sciendo
DOI 10.2478/jazcas-2021-0039

ON CORPUS-DRIVEN RESEARCH OF COMPLEX ADVERBIAL PREPOSITIONS WITH SPATIAL MEANING IN CZECH

AKSANA SCHILLOVÁ
Czech Language Institute, Czech Academy of Sciences, Prague, Czech Republic

SCHILLOVÁ, Aksana: On corpus-driven research of complex adverbial prepositions with spatial meaning in Czech. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 425 – 433

Abstract: Complex adverbial prepositions with spatial meaning have not been sufficiently studied so far in Czech. To establish a set of these expressions in their actual usage, the resources of the Czech National Corpus were used in this study. The research has shown that the SYN2020 corpus is a relevant tool for searching for two-word expressions with a LOCATIVE ADVERB – SIMPLE PREPOSITION structure that have the same function as a one-word locative preposition. The article describes a method for the extraction of these expressions from the corpus, as well as a method for the collection of their quantitative data using corpus tools. As a result of the research, a list of expressions that are presumably complex prepositions is provided.

Keywords: complex preposition, locative adverb, spatial meaning, Czech language, Czech National Corpus

1 INTRODUCTION

The article deals with the use of the Czech National Corpus (CNC) in the study of two-word expressions that have a LOCATIVE ADVERB – SIMPLE PREPOSITION structure and can function as a complex preposition with spatial meaning. In Czech linguistic literature, only four adverbial complex prepositions with spatial meaning are described, or at least mentioned: *stranou od* 'aside from' ([1], [2], [3], [4], [5]), *daleko od* 'far from' [3]; *napravo od* 'to the right of' and *nalevo od* 'to the left of' [1]. Since a number of these units are detected and described in other Slavic languages (see the next section), it can be assumed that these complex prepositions also occur in Czech but have not been studied yet. Cf. the explanatory dictionary of Belarusian prepositions [6] that presents a list of "adverbial-prepositional constructions that function as prepositions" consisting of 127 units, 94 of which have a spatial meaning [6, pp. 165–166].

Since this type of complex prepositions in Czech has not been sufficiently covered in linguistic literature, the study was focused on the search for these complex units in actual language use, namely in a language corpus. The search was carried out on the SYN2020 corpus that is a 100m representative corpus of contemporary written Czech available within the CNC project [7].

This paper describes the extraction of statistically significant ADVERB – PREPOSITION combinations from the corpus using such corpus tools as advanced queries, positive/negative filters, frequency distribution, collocation candidates, association measures (T-score, MI-score, logDice), etc. The extracted combinations are non-random from statistical point of view and can be considered candidates for prepositionalization due to their fixity, idiomaticity, and frequency.

The expressions with the ADVERB – PREPOSITION structure have been already discussed in a 1977 article by Kroupova [8] as having the potential for prepositionalization. At the present day, we can evaluate these expressions in the light of corpus data.

2 THEORETICAL BACKGROUND

The present paper is mainly inspired by the following motivations:

1) Czech adverbs, as well as complex prepositions, are insufficiently covered in linguistic literature.

As for adverbs, only a few pages are devoted to their general description in Czech grammars (see [3], [9], [10], [11]). Moreover, the grammars focus mainly on description of adverb formation and degrees of comparison of adverbs. However, for example, the valence properties of adverbs have not been studied in detail [12], neither the limits of this word class, and the recognition criteria for adverbialization, too, remain terra incognita ([13], [14], [15]).

As for Czech complex prepositions, the monographs [1] and [5] make a significant contribution to their description. Nevertheless, these works primarily focus on the description of prepositional expressions that are derived from nouns, cf. the lists of prepositions represented in [1, pp. 323–333] and in [5, pp. 39–49]. The Czech prepositional system, taken as a whole, does not have yet a comprehensive linguistic description at the level of a monograph. Besides, there is no dictionary of Czech prepositions and prepositional expressions, while there are, for example, dictionaries of Ukrainian, Belarusian, Russian, Polish prepositions, and their analogues ([6], [16], [17], [18], [19]).

2) The study of complex adverbial prepositions, which are a neglected area of Czech grammar, can contribute to the lexical database of the LEMUR project [20].

The main task of this project is to create a new electronic linguistic resource, namely a database of Czech multiword expressions, which will subsequently be useful for many reasons, e.g., for teaching Czech as a foreign language, for lexicography, for the creation and improvement of natural language automatic processing tools ([20], [21], [22]). In this database, Czech multiword expressions will be presented and comprehensively described. The project also develops a typology of these units, in which complex prepositions are considered one of the syntactic types of Czech multiword expressions [22, p. 44].

3 DATA EXTRACTION FROM THE SYN2020 CORPUS

3.1 Methods of prepositionalization candidates extraction from the corpus

When we use representative SYN-series corpora [23], we must remember that these corpora contain not only original Czech texts but also texts translated into Czech from other languages.[1] Thus, to exclude the influence of a source language on the results of the present study, a subcorpus within the SYN2020 corpus was created consisting only of original Czech texts (size in tokens ca. 80m).

In this subcorpus, the search was carried out through the following advanced query: [tag="D.*"][tag="R.*"]. It means that all the possible ADVERB – PREPOSITION combinations which occur in the corpus were searched.[2] As a result, a concordance spanning 474.254 hits was generated.

The next step was to obtain a frequency list of combinations that were found. To make the frequency list, the menu item Frequency (Frequency Custom… > Frequency distribution) was used. The resulting list consisted of more than 20 thousand combinations arranged in descending order of their absolute frequency.

Since there is no semantic annotation in the SYN corpora, combinations with spatial meaning were selected manually from the list. The selection was made from the first thousand most frequent combinations. In a future study, this sample can be expanded to include less frequent combinations that have a rank higher than 1000.

3.2 Methods of quantitative data collection

After extracting data on the absolute frequency of the ADVERB – PREPOSITION combinations (see above), their relative frequency was calculated, i.e., a correlation between the absolute frequency of the entire combination and that of its adverb.[3] A relative frequency that is above average can be considered as a sign of stability of a given combination, cf. [1, p. 50].

In addition to the data on the absolute and relative frequency of the expressions studied in this paper, their statistical values were also analysed. For this purpose, using the menu item Collocations, collocation lists for the locative adverbs that are members of these expressions were obtained. The span of collocations was restricted to the first position to the right from a key word (a locative adverb). From the

[1] On the language of translations into Czech (see [24], [25], [26]); on the specialized JEROME linguistic corpus for analysing translated Czech, see [27].

[2] Previously, a similar search was carried out for denominal complex prepositions on the SYN2000 corpus by Blatná [1, p. 11]. Cf., the use of the p-collocation tool to search for Czech verbal participles as candidates for prepositionalization by Richterová [28].

[3] In other words, for each combination under study, it was checked how many times the adverb has occurred in the corpus and how many times the corresponding ADVERB – PREPOSITION combination has occurred in the corpus. Then, the percentage was calculated using a simple mathematical formula: Rel. freq. = A × 100 ÷ B, where A is the absolute frequency of the combination, B is the absolute frequency of the adverb that is a component of this combination.

collocation lists, simple prepositions were selected and data on the association measures (MI, T-score, logDice)[4] of the LOCATIVE ADVERB – SIMPLE PREPOSITION collocations were extracted.

Association measures help to identify which co-occurrences of words in a corpus are regular and non-random, namely from the point of view of statistics. This means, in terms of the ratio of the number of occurrences of collocations to the number of occurrences of their components taken separately and to the total number of all words in a corpus, see [29, pp. 103–105].

In the present study, it is assumed that the higher the measures of the statistical association between a locative adverb and a simple preposition, the more likely it is that the collocation is not a free combination of words but a fixed multiword expression that has the same function as a simple preposition.

At the next stage, the data of all the analysed combinations were compared relative to each other. As a result, it was revealed which of the combinations are statistically significant and hence can be considered potential complex prepositions. According to Petkevic et al. [22], complex prepositions show the so-called statistical idiomaticity, which means that these expressions are usually not semantically idiomatic but have an above-average frequency and a restricted collocability [22, p. 52].

4 RESULTS

In the subcorpus of original Czech texts (size in tokens ca. 80m), which was created within the SYN2020 corpus, 20.072 ADVERB – PREPOSITION combinations were found.

In the top ten most frequent combinations, there are two expressions that are already considered as complex prepositions in linguistic literature. These are *spolu s* ‘together with’ (rank 2) and *společně s* ‘jointly with’ (rank 7).[5] See Fig. 1.

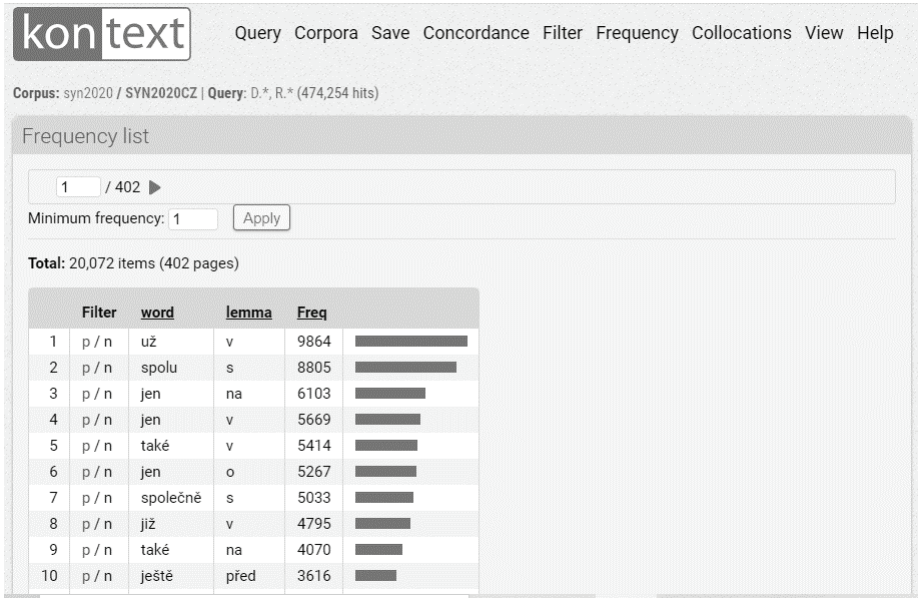
From the first thousand items of this frequency list, combinations with locative adverbs were manually selected: 61 expressions in total. The most frequent expressions are the following ones (their absolute frequency is given in brackets): *daleko od* ‘far from’ (779), *blízko k* ‘close to’ (371), *hluboko do* ‘deep into’ (353), *vysoko nad* ‘high above’ (317), *severně od* ‘north of’ (317), *daleko za* ‘far behind’ (298), *jižně od* ‘south of’ (292).

Some expressions are clearly distinguished from the other ones in terms of their relative frequency. This is a group of expressions that are used to refer to a location according to the cardinal direction (north, east, south, or west). These are the following ones: *severovýchodně od* ‘northeast of’, *severozápadně od* ‘northwest of’,

[4] For what these statistical measures mean and how they are calculated, see the corpus wiki accessible at: https://wiki.korpus.cz/doku.php/pojmy:asociacni_miry.

[5] The expression *spolu s* ‘together with’ is considered as a complex preposition in [1], [2], [3], [8], [9], [10], [30]; *společně s* ‘jointly with’ – in [1], [2], [5], [30].

jihovýchodně od ‘southeast of’, *jihozápadně od* ‘southwest of’, *východně od* ‘east of’, *severně od* ‘north of’, *jižně od* ‘south of’, *západně od* ‘west of’, which have a relative frequency of 82 to 92 %.



	Filter	word	lemma	Freq
1	p / n	už	v	9864
2	p / n	spolu	s	8805
3	p / n	jen	na	6103
4	p / n	jen	v	5669
5	p / n	také	v	5414
6	p / n	jen	o	5267
7	p / n	společně	s	5033
8	p / n	již	v	4795
9	p / n	také	na	4070
10	p / n	ještě	před	3616

Fig. 1. The beginning of the frequency list of the ADVERB – PREPOSITION combinations

For example, the adverb *severovýchodně* ‘northeast’ occurs in the subcorpus 86 times in 48 texts. However, in 79 cases, it co-occurs with the preposition *od* ‘of’, as in example (1) below, which is 92 % of the total number of the occurrences of the adverb in the subcorpus.

As for collocates on the right, the combination *severovýchodně od* ‘northeast of’ co-occurs exclusively (a) with toponyms, e.g., *Brno* ‘Brno’, *Bratislava* ‘Bratislava’, *Krakov* ‘Krakow’, *Japonsko* ‘Japan’, *Beringovy užiny* ‘Bering straits’, etc., 67 % of all the cases, or (b) with nouns referring to a place or an object in space, e.g., *obec* ‘municipality’, *vesnice* ‘village’, *město* ‘town’, *kostel* ‘church’, *nádraží* ‘station’, etc., 33 % of all the cases.

As for collocates on the left, the combination *severovýchodně od* ‘northeast of’ does not show the restricted collocability. In addition, it can be used as an element of an adverbial modifier separated by commas, as in example (2) below.

(1) *Farma leží* ***severovýchodně od Pekingu*** *a je největším zařízením svého druhu v Asii.* ‘The farm is located northeast of Beijing and is the largest facility of its kind in Asia.’

(2) *Na polích kolem Ovčár,* ***severovýchodně od Kolína****, můžeme i dnes najít valouny chalcedonu, jaspisu i křemene, pocházejících rovněž z labských teras.*
'In the fields around Ovcary, northeast of Kolin, we can still find boulders of chalcedony, jasper and quartz, which also come from the Elbe terraces.'

Moreover, the combination *severovýchodně od* 'northeast of' has a high association measure MI-score: 9.402. The other combinations of this semantic group (*severozápadně od* 'northwest of', *jihovýchodně od* 'southeast of', *jihozápadně od* 'southwest of', *východně od* 'east of', *severně od* 'north of', *jižně od* 'south of', *západně od* 'west of') also have MI > 7. Note that the boundary MI = 7 is considered relevant for systemic collocations [31]. Thus, from statistical point of view, these word combinations are systemic, fixed, and non-random.

The statistical data of all the combinations were summarized in one table and sorted by MI (from the highest to the lowest value). A fragment of the table (its beginning) that includes combinations with MI > 7 is presented in Table 1, where *Rank* is the position of the combination in the frequency list of bigrams [tag="D.*"] [tag="R.*"] sorted by their absolute frequency in descending order; *Abs. freq.* is the absolute frequency of the combination in the subcorpus, *Rel. freq.* is the percentage ratio of the absolute frequency of the combination to the absolute frequency of the adverb that is the left component of the combination.

Rank	**Combination**	**Abs. freq.**	**Rel. freq., %**	**MI ↓**	**T-score**	**logDice**
976	*nízko nad* 'low above'	75	14,8	10.743	8.655	5.810
259	*vysoko nad* 'high above'	317	18,5	10.445	17.792	7.873
422	*východně od* 'east of'	190	89,2	9.415	13.764	5.146
938	*severovýchodně od* 'northeast of'	79	91,9	9.402	8.875	3.881
285	*jižně od* 'south of'	292	83,9	9.389	17.063	5.765
260	*severně od* 'north of'	317	86,6	9.384	17.778	5.883
419	*západně od* 'west of'	192	82,4	9.379	13.836	5.161
953	*severozápadně od* 'northwest of'	77	90,6	9.365	8.762	3.844
305	*hluboko pod* 'deep below'	274	18,5	9.346	16.528	7.425
877	*jihovýchodně od* 'southeast of'	84	89,4	9.276	9.150	3.969

Rank	Combination	Abs. freq.	Rel. freq., %	MI ↓	T-score	logDice
981	*jihozápadně od* 'southwest of'	75	89,3	9.274	8.646	3.806
827	*napravo od* 'right of'	90	22,9	9.202	9.471	4.069
597	*nedaleko od* 'not far from'	132	23,3	8.975	11.466	4.621
84	*daleko od* 'far from'	779	8,9	8.328	27.824	7.169
856	*dole pod* 'down below'	87	3,2	8.268	9.297	5.779
766	*vpravo od* 'right of'	97	3,6	7.961	9.809	4.175
213	*blízko k* 'near to'	371	12,1	7.896	19.181	5.132
978	*vlevo od* 'left of'	75	2,5	7.487	8.612	3.804

Tab. 1. Quantitative data of the LOCATIVE ADVERB – SIMPLE PREPOSITION combinations that have MI > 7

In a future study, the combinations as used in the corpus will be analysed in terms of their semantics, collocability, and syntactic behaviour. The qualitative analysis will help to establish (a) which of them are actually used as complex prepositions, (b) under what conditions they have a prepositional function, (c) at what stage of prepositionalization they currently are.

5 CONCLUSION

The present study has shown that the search and statistical tools of the SYN2020 corpus are appropriate to detect the ADVERB – PREPOSITION expressions that are presumably used as complex prepositions with spatial meaning. The quantitative data extracted from the corpus serve primarily as an indicator of the fixity, regularity, and non-randomness of these expressions, which allows them to be detected.

Nevertheless, it should be noted that the quantitative data alone are not sufficient to claim that the expressions investigated in the present paper are prepositions. For this purpose, it is necessary to develop a special methodology that will be based on a close qualitative analysis of their actual usage. There are already some developments for recognizing the prepositional function of Czech denominal expressions, see [1] and [5]. However, the specific features of an adverb as a part of speech (the indeclinability, the heterogeneity of this word-class in terms of origin, semantics, syntactic functions, etc.) require specific analysis methods, which have

not yet been developed for the Czech language. The development of such methodology will be a task for further research.

References

[1] Blatná, R. (2006). Víceslovné předložky v současné češtině. Praha: NLN, Nakladatelství Lidové noviny, 351 p.

[2] F. Čermák, J. Hronek and J. Machač (eds.). (1988). Slovník české frazeologie a idiomatiky: Výrazy neslovesné. Praha: Academia, 511 p.

[3] Karlík, P., Nekula, M., Rusínová, Z., and Grepl, M. (2012). Příruční mluvnice češtiny. Praha: NLN, Nakladatelství Lidové noviny, 799 p.

[4] M. Komárek, J. Kořenský, J. Petr and J. Veselková (eds.). (1986). Mluvnice češtiny 2: Tvarosloví. Praha: Academia, 536 p.

[5] Kroupová, L. (1985). Sekundární předložky v současné spisovné češtině. Praha: Ústav pro jazyk český ČSAV, 155 p.

[6] Šuba, P. (1993). Tlumačalny slounik belaruskich prynazounikau. Minsk: Narodnaja asveta, 168 p.

[7] 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., and Škrabal, M. (2020). SYN2020: reprezentativní korpus psané češtiny. Ústav Českého národního korpusu FF UK, Praha. Accessible at: http://www.korpus.cz.

[8] Kroupová, L. (1977). Další sporné případy sekundárních předložek. In Naše řeč, 60(2), pages 68–75.

[9] Čechová, M., Dokulil, M., Hlavsa, Z., Hrbáček, J., and Hrušková, Z. (2011). Čeština – řeč a jazyk. Praha: SPN – pedagogické nakladatelství, 442 p.

[10] Štícha, F. et al. (2018). Velká akademická gramatika současné češtiny. I(1). Praha: Academia, 763 p.

[11] Štícha, F. et al. (2013). Akademická gramatika spisovné češtiny. Praha: Academia, 974 p.

[12] Sláma, J., and Štěpánková, B. (2019). On the Valency of Various Types of Adverbs and Its Lexicographic Description. In Jazykovedný Časopis (Philological Journal) [Online], 70(2), pages 158–169.

[13] Vondráček M. (2020). Výrazy typu *běda* z pohledu slovnědruhového. In Lingvistika – Korpus – Empirie. Praha: Ústav pro jazyk český, pages 93–102.

[14] Vondráček, M. (2018). Slovnědruhové přechody, sporná slovnědruhová klasifikace. In F. Štícha et al., Velká akademická gramatika současné češtiny. I(1). Praha: Academia, pages 100–107.

[15] Vondráček, M. (1999). Příslovce a částice – hranice slovního druhu. In Naše řeč, 82(2), pages 72–78.

[16] Zagnitko, A., Daniluk, I., Sitar, G., and Sukina, I. (2007). Slovnik ukrajinskich prijmennikiv. Sučasna ukrajinska mova. Doneck: TOV VKF "BAO", 416 p.

[17] Kanjuškevič, M. (2008). Belaruskija prynazouniki i ich analahi. Hramatyka realnaha užyvannja. Materyjaly da slounika. Hrodna: HrDU, 492 p.

[18] Vsevolodova, M., Vinogradova, E., and Čaplygina, T. (2018). Russkie predlogi i sredstva predložnogo tipa. Materialy k funkcionalno-grammatičeskomu opisaniju realnogo upotreblenija. Moskva: URSS, 800 p.

[19] Lachur, Cz. (2019). Polskie przyimki wtórne i jednostki o funkcji przyimkowej w użyciu realnym. Materiały do słownika (w zestawieniu z językiem rossijskim). Tom 1. Kępa, 425 p.
[20] Hnátková, M., Jelínek, T., Kopřivová, M., Petkevič, V., Rosen, A., Skoumalová, H., and Vondřička, P. (2019). Lexical database of multiword expressions in Czech. In V. Zakharov (ed.), Trudy meždunarodnoj konferencii "Korpusnaja lingvistika – 2019", St. Petersburg: Saint Petersburg University Press, pages 9–16.
[21] Hnátková, M., Jelínek, T., Kopřivová, M., Petkevič, V., Rosen, A., Skoumalová, H., and Vondřička, P. (2018). Lepší vrabec v hrsti nežli holub na střeše. Víceslovné lexikální jednotky v češtině: typologie a slovník. Korpus – gramatika – axiologie, 9(17), pages 3–22.
[22] Petkevič, V., Kopřivová, M., Hnátková, M., Jelínek, T., Kopřiva, P., Rosen, A., Skoumalová, H., and Vondřička P. (2020). Typologie víceslovných jednotek v češtině a frekvenční zastoupení jejich hlavních vlastností v žánrově vyváženém korpusu. In Studie z aplikované lingvistiky/Studies in Applied Linguistics, 11, pages 37–62.
[23] Hnátková, M., Křen, M., Procházka, P., and Skoumalová, H. (2014). The SYN-series corpora of written Czech. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavík: ELRA, pages 160–164.
[24] A. Čermáková, L. Chlumská and M. Malá (eds.). (2016). Jazykové paralely. Praha: Nakladatelství Lidové noviny, 290 p.
[25] Chlumská, L. (2017). Překladová čeština a její charakteristiky. Praha: Nakladatelství Lidové noviny, 149 p.
[26] Chlumská, L., and Richterová, O. (2014). Překladová čeština v korpusech. In Naše řeč, 97(4–5), pages 259–269.
[27] Chlumská, L. (2013). JEROME: jednojazyčný srovnatelný korpus pro výzkum překladové češtiny. Ústav Českého národního korpusu FF UK, Praha. Accessible at: http://www.korpus.cz.
[28] Richterová, O. (2016). Identifikace posunů ve slovnědruhové příslušnosti: nejen na paralelních korpusech. In A. Čermáková, L. Chlumská and M. Malá (eds.), Jazykové paralely. Praha: Nakladatelství Lidové noviny, pages 95–144.
[29] Čermák, F. (2017). Korpus a korpusová lingvistika. Praha: Univerzita Karlova, nakladatelství Karolinum, 268 p.
[30] Cvrček, V. et al. (2015). Mluvnice současné češtiny 1. Jak se píše a jak se mluví. Praha: Univerzita Karlova v Praze, nakladatelství Karolinum, 416 p.
[31] J. Kocek, M. Kopřivová and K. Kučera (eds.). (2000). Český národní korpus: Úvod a příručka uživatele. Praha: FF UK v Praze, 156 p.