MORPHOLOGICAL DISAMBIGUATION OF MULTIWORD EXPRESSIONS AND ITS IMPACT ON THE DISAMBIGUATION OF THEIR ENVIRONMENT IN A SENTENCE[1]

MILENA HNÁTKOVÁ – VLADIMÍR PETKEVIČ
Faculty of Arts, Charles University, Prague, Czech Republic

HNÁTKOVÁ, Milena – PETKEVIČ, Vladimír: Morphological Disambiguation of Multiword Expressions and Its Impact on the Disambiguation of Their Environment in a Sentence. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 145 – 155.

Abstract: This study concerns the impact of the collocation/phraseme disambiguation component within the complex system of the rule-based morphological disambiguation of Czech. This system constitutes one of the two main disambiguation subsystems that are responsible for the morphological disambiguation of the corpora of synchronic Czech within the Czech National Corpus project. We will show that although the part of texts constituted by collocations/phrasemes (generally multiword expressions – MWEs) is relatively small and consequently the errorfree morphological disambiguation of MWEs covers only a small portion of textual material, such perfectly disambiguated fragments in sentences help to improve the disambiguation of the rest, non-MWE part of sentences.

Keywords: multiword expressions, lexical database, morphological analysis, morphological ambiguity, morphological disambiguation, process of disambiguation, Czech National Corpus

1 INTRODUCTION

The series of corpora of synchronic Czech within the Czech National Corpus, viz. SYN2005, SYN2010, SYN2013PUB, SYN2015, versions of SYN,[2] are morphologically disambiguated by a complex process in which two main components cooperate: the rule-based disambiguation system called *LanGr* ([5], [6], [3], [4], [7], [8], [9]) and the stochastic tagger called *Featurama* (`https://sourceforge.net/projects/featurama/`). This hybrid disambiguation system is activated immediately after morphological analysis: individual morphological homographs are subject to the disambiguation of

(i) lemmas, and
(ii) morphological tags, including part-of-speech tagging.

2 THE DISAMBIGUATION PROCESS

The first disambiguation component, the LanGr system, consists of ca. 2 600 hand-crafted linguistic rules that are

[1] This paper was financially supported by the grant Between Lexicon and Grammar (Mezi slovníkem a gramatikou – Grant Agency of the Czech Republic, reg. no. 16-07473S).

[2] `http://korpus.cz`

DE GRUYTER OPEN

DOI 10.1515/jazcas-2017-0025

(a) developed on the basis of linguistic introspection and checked on corpus data, and also
(b) non-automatically inferred from corpus data.

Linguistic rules are written in a special programming language and their performance consists in the context-based gradual deletion of incorrect lemmas and tags assigned to individual tokens. First, the *LanGr* system processes the output of morphological analysis which assigns every token all of its tags and lemmas; the recall of morphological analysis is currently 99.25%. As the morphological analyzer assigns all tokens all of its lemmas and tags regardless of the context, the tokens are assigned the highest amount of incorrect tags, i.e. the precision is lowest possible on disambiguation input. The disambiguation consists in keeping the best possible recall (close to 100%) and in gradually increasing precision by removing lemmas and tags that are incorrect in the given context.

Disambiguation rules are contained in two main groups:
a) safe rules organized in two subgroups: Safe0 containing entirely safe rules and Safe1 containing slightly less safe rules
b) heuristic rules (Heu).

An input sentence is gradually more and more disambiguated by the rules' application until – ideally – a full disambiguation is achieved, i.e. each token is assigned the only correct lemma and tag. If the rule-based tagger is unable to entirely delete all inappropriate tags and lemmas in the input sentence, the remaining incorrect ones are removed by the second disambiguation component: stochastic tagger Featurama.

The process of the rule-based morphological disambiguation also involves the collocational module *Phras* ([1], [2]), identifying and properly disambiguating multiword expressions (MWEs). Thus, the following modules take part in the disambiguation process:
(i) LanGr tagger based on manually written rules;
(ii) Phras module using a lexical database of maximally disambiguated MWEs;
(iii) parameterizable stochastic tagger, currently Featurama.

The cooperation of the modules consists in the following sequence of operations applied to a sentence:

1733853790 **1st step**: The output of morphological analysis is processed by entirely safe rules (Safe0 group). The rules gradually disambiguate the sentence, i.e. the number of incorrect tags decreases. The process continues till there is nothing to disambiguate, i.e. till the rules in recurrent cycles exhaust their disambiguation capacity.

1733853790 **2nd step**: *Phras* module is invoked: it identifies MWEs in the sentence and performs disambiguation of their components as much as possible.

1733853790 **3rd step**: The set of safe rules Safe0 is reapplied. After these rules finish their job, i.e. they are not able to disambiguate any more, the 1733853790 4th step follows.

1733853790 **4th step**: Three sets of rules, i.e. Safe0, Safe1 and the set of heuristic rules Heu, are applied in cycles to disambiguate the sentence till they cannot disambiguate any more;

1733853790 **5th step**: The remaining incorrect tags intact up to now by the *LanGr* system are removed by the stochastic tagger Featurama and a postprocessing phase (see below).

Table 1 presents a quantitative contribution (in %) of each subsystem within the entire morphological analysis and disambiguation system, where the subsystems are as follows:

Morph – morphological analysis

Safe0 – safe rules: Safe0 (cf. 1st and 1733853790 3rd step above)

Phras – phraseme module Phras processing MWEs (cf. 1733853790 2nd step above)

SSH – the sets of rules Safe0, Safe1 and Heu applied together (cf. 1733853790 4th step above)

Tagger – stochastic tagger Featurama (cf. 1733853790 5th step above)

Post – postprocessing phase (verbal aspect added; possible reinterpretation of controversial part-of-speech annotation, e.g. adverb/particle; finalization of named entities processing...).

after	% of all tokens / incrementally	% of all words / incrementally
Morph	22.07	25.88
Safe0	31.46 / 53,53	36.88 / 62.76
Phras	**0.69 / 54.22**	**0.80 / 63.56**
Safe0	**2.17 / 56.39**	**2.54 / 66.10**
SSH	6.55 / 62.94	7.68 / 73.78
Tagger	22.13 / 85.07	25.94 / 99.72
Post	0.23 / 85.30	0.27 / 99.99

Tab. 1. Contribution of individual subsystems to the entire disambiguation of the texts contained in the sample Newton corpus of journalistic texts (the size in tokens including punctuation marks: 1 735 482 098; words: 1 480 369 445). The contribution is measured merely by the number of achieved unambiguous tags assigned to words after each phase of processing, the quality of disambiguation (recall and precision in the strict sense of the word) is not accounted for here.

In the middle column, the ratio of fully disambiguated tags of words in % after each phase of processing is presented with respect to all tokens (= all corpus positions including punctuation). In the right column, the ratio with respect to word forms only (i.e. without punctuation) is shown. Thus, morphological analysis identifies 22.07% of all tokens and 25.88% of all words as morphologically unambiguous word forms. The Safe0 rule group is able to disambiguate further 31.46% words that were ambiguous after morphological analysis etc. till all words (= 85.30% of all tokens) are unambiguously disambiguated (the rest of the tokens, i.e. 14.70%, is constituted by punctuation tokens). The figures in the right column have the same meaning as in the middle column but they are counted with respect to words only.

Table 2 shows the average number of tags assigned to tokens (words + punctuation marks) after each stage of processing. The figures in the second column mean that the average number of tags assigned to tokens by morphological analysis is almost 11; if punctuation is not taken into account the average number is 12.28, and if only ambiguous words are considered, the average number is more than 16. Table 2 demonstrates the paramount importance of the Safe0 set of rules that is able to decrease the average number of tags assigned by morphological analysis to tokens, words and ambiguous words to 2.81, 3.08 and 5.80, respectively.

after	All tokens counted	only words counted	only ambiguous words counted
Morph	10.62	12.28	16.21
Safe0	2.81	3.08	5.80
Phras	**2.76**	**3.03**	**5.75**
Safe0	**2.54**	**2.77**	**5.40**
SSH	1.92	2.04	4.17

Tab. 2. Average number of tags per word form achieved in the same annotated Newton corpus

Table 1 and Table 2 present, in fact, the measure of precision in a very coarse way since only the ratio of deleted tags in % is shown without taking into account whether only incorrect tags were deleted.

In Table 3 we present the recall after each processing step.

after	recall
Morph	99.25%
Safe0	99.09%
Phras	**99.07%**
SSH	98.82%

Tab. 3. The recall of (i) the morphological analyzer (Morph), (ii) the safe rules (Safe0), (iii) the MWE module (Phras), (iv) Safe0+Safe1+Heu(ristic) rules (SSH)

We see that the recall decreases very slightly: Safe0 rules make only 0.16% errors, the error rate of the MWE module is only 0.02%. The entire rule-base disambiguation system decreases recall after morphological analysis by only 0.43% (99.25 – 98.82).

The accuracy (recall + precision) of the entire disambiguation system, i.e. including the Featurama tagger and the postprocessing phase Post, is ca. 95.1%.

It is to be noted that the disambiguation system described above is not used in syntactic parsing. Stochastic parsers applied to Czech reduce morphological ambiguity to a large extent but the recall they achieve in morphological disambiguation proper is always lower (ca. 93%) than the recall of the system just depicted.

3 THE PHRAS MODULE

Now we will focus on the Phras module disambiguating MWEs in more detail. In Table 1 we see that it contributes to the overall disambiguation success rate only marginally (0.69%). However, if Phras is applied, i.e. if a sentence contains a MWE that is contained in the MWE lexical database exploited by Phras, it paves the way for the Safe0 rules that are able to remove further 2.17% tags thus allowing for further disambiguation. The average number of tags assigned to tokens, words and ambiguous words decreases by 0.05% (cf. Table 2) after the Phras module is invoked.

The performance of the Phras module will be demonstrated on examples (taken from sentences contained primarily in the SYN2015 corpus) showing how Phras

(i) disambiguates MWEs themselves (par. 3.1),

(ii) contributes to the disambiguation of the environment of MWEs in a sentence (par. 3.2).

3.1 Disambiguation of MWEs

Phras exploits the lexical database of fully or partially disambiguated MWEs. The fixed part of these expressions is fully disambiguated, the variable inflectional part is disambiguated only partially, but as much as possible. We will present two motivating examples demonstrating part-of-speech and case disambiguation.

Example 1

In the MWE

(1) *brány pekla*

gates$_{\text{Noun-Npl.Fem/Apl.Fem/Vpl.Fem}}$ of_hell$_{\text{Noun.Gsg.Neut}}$

there is a word form *brány* ‘gates’ 1733853794 that is, morphologically, part-of-speech ambiguous – it is:

(i) genitive singular (Gsg), or nominative/accusative/vocative plural (Npl/Apl/Vpl)[3] of the feminine noun *brána* ‘gate’

(ii) passive participle in feminine plural / masculine inanimate plural of the verb *brát* ‘take’.

In (1), the form *brány* is, however, a part-of-speech unambiguous feminine noun in plural and three cases: Npl/Apl/Vpl since the entirely unambiguous morphological interpretation depends on a textual context.

The other word in (1), *pekla*, is also part-of-speech ambiguous since it is:

(i) Gsg/Npl/Apl/Vpl of the neuter noun *peklo* ‘hell’

(ii) past participle in feminine singular / neuter plural of the verb *péci* ‘bake’.

In (1), the word form *pekla* is unambiguous: Gsg of the neuter noun *peklo.*

Example 2

In the MWE lexical database entry for the MWE,

(2) *ekonomický růst*

economic$_{\text{Adj-Nsg.MascInan/Asg.MascInan}}$ growth$_{\text{Noun-Nsg.MascInan/Asg.MascInan}}$

‘economic growth’

the form *ekonomický* ‘economic’ is a part-of-speech unambiguous adjective in Nsg/Asg masculine inanimate; the part-of-speech ambiguous form *růst* ‘growth / to grow’ is disambiguated as a masculine inanimate noun in Nsg/Asg (‘growth’), rather than the infinitive of the verb (‘to grow’). Moreover, this database entry contains information that both forms agree in number, gender and case.

The Phras module is also very helpful in disambiguating proverbs and other sentential idioms as is shown in the following example.

[3] The remaining cases in the declension system of Czech are: dative (D), locative (L) and instrumental (I).

Example 3

In the process of morphological disambiguation, the proverb:

(3) *Komu není rady, tomu není pomoci*.

To_whom is_not advice$_{\text{Gsg.Fem}}$, to_that$_{\text{DsgMasc}}$ is_not help$_{\text{Gsg.Fem}}$.

'There are none so deaf as those who will not hear'

contained in the MWE lexical database is first processed by the Safe0 rules. They cannot cope with two nouns in the genitive of negation (constructions with the genitive of negation are rare in modern Czech, being associated only with a limited set of nouns), namely *rady*$_{\text{Gsg.Fem}}$ 'advice' and *pomoci*$_{\text{Gsg.Fem}}$ 'help', because the word *rady* and *pomoci* can also be a form of the masculine animate noun *rada* 'counsellor' and the infinitival form of the verb *pomoci* 'to help', respectively. Moreover, the form *tomu* 'to_that' is not only dative singular (Dsg) masculine form of the pronoun *ten* 'that', but also Dsg neuter form of the pronoun *to* 'it'. The collocational module resolves all these ambiguities and entirely disambiguates the proverb.

3.2 Disambiguation of MWEs' Context

On several examples, we will show how disambiguation of MWEs performed by the Phras module can improve disambiguation of their sentential context. These randomly chosen examples are to elucidate the main objectives of the disambiguation of MWEs and of their content: part-of-speech disambiguation, primarily deciding between nouns, verbs and adjectives, and case disambiguation (concerning nouns, adjectives, pronouns and numerals), which is the most difficult subtask of the whole disambiguation process.

Example 4

In sentence:

(4) *Asadův režim* ***nenese odpovědnost*** *za použití zbraní hromadného ničení.*

Asad régime$_{\text{Noun.Nsg.MascInan}}$ **not_bears responsibility**$_{\text{Noun.Asg.Fem}}$ for exploitation of_ weapons of_mass destruction.

'Asad régime **does not bear responsibility** for the exploitation of the weapons of mass destruction.'

Phras identifies the pair *nenese odpovědnost* ('not_bears responsibility') of the MWE ***nést odpovědnost*** 'bear responsibility' and disambiguates its components. The unambiguous present 3$^{\text{rd}}$ person singular negative form *nenese* of the transitive verb *nést* 'bear' poses no disambiguation problem, but the feminine noun *odpovědnost* 'responsibility' can morphologically be Nsg/Asg. The masculine inanimate noun *režim* 'régime' is case ambiguous in the same way. As the disambiguated MWE is contained in the MWE lexical database, Phras unequivocally disambiguates *odpovědnost* in (4) as Asg. The general rules cannot solve, on the basis of sole syntax, the classical disambiguation problem in Czech consisting in the disambiguation of the pattern:

Noun1$_{\text{Nom/Acc}}$ Verb$_{\text{Trans.Pres.3rd.Sg}}$ Noun2$_{\text{Nom/Acc}}$

where either Noun1 and Noun2 is in the nominative and accusative case, respectively, or vice versa.

As Phras disambiguates *odpovědnost* as Asg, it fundamentally helps to disambiguate the sentence: as *odpovědnost* is in Asg, the noun *režim* 'régime' cannot be in non-prepositional Asg (the valency of the verb *nést* does not admit two accusative objects and, moreover, the noun *odpovědnost* cannot head an accusative nominal phrase having the syntactic function of adverbial) and that is why it is in Nsg. After such a correct disambiguation, it is then, e.g., no problem for a parser of Czech to assign proper syntactic functions to the nominal phrase *Asadův režim* ‘Asad régime’ (= subject) and to the nominal phrase *odpovědnost* ‘responsibility’ (= object). Thus the rest of the sentence is also influenced: there are no non-prepositional nouns as objects in accusative[4] in the sentence. In particular, the word *ničení* ‘destruction’ cannot be in non-prepositional accusative. The importance of the disambiguation of the MWE *nést odpovědnost* is thus clearly demonstrated. There are many such support verb (verbo-nominal) constructions in Czech as *nést odpovědnost* and the more such constructions are contained in the MWE lexical database, the better ad more accurate Phras is[5]. The disambiguation of such support verb constructions is of paramount importance especially in cases where some of the collocation components are not only case ambiguous (as in *odpovědnost*) but even part-of-speech ambiguous: e.g. the MWE *nabýt dojmu*$_{\text{Noun.Gsg.MascInan}}$ ‘get an impression’ contains the form *dojmu* that is morphologically Gsg/Dsg/Lsg of the masculine inanimate noun *dojem* ‘impression’, or 1st person singular present tense of the verb *dojmout* ‘impress’; the verbo-nominal construction *má*$_{\text{Trans.Pres.3rd.Sg}}$ *štěstí* (lit. ‘has happiness’ ‘(s)he is lucky’) contains the part-of-speech ambiguous word *má* (morphologically either 3rd person singular present tense of the verb *mít* ‘have’, or Nsg.Fem/Npl.Neut/Apl.Neut/Vpl. Neut of the possessive pronoun *můj* ‘my’; the verbo-nominal construction *svalit vinu*$_{\text{Noun.Asg.Fem}}$ ‘throw the blame (for something on someone)’ contains the part-of-speech ambiguous word *vinu* (morphologically, the form *vinu* is either Asg of the feminine noun *vina* ‘guilt’, or 1733853796 1st person singular present tense of the reflexive verb *vinout se*, ‘to wind’) etc. If such words were erroneously part-of-speech disambiguated, undoubtedly the disambiguation of other words in sentences containing such MWEs would be badly affected.

Example 5

In sentence:

(5) *Manželé přijedou na plzeňské* ***hlavní nádraží*** *parním vlakem.*

Married_couple will_arrive to$_{\text{Prep.Acc}}$ Pilsen$_{\text{Adj.Asg.Neut}}$ main$_{\text{Adj.Asg.Neut}}$ railway_station$_{\text{Noun.Asg.Neut}}$ by steam engine.

‘The married couple will arrive at Pilsen main railway station by steam engine.’

there is a frequent collocation *hlavní nádraží* ‘main railway station’, where the word *hlavní* ‘main’ is an adjective agreeing with the noun *nádraží* ‘railway station’ in number (singular/plural), gender (neuter) and case (nominative/accusative/vocative). However, it can also be a form of the feminine noun *hlaveň* ‘barrel’ in Isg/Gpl. If the

[4] Generally, nominal phrases can also head adverbials of time (duration) or regard in accusative as their attributes, but the set of head nouns governing such adverbials is limited.

[5] Some examples of support verb constructions: *vynést rozsudek* ‘pronounce judgement’, *upírat zrak* ‘fix one’s eyes on someone’, *nabýt dojmu* ‘get an impression’, *mít štěstí* ‘be lucky’, *mít smysl* ‘have sense’, *mít pocit* ‘have a feeling’, *mít právo* ‘be entitled to’, *mít naději* ‘have a hope’, *dávat přednost* ‘have preference (for something)’, *svalit vinu* ‘throw the blame (for something) on someone’, *učinit rozhodnutí* ‘make a decision’...)

disambiguation system chose this nominal interpretation of the word *hlavní* rather than the adjectival one, the disambiguation of the context would be wrong: the prepositional phrase (PP) ***na plzeňské hlavní nádraží*** (lit. ‘to Pilsen main railway station’) would be incorrectly split into three parts:

(i) PP *na*$_{\text{Prep-Acc/Loc}}$ *plzeňské*$_{\text{Adj}}$ ‘to Pilsen’[6]

(ii) feminine noun *hlavní* ‘barrel’ that is morphologically in Isg/Gpl and does not agree with the PP *na plzeňské* in case since the preposition *na* generally requires accusative/locative and the adjective *plzeňské* morphologically is, i.a., Gsg.Fem/Dsg.Fem/Lsg.Fem/Npl.Fem/Apl.Fem/Vpl.Fem/Nsg.Neut/Asg. Neut/Vsg.Neut

(iii) neuter noun *nádraží* ‘railway station’ that is morphologically Nsg/Gsg/Dsg/ Asg/Vsg/Lsg/Npl/Gpl/Apl/Vpl.

Thus, the rule-based system would not know how to disambiguate the ambiguous PP *na plzeňské*. Moreover, the case and number of the neuter noun *nádraží* could hardly be identified. If, on the contrary, Phras disambiguates *hlavní nádraží* as a collocation, i.e. *hlavní* as an adjective ‘main’ coforming a nominal phrase with the noun *nádraží* (the adjective *hlavní* agrees with the noun *nádraží* in number, gender and case), the subsequently applied rules can assume that the sequence ***na plzeňské hlavní nádraží*** complies with the PP pattern:

Prep$_{\text{Acc}}$ Adj$_{\text{Asg.Neut}}$ Adj$_{\text{Asg.Neut}}$ Noun$_{\text{Asg.Neut}}$

This means that the rules can recognize the sequence as one PP in Asg.[7]

Example 6

In sentence:

(6) *Prioritou je* ***zajištění odbytu****.*

Priority is$_{\text{Verb.Pres.3rd.Sg}}$ securing$_{\text{Noun.Nsg.Neut}}$ of_sales$_{\text{Noun.Gsg.MascInan}}$.

‘The priority is the securing of sales.’

there are three part-of-speech ambiguous words:

(a) *je* is:

(i) 3rd person singular present tense form of the verb *být* ‘be’

(ii) Asg of the 3rd person neuter personal pronoun *ono* ‘it’

(iii) Apl of the 3rd personal pronoun (all genders) *oni/ony/ona* ‘they’;

(b) *zajištění* is:

(i) Npl.MascAnim/Vpl.MascAnim form of the adjective *zajištěný* ‘secured’

(ii) Nsg/Gsg/Dsg/Asg/Vsg/Lsg… of the deverbal neuter noun *zajištění* ‘securing’;

(c) *odbytu* is:

(i) Gsg/Dsg/Lsg of the masculine inanimate noun *odbyt* ‘sales’

(ii) passive participle of the transitive verb *odbýt* ‘do sloppily’ in Asg.Fem.

General disambiguation rules can hardly correctly disambiguate the nominal phrase (NP) *zajištění odbytu* ‘securing of sales’ as well as its immediate context: the

[6] For simplicity reasons, we omit the interpretation of the word form *plzeňské* as a (deadjectival) noun.

[7] They can, moreover, disambiguate *plzeňské* as an adjective rather than as a deadjectival noun (univerbization: *plzeňské pivo* ‘Pilsner beer’ → *plzeňské* ‘Pilsner’).

word *je*. The Phras module identifies the pair *zajištění odbytu* as a collocation where *zajištění* is disambiguated as a noun rather than as an adjective, and *odbytu* as a noun in Gsg rather than as a verbal passive form. For the subsequent rules it will then be much easier to identify the entire NP *zajištění odbytu* ‘securing of sales’ as an NP where *zajištění* is in Nsg, and also *je* as a verbal predicate in singular rather than as a personal pronoun.

Example 7

The sentence:
(7) *Nasypala dovnitř* ***prací prášek****.*
She poured inside$_{\text{Adv}}$ washing$_{\text{Adj.Asg.MascInan}}$ powder$_{\text{Noun.Asg.MascInan}}$.
‘She poured inside the washing powder.’
is difficult to disambiguate since there are two part-of-speech ambiguous forms (*dovnitř* ‘inside / into’, *prací* ‘washing / of-works’), and especially in such structures the processing of collocations can be very helpful. The word *dovnitř* is either (i) a preposition (‘into’) requiring genitive, or (ii) an adverb (‘inside’); the word *prací* is either (i) Isg/Gpl form of the feminine noun *práce* ‘work’, or (ii) a very number-case ambiguous form of the soft adjective (the root ending in *-í*) *prací* ‘washing’. The rules could incorrectly identify the form *prací* as Noun.Gpl.Fem ‘work’, and the word form *dovnitř* as a preposition taking genitive and thus the pair *dovnitř prací* could be identified as a genitive prepositional phrase with the meaning ‘into the works’. If, on the contrary, Phras correctly identifies the pair *prací prášek* as a noun phrase (‘washing powder’) contained in the lexical database where *prací* is an adjective agreeing with the noun *prášek* in number (= singular), gender (= masculine inanimate) and case (= nominative/accusative), the subsequent rules will exclude *dovnitř* as a preposition taking genitive, thus interpreting *dovnitř* only as an adverb.

Example 8

The word *místo* is ambiguous between a noun ‘place’ and a preposition ‘instead of’. A correct disambiguation of this very frequent word is crucial for the errorfree disambiguation of clauses where the word *místo* appears. The main disambiguation problem consists in that the noun *místo* often collocates with a NP in the genitive case and the preposition *místo* takes genitive, too. If typical collocations with the noun *místo* are contained in the MWE lexical database exploited by the Phras module, the disambiguation of sentences containing such collocations is much better. For instance, in sentence
(8) *Policie obrátila* ***místo činu*** *vzhůru nohama*.
Police$_{\text{Noun.Nsg.Fem}}$ reversed place$_{\text{Noun.Asg.Neut}}$ of_crime$_{\text{Noun.Gsg.MascInan}}$ upwards with legs.
‘The police put the scene of crime out of joint.’
the Phras module identifies *místo* as a noun rather than as a preposition since it uses a partially disambiguated collocation *místo činu* (lit. ‘place of crime’, ‘scene of crime’) contained in the lexical database: *místo*$_{\text{Noun-Nsg.Neut/Asg.Neut/Vsg.Neut}}$ *činu*$_{\text{Noun.Gsg.MascInan}}$, where the components have disambiguated morphological properties as indicated. In sentence (8), the noun *místo* is unambiguously in Asg since it does not agree with the feminine singular predicate *obrátila* ‘reversed’ in gender and therefore

it cannot be the subject in the nominative case. As *místo* is in accusative, *policie* ‘police’ cannot be in accusative (the verb *obrátila* cannot take two objects in accusative), it can only be in Nsg (correct), or Gsg/Vsg (incorrect), or Npl/Apl/Vpl (incorrect). If *místo* were erroneously identified as a preposition, the accusative reading of the form *policie* could not be syntactically excluded.

Most frequent right nominal collocations with the noun *místo* are as follows: *místo určení*[8] *‘*destination’, *místo činu ‘*scene of crime’, *místo nehody* ‘accident site’, *místo konání ‘*venue’, místo *narození* ‘place of birth’, *místo nálezu* ‘place of finding’, *místo spolujezdce* ‘passenger seat’, *místo odpočinku* ‘resting place’. Such collocations contained in the lexical database and exploited by the Phras module often help to disambiguate the context of these collocations.

Example 9
In sentence:
(9) *Řeč je o dobrovolných* ***dárcích krve***.
Talk is about voluntary$_{\text{Adj.Lpl.MascAnim}}$ donors$_{\text{Noun.Lpl.MascAnim}}$ of blood$_{\text{Noun.Gsg.Fem.}}$
‘Voluntary blood donors are being talked about.’
the Phras module identifies *dárcích* as Lpl of the masculine animate noun *dárce* ‘donor’ since it is a component of the MWE *dárce*$_{\text{Noun.MascAnim}}$ *krve*$_{\text{Noun.Gsg.Fem}}$ ‘blood donor’. In the nominal Lpl phrase *o dobrovolných dárcích krve* ‘about voluntary donors of blood’ the adjective *dobrovolných* ‘voluntary’ is also in Lpl.MascAnim (due to agreement with *dárcích* in number, gender and case). However, the form *dárcích* is, morphologically, also Lpl form of the masculine inanimate noun *dárek* ‘present’. Without knowing the existence of the MWE *dárce krve* the rules could erroneously disambiguate and lemmatize the form *dárcích* as a Lpl.MascInan form of *dárek*. If so, the form *dobrovolných* ‘voluntary’ would then be erroneously also disambiguated as Lpl.MascInan (due to agreement). Moreover, the morphologically ambiguous form *krve* is correctly disambiguated as Gsg.

4 CONCLUSION

In the paper, we have demonstrated the significance of MWEs’ morphological disambiguation – performed by a special Phras module on the basis of a lexical database containing (partially) disambiguated MWEs – for the successful disambiguation of the other, non-MWE parts of sentences containing MWEs, the disambiguation being performed by subsequent disambiguation rules. Further work will consist in improving the collaboration of the Phras module with the general rules as to the division of labour: which MWEs are to be processed by general rules and which should be included in the lexical database and processed by the Phras module. Furthermore, the database will constantly be enhanced.

[8] The neuter nominal forms *určení* (‘destination’) and *narození* (‘birth’) in bold are, generally, part-of-speech ambiguous: they are also adjectives (‘determined’ and ‘born’) in Npl.MascAnim/Vpl.MascAnim.

References

[1] Hnátková, M. (2006). Typy a povaha komponentů neslovesných frazémů z hlediska lexikálního obsazení (Types and nature of components of non-verbal phrasemes from the viewpoint of lexical elements). In *Kolokace*. *Studie z korpusové lingvistiky*, pages 142–167, Nakladatelství Lidové noviny – Ústav Českého národního korpusu, Praha.

[2] Hnátková, M. and Kopřivová, M. (2014). From Dictionary to Corpus. In Jesenšek, V. and Grzybek, P., editors, *Phraseology in Dictionaries and Corpora, ZORA 97*, pages 155–168, Maribor, Slovenia.

[3] Jelínek, T. (2008). Nové značkování v Českém národním korpusu (New annotation in the Czech National Corpus). *Naše řeč*, 91(1):13–20.

[4] Jelínek, T. and Petkevič, V. (2011). Systém jazykového značkování současné psané češtiny (The system of linguistic annotation of contemporary written Czech). In *Korpusová lingvistika Praha 2011, sv. 3: Gramatika a značkování korpusů*, pages 154–170, Nakladatelství Lidové noviny / Ústav Českého národního korpusu, Praha, Czech Republic.

[5] Květoň, P. (2006). *Rule-Based Morphological Disambiguation (Towards a Combination of Linguistic and Stochastic Methods)*. PhD thesis. MFF UK, Praha.

[6] Petkevič, V. (2006). Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary. In Šimková, M., editor, *Insight into the Slovak and Czech Corpus Linguistics*, pages 26–44, Veda (Publishing House of the Slovak Academy of Sciences & Ludovít Štúr Institute of Linguistics of the Slovak Academy of Sciences), Bratislava, Slovakia.

[7] Petkevič, V. (2014). Problémy automatické morfologické disambiguace češtiny. *Naše řeč*, 97(4–5):194–207.

[8] Petkevič, V. (2014). Ambiguity, language structures and corpora. In *La linguistique* (coordonné par Radimský, J. and Pešek, O., editors, Le Cercle linguistique de Prague – II. D’hier à aujourd’hui), vol. 50, 2014-2, pages 63–82, Presses Universitaires de France.

[9] Petkevič V. (2014). *Morfologická homonymie v současné češtině*. Nakladatelství Lidové noviny – Ústav Českého národního korpusu, Praha.