each newspaper or magazine, and this practice also includes the use of comments.

This study follows up and expands on the unpublished pilot study [3], conducted on uncategorized journalistic texts from the SYN v6 corpus [4]. Preliminary results showed that the three most frequent comments (laughter, smile, thinking) covered 95% of all comments (see Table 4). From the formal point of view, comments are mainly composed of a verb, a verb with an adverb, or, alternatively, a more complex structure (e.g., *začne se hlasitě smát* 'starts to laugh loudly'). Aiming to verify to what extent these results are valid only for the written interviews, we used the new section attribute within the SYN v8 corpus [2]. Further, the results from both written corpora are compared to the transcribers' paralinguistics comments included in the spoken corpora of Czech.

In general, our paper aims to reveal which types of comments are incorporated the written interviews and how they are structured. We also consider the overall motivation of paralinguistic comments in the texts.

2.2 Data and methodology

This study is based on three corpora of today's Czech. Within the context of the written corpora SYN v6 and v8, we focus on the journalistic texts only, as we presume a higher frequency of paralinguistic comments there. SYN v6 has no particular category for written interviews; therefore, we had to work with the whole journalistic subcorpus (4.36G), using the following CQL query:

[word=”\(“][word=”(?i)[aábcčdďeěéfghiíjklmnňoópqrřsštťuúůvwxyýzž]{1,20}” & pos!=”[XC]” & tag!=”.{14}8.”]

On the contrary, the NMGi subcorpus (over 2M tokens) can be delimited within the newer SYN v8 corpus, and the CQL query syntax is much more straightforward:

[word=”\(“] within <text section=”rozhovory” />

The ORTOFON v2 corpus [5] (2.5M) represents the synchronic spoken language. We benefit from the fact that the comments are tagged as the “M” part of speech.

[pos=”M”]

The results from both written corpora were manually filtered, focusing on the search for relevant results, i.e., the comments which describe the interviewee’s paralinguistic behaviour.

2.3 Results

Firstly, we compare the content of comments within the SYN v6: NMG subcorpus with the SYN v8 corpus. Although there are more than 6M hits, most of them needed to be filtered out. Table 1 shows the ten most frequent types. The relevant occurrences are in bold.

	SYN v6: NMG	**ipm**
1.	*(na snímku)* ‘on the photograph’	28.1
2.	*(vlevo)* ‘on the left’	19.8
3.	***(smích)*** **‘laughter’**	15.4
4.	*(vpravo)* ‘on the right’	14.5
5.	***(směje se)*** **‘is laughing’**	6.6
6.	*( )* [website removed]	5.3
7.	*(ne)* ‘no/non-’	4.9
8.	***(úsměv)*** **‘smile’**	4.4
9.	*(uprostřed)* ‘in the middle’	4.0
10.	*(ANO)* [abbreviation of Czech political party]	3.9

Tab. 1. Top 10 most frequent chunks in parentheses in the SYN v6: NMG subcorpus

Table 1 illustrates what kind of information is mainly captured within parentheses as comments in NMG. These results are similar in both written corpora.

Besides paralinguistic behaviour, it can be a caption of a photograph or a picture (nr. 1, 2, 4, 9 in Table 1), a quotation of a website (nr. 6), or one's affiliation to a political party (nr. 10). Also, the negation particle *ne* (nr. 7) was found in texts quite often, e.g., *Vláda (ne)schválila daňovou reformu* 'The government has (not) approved the tax reform'.

Filtered results show the prevalence of laughter or smiles within both corpora (see also Table 2 below). Looking closer, we could identify the varied modes/phases of laughing and smile and/or different length of their duration. The third most frequent behaviour is the process of thinking, also of various lengths or efforts. Other types of comments describe pauses, gestures, facial expressions, or sighs – in general, the nonverbal physical or physiological behaviour. There are also the interviewer's comments on the interviewee's speech (e.g., *hledá v mobilu* 'is searching on the mobile' *skáče/skočí do řeči* 'is cutting in', *nesouhlasně vrtí hlavou* 'shakes his/her head in disapproval') and noises from outside (a phone ringing), other people's reactions (*jeho žena přikyvuje* 'his wife nods'). The mental state of the speaker is described by adverbs (e.g., *smutně* 'sadly', *pobaveně* 'amusedly', *zklamaně* 'disappointedly'), which may be added to the verbal comment, too.

rank	**SYN v6: NMG**	**ipm**	**SYN v8: NMGi**	**ipm**
1.	*(smích)* 'laughter'	15.4	*(smích)* 'laughter'	145.3
2.	*(směje se)* 'is laughing'	6.6	*(směje se)* 'is laughing'	50.8
3.	*(úsměv)* 'smile'	4.4	*(úsměv)* 'smile'	30.3
4.	*(usmívá se)* 'is smiling'	2.1	*(usmívá se)* 'is smiling'	9.8
5.	*(usměje se)* 'smiles'	1.3	*(skáče do řeči)* 'cuts in'	2.8
6.	*(rozesměje se)* 'laughs'	0.3	*(ukazuje něco/na něco)* 'points at sb/sth'	1.4
7.	*(přemýšlí)* 'is thinking'	0.2	*(usměje se)* 'smiles'	0.9
8.	*(pousměje se)* 'half-smiles'	0.2	*(se smíchem)* 'with laughter'	0.9
9.	*(zamyslí se)* 'thinks'	0.1	*(vehementně předvádí)* 'demonstrates vehemently'	0.5
10.	*(důrazně)* 'strongly'	0.1	*(nesouhlasně vrtí hlavou)* 'shakes his/her head in disapproval'	0.5
11.	*(odmlčí se)* 'falls silent'	0.1	*(přemýšlí)* 'is thinking'	0.5

rank	SYN v6: NMG	ipm	SYN v8: NMGi	ipm
12.	*(s úsměvem)* 'with a smile'	<0.1	*(rozesměje se)* 'laughs'	0.5
13.	*(zasměje se)* 'laughs'	<0.1	*(hledá v mobilu)* 'is searching on the mobile'	0.5
14.	*(skočí do řeči)* 'cuts in'	<0.1	*(zamýšlí se)* 'think about sth'	0.5
15.	*(kroutí hlavou)* 'shakes one's head'	<0.1	*(s úsměvem)* 'with a smile'	0.5
16.	*(chvíli přemýšlí)* 'is thinking for a while'	<0.1	*(pokyvuje hlavou)* 'nods'	0.5
17.	*(dlouho přemýšlí)* 'is thinking for a long time'	<0.1	*(dlouze přemýšlí)* 'is thinking for a long time'	0.5
18.	*(se smíchem)* 'with laughter'	<0.1	*(smutně se usmívá)* 'smiles sadly'	0.5
19.	*(povzdechne si)* 'sighs'	<0.1	*(chvilku přemýšlí)* 'is thinking for a while'	0.5
20.	*(skáče do řeči)* 'is cutting in'	<0.1	*(úsměv na tváři mluvčího)* 'smile on the speaker's face'	0.5

Tab. 2. Comparison of the top 20 paralinguistic comments in the SYN v6: NMG and SYN v8: NMGi subcorpora

The results from the written corpora (Table 2) partially correspond with the comments of paralinguistic comments in the ORTOFON v2 corpus (Table 3). These comments are added during the process of transcription and focus on the sounds that could influence spontaneous dialogue. There are not only paralinguistic, nonverbal expressions of a speaker, such as breathing in and out, but also sounds accompanying the speech, e.g., noise from the street, knocking on the door, clearing the throat, etc. Table 3 shows that laughter is the third most frequent comment and a faint smile the fifth one. This type of comment is the only one that is exactly the same as in the written data.

rank	ORTOFON v2	ipm	rank	ORTOFON v2	ipm
1.	*(nadechnutí)* 'breathing in'	8,143	11.	*(hlasitý hovor v pozadí)* 'loud talking in the background'	580
2.	*(rušivý zvuk)* 'disruptive sound'	5,366	12.	*(citoslovce)* 'interjection'	541
3.	*(smích)* 'laughter'	4,483	13.	*(smích více mluvčích najednou)* 'collective laughter of multiple speakers'	376
4.	*(hluk v pozadí)* 'noise in background'	2,682	14.	*(povzdech)* 'sigh'	362

rank	ORTOFON v2	ipm	rank	ORTOFON v2	ipm
5.	*(pousmání)* 'faint smile'	2,531	15.	*(zvuk z rádia)* 'sound from a radio'	339
6.	*(mlasknutí)* 'lip smacking'	1,583	16.	*(ruch z ulice)* 'noise from the street'	318
7.	*(cinkání nádobí)* 'clinking dishes'	1,553	17.	*(zvuky při jídle)* 'sounds during eating'	308
8.	*(odkašlání)* 'clearing the throat'	690	18.	*(polknutí)* 'swallow'	294
9.	*(vydechnutí)* 'breathing out'	646	19.	*(mluví ke zvířeti)* 'talks to animal'	272
10.	*(klepání)* 'knocking'	592	20.	*(zvuk z televize)* 'sound from TV'	266

Tab. 3. Top 20 most frequent paralinguistic comments in the ORTOFON v2 corpus

2.4 Summary

This study focused on the paralinguistic comments in written journalistic texts. The most frequent comments include laughter, smile, and thinking (Table 4).

	SYN v6: NMG	SYN v8: NMGi
laughter	72%	80%
Smile	21%	17%
thinking	2%	1%
Other	5%	2%

Tab. 4. The types of comment according to their meaning within written corpora

The frequency of laughter and smile indicates that these comments are considered essential for readers to understand the tone of the interview properly. We assume that this is a primary motivation for their incorporation into texts despite a somewhat foreign nature in the stream of speech and potential difficulties with their conversion into words.[1] The presence of paralinguistic comments in NMGi is, at least from the perspective of frequency, a typical feature for this register (similar to scripts).

[1] The laughter or smile comments are sometimes replaced with emoticons, e.g., – *Která historka vás v poslední době pobavila? – Ta, že jste si za mnou přišel pro rozhovor. :)* '– Which story has amused you lately? – That you came to me for an interview. :)'.

3 CASE STUDY 2: THE DETERMINER *TEN* IN NMG INTERVIEWS

3.1 Introduction

The second case study concerns the specific role of the demonstrative *ten*[2] as a determiner. Although Czech belongs to languages without a definite article, its function is repeatedly attributed to the pronoun *ten*, and there are recurring hypotheses about the gradual emergence of the category of definiteness in, predominantly spoken, Czech (as early as in 1917: [6], most recently [7]). We want to verify this hypothesis – purely quantitatively at the moment – on the spoken corpora of Czech, including their specific segment of NMGi. We limit ourselves to a quantitative analysis only, which should then be supplemented by a deeper qualitative analysis, e.g., in a similar way as described in [7].

3.2 Data and methodology

To get an insight into behaviour of the pronoun *ten* as a means of deixis, we used the following CQL query:

```
(1:[lemma=”ten” & tag=”..M.*”] [pos=”[AP]”]{0,3} 2:[tag=”N.M.*”]) |
(1:[lemma=”ten” & tag=”..I.*”] [pos=”[AP]”]{0,3} 2:[tag=”N.I.*”]) |
(1:[lemma=”ten” & tag=”..F.*”] [pos=”[AP]”]{0,3} 2:[tag=”N.F.*”]) |
(1:[lemma=”ten” & tag=”..N.*”] [pos=”[AP]”]{0,3} 2:[tag=”N.N.*”]) & 1.case=2.
case
```

We are looking for a combination of the pronoun *ten* and a noun (with up to 3 potentially inserted tokens). At the same time, both expressions must match in the case (condition 1.case = 2.case) and gender (which has no attribute in the corpora, thus it is necessary to specify via a tag successively all four genders: masculine (in) animate M/I, feminine F, neuter N). We are aware that unwanted hits remain in our dataset (besides cases of actual deixis, these are, e.g., multiword expressions such as *v tom případě* ‘in that case’, *tou dobou* ‘at the time’, etc.). Still, we intentionally do not want to advance to their manual filtering as we are only interested in obtaining and comparing the relative frequency (ipm) of this phenomenon in different corpora.

3.3 Partial results

The results summarized in Table 5 correspond to our initial hypothesis: the structure is most abundantly represented in spoken data, and within them, most often in ORATOR, the corpus of monologues about which speakers are informed in advance and for which they can prepare [8]. The speaker’s effort to clearly identify the noun in question and/or to emphasize it duly in the structure of their lecture can explain the higher occurrence of the structure. We cannot exclude a double deixis –

[2] The lemma *ten* (the singular nominative form for the masculine gender) also includes the forms of the feminine and neuter gender *ta* and *to*, as well as plural forms *ti*, *ty*, and *ta*. In all the corpora we use in this chapter, the lemmatization is unified, including all these forms into the given lemma.

a verbal and physical one, i.e., real pointing to the subject using a pointer or a finger (in a presentation, on a blackboard, etc.), parallel with the utterance of the demonstrative noun. Another possible explanation includes the speaker's attempt to be informal or gain more time to word an idea by adding redundant expressions. In written corpora, this structure most often appears in fiction, while there is a significant difference between the subcorpus of interviews and NMG in general (see lines 4 and 3).

(sub)corpus	size in tokens	hits	frequency (ipm)
SYN2020: FIC	33.3M	188,013	1,543
SYN2020: NFC	33.3M	46,811	384
SYN2020: NMG	33.3M	53,745	441
SYN v8: NMGi	2.2M	6,069	2,827
ORTOFON v2	2.1M	20,794	8,121
ORAL v1	5.4M	53,500	8,410
ORATOR v2	1.2M	17,296	11,263

Tab. 5. The frequency of *ten* structures in the selected Czech corpora

Let us start with spoken language represented by the ORTOFON v2 corpus. The speaker in example (1) has a prominently higher ipm of the *ten* structure compared to ipm for the whole corpus (13,115 vs. 8,121), and it is obviously a salient feature of his idiolect.

(1) 15T008N, Ignác V. [talks about his visit to Jerusalem and comments on the behavior of Orthodox Jews at the Wailing Wall]
*takže **ty** bezpečnostní **ty** tam jsou jako hodně no tak jsme prošli .. a pak jsme byli fakt u **tý** Zdi nářků a tam to je úplně teda speciální tam .. fakt choděj jenom **ti** ortodoxní Židi takový **ty** dlouhý kabáty .. fakt choděj jenom **ti** ortodoxní Židi takový **ty** dlouhý kabáty .. a prostě u **té** Zdi nářků všichni tak jako se kolébají a říkají **tu** modlitbu*

On the contrary, the next speaker's utterance is laconic in terms of the frequency of the pronoun *ten* (ipm 2,575):

(2) 20A011N, Dušan M.[3] [depicts a difficult traffic situation at an intersection while driving a car]

[3] There is an incorrect piece of information in the database, this is in fact a female speaker. Potential yet unrealized occurrences of the pronoun are denoted by **(0)**, or **(0?)** in questionable cases, respectively.

tak jsem jela a říkám Honzíku tak . dobrý .. to snad dám .. snad to nikde tatínkovi nepoškodím .. bude to dobrý zavezu tě do **(0)** *školy . maminka tě odveze nakoupí přijede dom .. no nicméně jsem přijela .. do Ole* @ do* **(0?)** *Olešnice na* **(0?)** *křižovatku .. @ viděla jsem že jede . velkej traktor tak říkám .. tak . ho pustím udělám dobrej skutek pustím ho .. pustila jsem* **(0)** *traktor .. rozjela jsem se . a* **(0)** *auto mně chcíplo . uprostřed* **(0)** *křižovatky ... za mnou auto vedle mě auto vedle mě dělníci všichni se mi @ mně smáli já jsem .. červenala začala jsem .. panikařit t* .. nezačala jsem panikařit začala jsem nadávat proč mi* ***to*** *auto půjčuje ..*

The following sample comes from the ORATOR v2 corpus, which shows the highest ipm (11,263) within the spoken corpora.

(3) 18X058F, Pavelka S. (ipm 17,670) [comments on a graph showing a decreasing amount of exercise for current children compared to previous years]
zkrátka ukazuje se asi to že tady někde okolo ***toho*** *roku devadesát šest .. už to množství* ***těch*** *realizovanejch pohybovejch aktivit .. kleslo pod* ***tu*** *biologickou potřebu .. a teď už jsme v období . kdy jenom se zhoršujeme .. otázka je jak se z toho dostat dál a jak tomu .. asi .. u jednoho 3 .. co jsou doporučení vokolo* ***toho*** *pohybu ..*

Naturally, we do not find spoken language only in spoken corpora. With some retreat from authenticity, we can find it also in written corpora, chiefly in fictional speech (as opposed to narrative). No matter how successful a writer is in pursuit of representing genuine speech, it is easily detectable in the corpus, be it by quotation marks or similar means. The situation in NMGi is different. Although we cannot be entirely sure (if we did not directly hear the speech from the authentic recording) to what extent it was corrected, paraphrased, re-stylized, etc. by an interviewer, an editor, or a proofreader, a general idea of the degree of authenticity of the quoted statement can often be made. As in the following example from NMGi (along with examples (1) and (3)), it shows occurrences of the pronoun *ten* not really corresponding to its primary (i.e., deictic) function:

(4) Sedmička, č. 38/2018 [a singer presents his new album]
A ještě musím k ***té*** *desce říct jednu věc . Je důležité , že to zpívám v češtině . Teď je trend zpívat v angličtině a* ***ty*** *věci v ní samozřejmě byly napsány , jsou to zahraniční věci , ale moje specialita je zpívat v češtině . Některé věci se tak k lidem dostanou snáz . Líp to pochopí . Navíc* ***ta*** *práce s češtinou je o hodně těžší než to zpívat anglicky . V angličtině to jde samo . Zazpíváte větu a jste světový zpěvák . Ale u* ***té*** *češtiny musíte prokázat interpretační zralost a schopnost .*

3.4 Summary

Our second topic is not a full-fledged case study, but rather a quantitative probe in regards to available corpus data. Primarily, we aimed to indicate the connection between the use of the demonstrative *ten* and specific registers and to show a descending tendency from genuinely spoken data, where *ten* is a strong indicator of spontaneity and/or emotionality, to written data. Nevertheless, even in written registers, e.g., in NMGi, they may play a prominent role: their increased frequency contributes to possible grammaticalization of the pronoun *ten* into a definite article. Understandably, deeper qualitative analysis, preceded by manual filtration of objectionable results, is necessary.

4 CONCLUSION AND FUTURE WORK

The specificity of the NMGi register is obvious, and we must always be aware that it is an inauthentic, further modified linguistic imprint of a speaker. Therefore, it should not be confused with genuine data in spoken corpora. The extent of the interventions by an interviewer, an editor, a proofreader, or other members of editorial staff can only be speculated, and we do not know of any study that would address this issue, at least for Czech. It would certainly be illuminating to have authentic recordings available and then compare them with the final form of interviews. What do the editors consider undesirable in the interviewee's speech (in terms of content and/or language), what the reader "stand", etc.? These are only a few of the many research questions waiting to be answered.

Our paper aimed to demonstrate the usefulness of extending the list of the section attribute in the Czech NMG subcorpora by interviews, truly a specific register worthy of linguists' attention. In the first case study, we examined how the diverse types of respondents' paralinguistic behaviour are recorded in interviews and with what frequency. In the second case study, we focused on the frequency of the pronoun *ten*, which is significantly more salient in spoken utterances than in written texts. That supports the hypothesis of its gradual grammaticalization, i.e., transformation into a definite article in spontaneous spoken Czech.

The potential that such a handily defined NMG section has is far from exhausted. To look for other phenomena (e.g., contact particles or various n-grams such as *já si myslím že* 'I think that', *je/není to o* 'it is (not) about', *na druhou stranu* 'on the other side', etc.) in the NMGi subcorpus and compare them with different registers is equally tempting.

ACKNOWLEDGEMENTS

This paper resulted from the implementation of the Czech National Corpus project (LM2018137) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

[1] Křen, M. et al. (2015). SYN2015: reprezentativní korpus psané češtiny. Praha: Ústav Českého národního korpusu FF UK.

[2] Křen, M. et al. (2019). Korpus SYN, verze 8 z 12. 12. 2019. Praha: Ústav Českého národního korpusu FF UK. Accessible at: http://www.korpus.cz.

[3] Komrsková, Z., and Škrabal, M. (2018). The role of paralinguistic behaviour, especially laughter in written interview vs. spoken dialogue. A corpus-based study. Poster at the Second International Conference on Sociolinguistics (ICS.2), 6–8 September 2018, Budapest.

[4] Křen, M. et al. (2017). Korpus SYN, verze 6 z 18. 12. 2017. Praha: Ústav Českého národního korpusu. Accessible at: http://www.korpus.cz.

[5] Kopřivová, M. et al. (2020): ORTOFON v2: Korpus neformální mluvené češtiny s víceúrovňovým přepisem. Praha: Ústav Českého národního korpusu FF UK. Accessible at: http://www.korpus.cz.

[6] Zubatý, J. (1917). Ten. Naše řeč, 1(10), pages 289–294.

[7] Dvořák, J. (2020). The emerging definite article ten in (informal spoken) Czech: a further analysis in terms of semantic and pragmatic definiteness. Naše řeč, 103(4), pages 297–319.

[8] Kopřivová, M. et al. (2020): ORATOR v2: Korpus monologů. Praha: Ústav Českého národního korpusu FF UK. Accessible at: https://www.korpus.cz.