Difference between written and spoken Czech:

Koláová, Veronika; Kolá, Jan; Mikulová, Marie

DOI:
10.1515/pralin-2017-0002

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Koláová, V, Kolar, J & Mikulová, M 2017, 'Difference between written and spoken Czech: The case of verbal nouns denoting an action', Prague Bulletin of Mathematical Linguistics, vol. 107, no. 1, pp. 19-38. https://doi.org/10.1515/pralin-2017-0002

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 107 APRIL 2017 19–38

Difference between Written and Spoken Czech: The Case of Verbal Nouns Denoting an Action

Veronika Kolářová,[a] Jan Kolář,[b] Marie Mikulová[a]

[a] Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
[b] Institute of Mathematics of the Czech Academy of Sciences

Abstract

The present paper extends understanding of differences in expressing actions by verbal nouns in corpora of written vs. spoken Czech, namely in the Czech part of the Prague Czech-English Dependency Treebank and in the Prague Dependency Treebank of Spoken Czech.

We show that while the written corpus includes more complex noun phrases with more explicit expression of adnominal participants, noun phrases in the spoken corpus contain more deletions and more exophoric references. We also carried out a quantitative analysis focusing on relative frequencies of combinations of participants modifying verbal nouns; although the written corpus shows higher relative frequencies, the order of the relative frequencies of particular combinations is the same in both types of communication.

1. Introduction

Differences between written and spoken language have come under scrutiny in linguistic research in English and other languages including Czech for decades (e.g., Halliday, 1967; Hausenblas, 1962; Chafe and Danielewicz, 1987). Though older studies were based on authentic spoken examples or even on collections of spoken texts compiled for the particular purpose, a new dimension of research of spoken language has been added by the development of large corpora of spoken communication. From the linguistic point of view, however, only few of the resources are POS tagged and/or lemmatized, or even include syntactic annotation. The Prague Dependency Treebank of Spoken Czech (which is the resource for our research, see Section 3), the Switchboard corpus in the Penn Treebank-3 (Marcus et al., 1999), Childes Database

 Corresponding author: kolarova@ufal.mff.cuni.cz
Cite as: Veronika Kolářová, Jan Kolář, Marie Mikulová. Difference between Written and Spoken Czech: The Case of Verbal Nouns Denoting an Action. The Prague Bulletin of Mathematical Linguistics No. 107, 2017, pp. 19–38.
doi: 10.1515/pralin-2017-0002.

	Written communication	**Spoken communication**
Expression	condensed / complex / intricate sentences	analytical / unelaborated flow of speech
Specific means of expression	hypotaxis, nominalisations	parataxis, repetitions, restarts, corrections, disfluencies
Segmentation	strict / clear boundaries between sentences	unclear sentence segmentation, juxtaposition
Deletions / ellipses	deletions / ellipses with context-dependent references	incompleteness, fragmentation, interruptions, extra-textual (exophoric) references
Valency	refinement of forms of participants	marked participants and forms

Table 1. Main differences in syntax between written and spoken communication

(MacWhinney, 2000; Sagae et al., 2004), or Corpus Gesproken Nederlands (Schuurman et al., 2003) are among the few exceptions to this rule.

In spite of this situation, the data-based research in various aspects of spoken language has recently become one of the central topics (e.g., Biber et al., 1999; Brazil, 1995; Roberts and Street, 1997; Leech, 2000). Within the spoken Czech research, which mostly happens on the national scene, real texts and dialogues are analysed and presented (e.g., Těšitelová, 1983), with focus on the specificity of the spoken word form (e.g., Šonková, 2008; Cvrček et al., 2010), spoken syntax (e.g., Müllerová, 1994; Hoffmannová, 2012; Mikulová and Hoffmannová, 2011), issues of valency (e.g., Mikulová et al., 2013) and the specificity of the social issues concerning the speakers and situations in which the analysed utterances were used (e.g., Hoffmannová et al., 1999; Hoffmannová and Müllerová, 2007; Čmejrková et al., 2004). Despite the numerous studies, the description of syntax of spoken Czech is still not as developed, consistent and comprehensive as the description of written Czech. This article aims at a description of differences between the written and spoken syntax, focusing on a special case of action-denoting verbal nouns in corpora of written and spoken Czech.

2. Differences between written and spoken language

On the basis of the studies mentioned in Section 1, we summarize the main differences in syntax between both types of communication (see also Table 1).

A prominent feature of spoken language besides its acoustic nature is its anchoring in the time and situation. The conditions of spontaneous speech production (presence of the addressee, speaking skills of the speakers, importance of non-verbal communication) lead to numerous repetitions, incomplete sentences, corrections and interruptions. Presence of the addressee, context and knowledge shared by the speakers play an important role, so it is possible to leave much unsaid or indirectly implied in the spontaneous speech. On the other side, writers receive no immediate feedback from their readers so there is more need to explain things clearly and unambiguously than in speech. Using longer sentences and many subordinate clauses, written texts are usually more complex and intricate than speech.

Verbal nouns denoting an action belong to nominalisations and they can be understood as reclassifications of their corresponding verbal clauses (Heyvaert, 2003). They help to form compact and condensed expression and when they are modified by their participants, they constitute complex noun phrases.[1] We suppose that a written text includes more complex noun phrases with more explicit expression and that noun phrases in spontaneous speech contain more deletions, using more exophoric references. We test this hypothesis in corpora of written and spoken Czech containing deep syntactic annotation (see Section 3).

3. Data: Corpora of written and spoken Czech with deep syntactic annotation

The syntactic behaviour of Czech verbal nouns can be studied most effectively in syntactically annotated corpora. One of the features characteristic of syntax of spoken language is its incompleteness and fragmentation (Hunyadi, 2013). We are convinced that the unexpressed elements should become an important part of the research of the differences between written and spoken communication. However, as elements that are not present on the surface layer of a sentence, their reconstruction relies on a theoretical framework that deals with the deep structure of sentences and therefore they are only exceptionally captured even in syntactically annotated corpora. In order to be able to search for the unexpressed elements in the syntactic structure of written and spoken communication, we use manually syntactically annotated corpora built within the theoretical framework of Functional Generative Description (FGD, Sgall et al., 1986) because they capture also deletions (see Section 4).

Further aspect we considered when selecting resources that best meet our requirements is the need of comparable data. The data should be comparable especially in its annotation scheme. Thus we chose two corpora from the Prague Dependency Treebank family which have comparable size and, moreover, which apply the same guidelines for annotation of valency of verbal nouns. These two corpora are (i) the

[1] In this paper, we use the term noun phrase for nominal constructions in which a noun is modified by its dependents, focusing on verbal nouns modified by their participants.

	PCEDT (written corpus)	PDTSC (spoken corpus)
Tokens	1 162 072	742 257
Sentences	49 208	73 835
Words per a sentence	23.6	10.1
Content verbs	99 186	102 868

Table 2. Comparison of the size of the used written and spoken corpora

Prague Dependency Treebank of Spoken Czech (PDTSC), and (ii) the Czech part of the Prague Czech-English Dependency Treebank (PCEDT). The unique opportunity of having a spoken and written resource with a comparable annotation enables us to carry out precise and reliable analysis of selected differences between the two types of communication.

(i) **The Prague Dependency Treebank of Spoken Czech 2.0** (PDTSC) is the upcoming release (planned to be published in 2017; Mikulová et al., in press).[2] The corpus offers a huge, unique material for a systematic analysis of syntax of spoken Czech on higher levels of linguistic abstraction, including deep syntactic annotation. PDTSC recordings consist of two types of dialogues. First, it contains slightly moderated testimonies of Holocaust survivors from the Malach project corpus.[3] The second part of the corpus consists of dialogues recorded for the Companions project.[4] It contains personal memories, but in a setting where the two dialogue participants chat over a collection of photographs. The spoken material is manually transcribed, edited for disfluencies, and then annotated syntactically (on the layer of surface syntax and deep syntax) while keeping the original transcript explicitly aligned with the edited version. This allows the morphological, syntactic and semantic annotation to be deterministically and fully mapped back to the transcript and audio. The PDTSC consists of 742 257 tokens and 73 835 sentences, representing 6 174 minutes of spontaneous dialogue speech.

(ii) **The Prague Czech-English Dependency Treebank 2.0** (PCEDT, Hajič et al., 2012) is a manually parsed Czech-English parallel corpus of 1.2 million tokens in 49 208 sentences for each language. The English part holds the Wall Street Journal (WSJ) section of the Penn Treebank (Marcus et al., 1999). The Czech part was translated from the English source sentence by sentence.

[2]The results of our search can differ from the future published corpus but the differences will be insignificant (because there are not a lot of substantial changes in the data now).

[3]`http://malach.umiacs.umd.edu/`

[4]`http://www.companions-project.org`

Table 2 shows that these two corpora are comparable in size (cf. the number of tokens) but it also reflects some differences between the written and spoken communication, especially the difference in number of sentences and their length (there are more sentences in the spoken corpus but the sentences are shorter on average than in the written corpus). Searching through the two corpora is carried out by the tool called PML-TQ (Štěpánek and Pajas, 2010).

We are aware that the two corpora are not representative samples of written and spoken communication. The Czech part of the PCEDT is a translation rather than an original written text, moreover it focuses on a very specialized semantic domain concerning trading. Neither our spoken corpus, PDTSC, contains completely spontaneous (i.e., unprepared) spoken production. However, despite these deficiencies, we take advantage of their deep syntactic annotation and show that even such imperfect samples of written and spoken communication reflect significant differences in expressing an action by verbal nouns. On the basis of the two corpora, we can specify the following distinctive features of denoting an action by verbal nouns in Czech written and spoken communication: degree of compact and condensed expression measured by frequency of verbal nouns (Section 6.1), noun phrase complexity (Section 6.2), degree of explicitness (Section 6.3), and finally deletions and exophoric references (Section 6.4).

4. Prague Dependency Treebank family: Annotation scheme

4.1. Valency

As mentioned above, one of the important features of the PDT-style annotation (Hajič et al., in press) is the fact that in addition to the morphological layer and to the syntactic annotation of the surface shape of the sentences the scenario includes a complex semantically based annotation on the highest, deep syntax layer (so-called tectogrammatical layer). The core component in the annotation is valency and one of the important features is the reconstruction of surface deletions on the tectogrammatical layer (the annotation guidelines are formulated in Mikulová et al., 2006; Mikulová, 2014). The valency theory for the theoretical framework of the FGD was formulated by Panevová (1974, 1975) and it has been detailed in numerous studies addressing especially valency of verbs (Panevová, 1998, 1999, 2014) and nouns (Piťha, 1980; Panevová, 2002; Kolářová, 2014). The following types of complementations (i.e. the individual dependency relations) are able to fill in the individual slots of the valency frames of verbs:

– inner participants or arguments that can be obligatory or optional: Actor, Patient, Addressee, Effect, Origin (e.g., *Vláda omezila těžbu uranu ze současných 950 tun na 500 tun ročně* 'The government restricted uranium mining from the current 950 tonnes to 500 tonnes per year');

– obligatory free modifications or adjuncts, especially those with the meaning of direction (e.g., *přijet někam* 'to arrive somewhere') or location (e.g., *přebývat někde* 'to dwell somewhere') and manner (e.g., *chovat se dobře* 'to behave well').

Within the concept of nominal valency in the framework of the FGD, the same repertoire of valency complementations is assumed for deverbal nouns denoting an action. The repertoire of valency complementations of non-deverbal nouns and deverbal nouns undergoing substantial shifts in their meaning is supplemented with some more modifications, especially with a special nominal participant Material (e.g., *skupina lidí* 'group of people', *jedno balení másla* 'one package of butter') and a free modification Appurtenance (e.g., *Petrovo auto* 'Peter's car', *oddělení odbytu* 'sales department').

The valency theory was applied to the PDT-corpora data which resulted in a very complex and detailed annotation scheme. Different meanings of words with valency that occur in the data are differentiated in a valency lexicon called PDT-Vallex[5] (Hajič et al., 2003; Urešová, 2012). Each PDT-Vallex entry describes a lexeme (represented by the "lemma") and its valency frame(s). One valency frame typically corresponds to one meaning (sense) of a word (i.e., a verb, a noun, an adjective, or an adverb). Although PDT-Vallex does not explicitly work with the term lexical unit, a meaning of a word with its particular valency frame corresponds to a lexical unit, understood roughly as 'a given word in a given sense' (Cruse, 1986). In PDT-Vallex, a valency frame encodes the core valency information, listing possible alternative forms of valency complementations and giving information about semantic roles, i.e., deep functions in terms of tectogrammatical functors of the FGD, esp. ACT for Actor, PAT for Patient, ADDR for Addressee, ORIG for Origin or EFF for Effect. Moreover, information about obligatoriness is assigned to each participant (optional participants are marked with the sign '?', see (1) and (2) in Section 5) and it is reflected in the deep structure of sentences in which the respective noun occurs as follows: nodes for valency complementations that are obligatory and thus present in the deep structure of the sentence are added into the data even though they are not present on the surface layer of the sentence. This is exactly where ellipsis meets valency: an unexpressed obligatory participant or free modification is treated as a surface deletion (valency ellipsis) and it is captured by adding a node to the tree (for more details concerning coreference types see Section 4.2). Nodes added for obligatory complementations that are not present on the surface layer of the sentence enable us to search for the unexpressed elements which we consider crucial for our research into the differences between written and spoken communication.

To summarize, the annotation of valency in the PDT-corpora consists of:

– determining and assigning a valency frame from PDT-Vallex;
– a corresponding semantic role (ACT, PAT, ADDR, etc.) is assigned to the nodes for valency complementations expressed on surface;

[5] http://hdl.handle.net/11858/00-097C-0000-0023-4338-F

– obligatory valency complementations unexpressed on the surface are captured by an added (newly created) node with an artificial lemma (for example #PersPron), and the corresponding semantic role is also assigned.

4.2. Coreference relations

The PDT-style annotation also captures various types of (co)reference relations. For each participant (when not expressed on the surface, it is captured by an added node, see Section 4.1), the annotator determines whether the participant has its antecedent in the text (core coreference relations) or whether there is a reference to a situation or reality external to the text (exophoric references), or whether there is no (co)reference. Within the core coreference relations, the two following types are distinguished: grammatical coreference (in which it is possible to pinpoint the antecedent according to grammatical rules) and textual coreference (where reference is determined not only by grammatical means, but also via context). (Co)reference relations are annotated in the case of personal and possessive pronouns, demonstrative pronouns *ten, ta, to* 'this/that', and in the case of unexpressed obligatory participants. Various types of coreference relations are captured by assignments of artificial lemmas of various types and via various types of coreference arrows from the participant (coreferring node) to its antecedent. An exophoric reference is represented as a short arrow pointing upwards (for more details see Zikánová et al., 2015).

Depending on the type of (co)reference relations, we distinguish the following types of obligatory participants of verbal nouns denoting an action (some of them are addressed in Section 6.3 and Section 6.4):

(a) a participant expressed by a noun, a possessive adjective, a prepositional phrase, a content clause or an infinitive (e.g., *Petrovo*.ACT *pití čaje*.PAT 'Peter's drinking of tea')

(b) a participant expressed by a pronoun (e.g., *jeho*.ACT *pití toho*.PAT 'his drinking of that') with one of the following types of (co)reference:
  (ba) grammatical coreference
  (bb) textual coreference
  (bc) exophoric reference
  (bd) no reference (in the case of idioms, e.g., *mít své opodstatnění*, lit. *to have its justification*, i.e., 'be justifiable')

(c) a participant unexpressed on the surface (e.g., *pití* 'drinking') with one of the following types of (co)reference:
  (ca) grammatical coreference
  (cb) textual coreference
  (cc) exophoric reference
  (cd) no reference (in the case of the so-called general participant, e.g., *tuk na smažení* 'frying fat').

5. Verbal nouns denoting an action

In this paper, we concentrate on expressing an action by Czech verbal nouns which are derived from verbs by productive means (suffixes *-(e)ní/tí*, as in *vykládání* 'explaining//unloading' or *pojetí* 'conception'). We do not consider another type of Czech deverbal nouns that also in some of their meanings denote an action, i.e., nouns derived from verbs by non-productive means including the zero suffix (such as *vykládka* 'unloading', *výklad* 'explanation/interpretation'). There are three reasons for working only with verbal nouns (i.e., with the productively derived nouns) in this study:

(i) They often have a meaning denoting an action;
(ii) All of them can be found in the data thanks to their unique suffixes *-(e)ní/tí*;
(iii) Their valency is annotated according to the same guidelines in both selected corpora.

We suppose that all verbal nouns denoting an action have an obligatory Actor (ACT). Nouns denoting an abstract result of an action usually also have an Actor but it might be optional rather than obligatory as illustrated by the examples of valency frames of the noun *omezení* 'restricting/restriction' from PDT-Vallex, see (1) for denoting an action of restricting, and (2) for an abstract result of the action, i.e., restriction, with an optional Actor. Verbal nouns denoting a thing do not have an Actor in their valency frame at all, cf. two meanings of the noun *pití* 'drinking/drink' in (3) and (4).

(1) *omezení* 'restricting'
ACT(Gen,Ins,Poss) PAT(Gen,Poss) ?ORIG(*z* 'from' + Gen) ?EFF(*na* 'to' + Acc)
postupné omezení těžby.PAT *uranu ze současných 950 tun*.ORIG *na 500 tun*.EFF *ročně*
'gradual restricting of uranium mining from the current 950 tonnes to 500 tonnes per year'
(2) *omezení* 'restriction'
?ACT(Gen,Poss) ?PAT(*proti* 'on' + Dat)
omezení vlády.ACT *proti exportérům*.PAT 'restriction of the government on exporters'
(3) *pití čaje*.PAT *Petrem*.ACT 'drinking tea by Peter'
(4) *tvrdé pití* 'strong drink'

Therefore, we assume that the best way to find all verbal nouns denoting an action in our data is to search for verbal nouns that are in the data modified by an Actor (either present on the surface or added as an unexpressed but obligatory element), see (5) for the query specified in PML-TQ. Using this method, we get all the nouns denoting an action. We also get occurrences of nouns denoting an abstract result of an action in which the Actor is expressed on surface, however we believe this fact has a negligible impact on the results of our inquiry. The numbers of found verbal nouns are given in Section 6.1.

(5) An example of a query specified in PML-TQ: searching for verbal nouns denoting an action

```
t-node $a :=
[ t_lemma ~ "^.*[nt]í([_-].*)?$",
  t-node  [ functor = "ACT" ],
  a/lex.rf a-node  [ tag ~ "^N.N.*$" ] ];
  >> distinct $a
  >> give count()
```

6. Action-denoting verbal nouns in corpora of written and spoken Czech

In this section, we present the results of our search in the data of the PCEDT and PDTSC. We analyze and compare frequencies of phenomena outlined in Sections 2 and 3 which are believed to differentiate between written and spoken communication (especially frequency of verbal nouns and noun phrase complexity, see Sections 6.1 and 6.2). Exploitation of the rich and elaborate annotation scheme of the PDT corpora and use of the powerful searching tool PML-TQ enables us to particularize the distinctive features and even introduce more detailed characteristics of written and spoken communication, especially in the domain of coreference (Sections 6.3 and 6.4).

6.1. Frequency of verbal nouns and their semantic domain

Table 3 shows that while the number of occurrences of content verbs is similar in both corpora, verbal nouns are considerably more frequent in the written corpus than in the spoken one (cf. 1595 lemmas with 16283 occurrences in the PCEDT vs. only 501 lemmas with 1359 occurrences in the PDTSC). We interpret the higher number of verbal nouns denoting an action in the PCEDT as a manifestation of more condensed expression which is characteristic of written communication in general.

Given the subject matters of the texts of the used corpora (see Section 3), lexical meanings of the most frequent lemmas of verbal nouns occurring in the PCEDT and PDTSC belong to different semantic domains (see Table 3; the most frequent verbal noun that occurs in both corpora is the noun *rozhodnutí* 'decision'). Capturing personal memories and testimonies, the spoken corpus describes actions in everyday life such as exercising, meeting, having a swim, skiing, birth, travelling, and learning. The main semantic domain of the written corpus (as determined by the texts included in the corpus) is trading and related actions such as increase, reduction, taking over, making decision, negotiation, financing.

6.2. Noun phrase complexity

We carried out a quantitative analysis of combinations of participants modifying verbal nouns in both corpora. Table 4 gives relative frequencies of combinations of

Phenomenon	**Occurrences in PCEDT (written corpus)**	**Occurrences in PDTSC (spoken corpus)**
Verbal nouns denoting an action	16 283	1 359
Content verbs	99 186	102 868
Number of verbs per 1 verbal noun	6.1	75.7
Semantic domain	Trading	Activities of everyday life
The most frequent lemmas of verbal nouns	obchodování (1 323), zvýšení (590), snížení (458), převzetí (437), **rozhodnutí (357)**, jednání (325), financování (258), prohlášení (221), očekávání (190), pojištění (181), zdanění (179), omezení (167), řízení (159), obvinění (152), uzavření (147), podnikání (136), hlasování (122), získání (121), oznámení (120), zlepšení (120), snižování (113), …	cvičení (73), setkání (44), koupání (37), lyžování (33), narození (32), cestování (23), učení (21), plavání (20), vaření (17), přijímání (16), povídání (14), posezení (14), osvobození (13), vítání (12), pití (11), vyprávění (11), čtení (11), fotografování (10), psaní (10), bydlení (9), hraní (9), přání (9), tancování (9), hlídání (8),vyučení (8), stravování (8), **rozhodnutí (8)**, …
Number of lemmas of verbal nouns	1 595	501

Table 3. Frequency of verbal nouns and their semantic domain

Combinations of participants expressed on surface	**PCEDT (written corpus)** Occurrences abs.	rel.	**PDTSC (spoken corpus)** Occurrences abs.	rel.
PAT	7 003	43 %	254	18 %
ACT	1 606	10 %	101	7 %
ACT + PAT	363	2 %	7	0.5 %
PAT + ADDR	126	0.8 %	2	0.2 %
0 expr. participants	6 860	42 %	996	73 %
Other combinations	325	2 %	5	0.4 %

Table 4. Combinations of (semantic roles of) participants expressed on surface

Number of expressed participants	**PCEDT** Occurrences		**PDTSC** Occurrences	
0	6 860	42 %	996	73 %
1	8 817	54 %	359	26 %
2	585	3.6 %	10	0.7 %
3	20	0.1 %	0	0 %
4	1	0.01 %	0	0 %

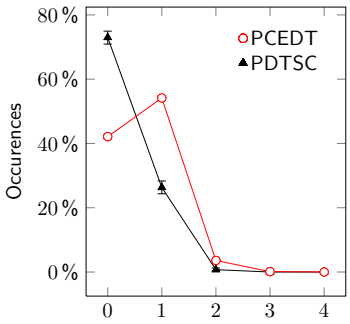


Table 5. Number of participants expressed on surface

participants expressed on the surface by any form, reflecting semantic roles of the participants. We can see that although the written corpus shows higher relative frequencies, the order of the relative frequencies of particular combinations is the same in both types of communication. The most frequent combination is the case when only Patient is expressed. The case when only Actor is expressed is the second most frequent combination, followed by the combinations Actor + Patient or Patient + Addressee, the latter of which is applicable only in the case of nouns that have an Addressee in their valency frame.[6]

Table 5 reflects the same data but focuses on the number of participants expressed on surface regardless of their semantic role. We can see that more complex noun

[6]This order of relative frequencies seems to be of general validity; for the case of verbal nouns representing five different semantic classes in the data obtained from the Prague Dependency Treebank 3.0 see (Kolářová, in press).

Phenomenon	Occurrences in PCEDT	Occurrences in PDTSC
Verbal nouns denoting an action	16 283	1 359
Chains of two verbal nouns (N_1 modified by N_2)	348 2.14 %	2 0.15 %
Chains of three verbal nouns (N_1 modified by N_2 with N_2 modified by N_3)	8	0

Table 6. Cumulation of verbal nouns

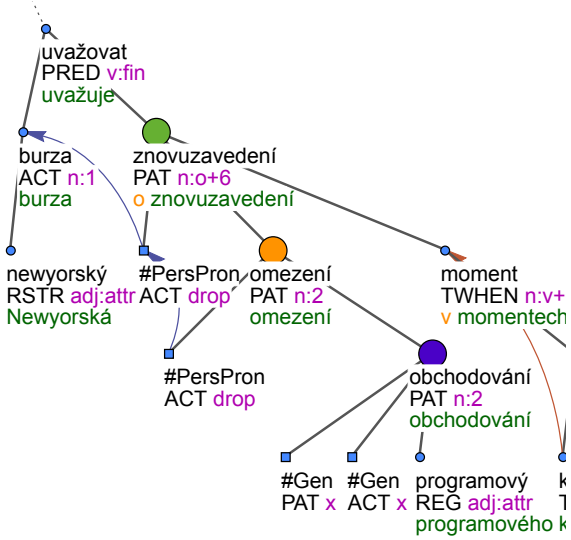
phrases are used in the written communication. The written corpus slightly prefers one expressed participant to no expressed participant, while in the spoken communication, actions are described mostly without specifying participants that take part in the situation. Combinations of two participants are rare even in the written corpus (relative frequency 3.6 % in the PCEDT and just 0.7 % in the PDTSC). Combinations of three participants appear only exceptionally.

Table 6 focuses on the case when a verbal noun is modified by another verbal noun (being a part of a prepositional phrase or in the form of prepositionless genitive). The data show that such a cumulation of verbal nouns is more frequent in the written corpus which corresponds again to the complexity of written communication. The written corpus contains 348 chains containing two verbal nouns (N_1 modified by N_2), which represents more than 2 % of all occurrences of verbal nouns denoting an action in the PCEDT. In the spoken corpus, only two such chains occur (representing just 0.15 % of all occurrences, cf. (6)). A chain containing three verbal nouns (N_1 modified by N_2 with N_2 modified by N_3) occurs only in the written corpus (8 occurrences, cf. (7) and Figure 1).

(6) *po dokončení sváření* 'after finishing welding' (PDTSC)
(7) *Nyní se obhájci UNESCO přimlouvají u prezidenta Bushe za zrušení rozhodnutí prezidenta Regana o odstoupení.* 'Now UNESCO apologists are lobbying President Bush to renege on President Reagan's decision to depart.' (PCEDT)

6.3. Degrees of explicitness

Gernsbacher (1990, p. 133—136) specifies the following scale of explicitness of anaphors (coreferring nodes): The most explicit anaphors are proper names, followed by common nouns. Pronouns are less explicit than common nouns and finally the least explicit of all referential forms are zero anaphors.



Newyorská	*burza*	*uvažuje*	*o*	*znovuzavedení*	*omezení*
New_York-NOM	*bourse-NOM*	*considers-PRS*	*about*	*reviving-LOC*	*curb-GEN*

programového	*obchodování*	*v*	*momentech,*	*kdy*	*je*	*trh*	*nestabilní.*
program-GEN	*trading-GEN*	*in*	*moments-LOC*	*when*	*is-PRS*	*market-NOM*	*volatile-NOM*

Figure 1. 'The Big Board is considering reviving a curb on program trading when the market is volatile.' (PCEDT)

We searched for participants of verbal nouns that are expressed on surface (i.e., the categories (a) and (b) in Section 4.2) and observed that among all the participants, pronouns (category (b)) are more frequent in the spoken corpus than in the written one (the relative frequency 17.0 % in the PDTSC vs. just 6.3 % in the PCEDT, see Table 7). We also checked their coreference types, that is whether their coreference is textual or grammatical. A pronoun with textual coreference is exemplified in (8) by the personal pronoun *jejich* 'their' referring to the noun *zprávy* 'messages'. A pronoun with grammatical coreference is illustrated in Figure 2 by the reflexive pronoun *svůj* 'its/their'. We suppose the Gernsbacher's scale of explicitness of anaphors can be supplemented by the opposition between pronouns with textual coreference and pronouns with grammatical coreference, the latter of which are believed to be more explicit anaphors. Our data show that pronouns with grammatical coreference (i.e., category (ba) in Section 4.2) are considerably more frequent in the written corpus than in the spoken one (cf. the relative frequency 44 % in the PCEDT and just 9.4 % in the PDTSC, see Table 7). While pronouns with textual coreference (category (bb) in Sec-

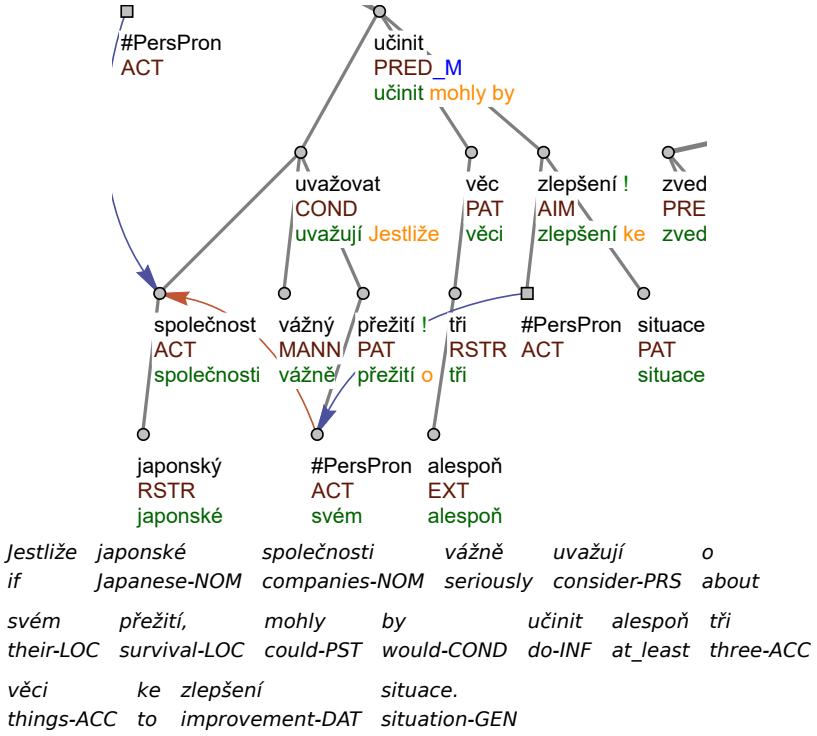


Figure 2. '*If the Japanese companies are seriously considering their survival, they could do at least three things to improve the situation.*' *(PCEDT)*

tion 4.2) account for 47.6 % in the PCEDT, they represent 81.25 % in the PDTSC. We interpret these results as a manifestation of a considerably lesser degree of explicitness in the spoken communication than in the written data.

(8) *Práce zahrnovala vybavování valutami, přijímání zpráv z těch cest a jejich předávání na jednotlivé odbory.* 'The work included providing by foreign currencies, receiving messages from the journeys and their passing on to the particular departments.' (PDTSC)

6.4. Deletions and exophoric references

In line with our expectations, there are more deletions of participants of verbal nouns in the spoken corpus than in the written one (i.e., 83.15 % in the PDTSC vs. 68 % in the PCEDT, see Table 7). Looking at the case of verbal nouns modified by no

Phenomenon	PCEDT (written corpus)	PDTSC (spoken corpus)
Participants expressed on the surface	32 %	16.85 %
Deletions of participants	68 %	83.15 %
Verbal nouns not modified by any participant from their valency frame	42 %	73 %
Participants expressed by pronouns (% of participants expressed on the surface)	6.3 %	17.0 %
– Pronouns with grammatical coreference	44 %	9.4 %
– Pronouns with textual coreference	47.6 %	81.25 %
Exophoric references (% of all participants)	0.34 %	13.37 %

Table 7. Phenomena related to degrees of explicitness

participant from their valency frame, we can see even bigger difference between the spoken and written type of communication (i.e., 73 % in the PDTSC vs. 42 % in the PCEDT, see Table 7).

Our data also reflect the difference between the context-dependent character of written language and the context-free nature of spoken communication. Adnominal participants with a context-dependent coreference refer to an antecedent in the previous context, being marked by either a textual or a grammatical coreference arrow. For example, in Figure 2, the noun *společnosti* 'companies' is referred to by a grammatical coreference arrow coming out from the node for the reflexive pronoun *svůj* 'its/their', and then this node is referred to by a textual coreference arrow coming out from the added Actor of the noun *zlepšení* 'improvement'. In contrast, adnominal participants with an extra-textual (exophoric) reference refer to something which is not mentioned in the text or speech, being indicated by a short upward arrow. For example, in the sentence illustrated by Figure 3, it is clear that children mentioned in the previous context guarded those who were at a summer camp and/or their equipment though it was not mentioned in the previous context at all; we can understand it just because we know that children usually do it when being at a camp. Adnominal participants with an exophoric reference (categories (bc) and (cc) in Section 4.2) are considerably more frequent in the spoken corpus than in the written data (13.37 % in the PDTSC vs. 0.34 % in the PCEDT).

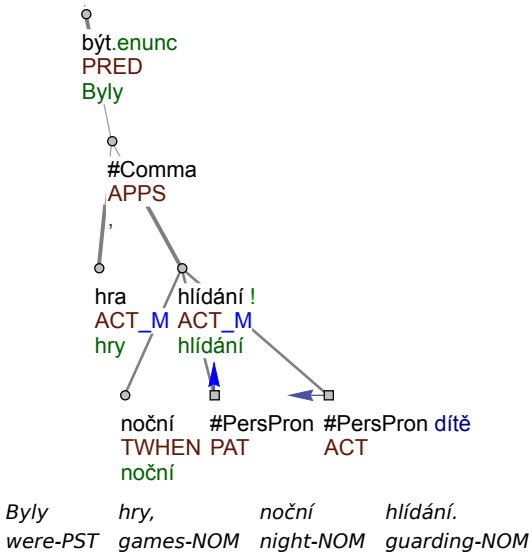


Figure 3. (To jste museli pro děti vymýšlet nějaký program? Pochopitelně. '*Did you have to think up any program for the kids? Of course.*') '*There were games, night guarding.*' *(PDTSC)*

6.5. Discussion of the results

We described differences in expressing an action by verbal nouns in the Czech part of the PCEDT (representing the written mode of communication) and in the PDTSC (representing the spoken mode of communication). Most of the results described in Sections 6.1 to Section 6.4 could not be observed without corpora with deep syntactic annotation, including annotation of deletions and coreference relations. The results confirm our hypothesis formulated in Section 2. In line with our expectations, written communication is more condensed and more complex even in the case of verbal nouns denoting an action. Although the written corpus shows higher relative frequencies, the order of the relative frequencies of particular combinations is the same in both types of communication. In the coreference domain, the complex annotation scheme enabled us to exploit the opposition between pronouns with textual and grammatical coreference (Section 6.3), demonstrating a significant difference in the degree of explicitness between written and spoken expression. The considerably higher number of deletions of participants and more frequent exophoric references give evidence about the incompleteness and context-free nature of spoken communication.

Written communication is characterized by the following features:
- Expression is more condensed, more verbal nouns per a content verb are used.
- Noun phrases are more complex which is reflected in the higher relative frequencies of combinations of participants expressed on the surface and in more frequent cumulation of verbal nouns. One expressed participant is preferred to no expressed participant.
- More explicit expressions are used (less pronouns in total, pronouns with grammatical coreference have almost the same relative frequency as pronouns with textual coreference).

In contrast, spoken communication has the following characteristics:
- Participants of verbal nouns are more often omitted on the surface and they have more often no antecedent in the context (more deletions, more extra-textual references).
- Typically no participant of a verbal noun is expressed on the surface.
- Less explicit expressions are used (approximately three times more pronouns than in written communication, considerably more pronouns with textual coreference than pronouns with grammatical coreference).

7. Conclusion

Our research into the differences between written and spoken Czech is focused on the case of verbal nouns denoting an action in the PCEDT and the PDTSC. Exploiting the annotation of valency and (co)reference relations, the results confirm our hypothesis predicting more complex noun phrases with more explicit expression in a written text and more deletions accompanied by more exophoric references in spoken communication. We support our results by numbers of occurrences of the studied phenomena in both corpora and we specify the differences between the two types of communication, providing valuable information which is hard to detect in corpora without deep syntactic annotation.

Acknowledgements

The research reported in the paper was supported by the Czech Science Foundation under the projects GA16-02196S and GA17-12624S. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071). The second author was supported by grant no. RVO 67985840 of Czech Academy of Sciences.

Bibliography

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. *Longman grammar of spoken and written English*. Longman, London, 1999.

Brazil, David. *A grammar of speech*. Oxford University Press, Oxford, 1995.

Chafe, Wallace and Jane Danielewicz. Properties of spoken and written language. In Horowitz, R.; Samuels, S. J., editor, *Comprehending Oral and Written Language*. Academic Press, New York, 1987.

Čmejrková, Světla, Lucie Jílková, and Petr Kaderka. Mluvená čeština v televizních debatách: korpus DIALOG. *Slovo a slovesnost*, 65(4):243–269, 2004.

Cruse, D Alan. *Lexical semantics*. Cambridge University Press, Cambridge, UK, 1986. ISBN 0-521-27643-8.

Cvrček, Václav et al. *Mluvnice soucasné cestiny*. Karolinum, Praha, 2010. ISBN 978-80-246-1743-5.

Gernsbacher, Morton Ann. *Language comprehension as structure building*. Erlbaum, Hillsdale, 1990.

Hajič, Jan, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In Nivre, J.; Hinrichs, E., editor, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68. Vaxjo University Press, Vaxjo, Sweden, 2003. ISBN 91-7636-394-5.

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160. European Language Resources Association, Istanbul, Turkey, 2012. ISBN 978-2-9517408-7-7.

Hajič, Jan, Eva Hajičová, Marie Mikulová, and Jiří Mírovský. Prague Dependency Treebank. In Ide, N.; Pustejovsky, J., editor, *Handbook on Linguistic Annotation*. Springer Verlag, Berlin, Germany, in press.

Halliday, Michael Alexander Kirkwood. *Intonation and grammar in British English*. Mouton, Hague, 1967.

Hausenblas, Karel. O studiu syntaxe běžně mluvených projevů. In *Otázky slovanské syntaxe: sborník brněnské syntaktické konference, 17.-21. IV. 1961*, 1962.

Heyvaert, Liesbet. *A cognitive-functional approach to nominalization in English*. Walter de Gruyter, Berlin, 2003. ISBN 3-11-017809-5.

Hoffmannová, Jana. Syntaktická stylistika mluvených projevů. In Hoffmannová, J.; Klímová, J., editor, *Čeština v pohledu synchronním a diachronním*. Karolinum, Praha, 2012.

Hoffmannová, Jana and Olga Müllerová. *Čeština v dialogu generací*. Academia, Praha, 2007.

Hoffmannová, Jana, Olga Müllerová, and Jiří Zeman. *Konverzace v češtině při rodinných a přátelských návštěvách*. Trizonia, Praha, 1999.

Hunyadi, Laszlo. Incompleteness and fragmentation in spoken language syntax and its relation to prosody and gesturing: Cognitive processes vs. Possible formal cues. In *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on*, pages 211–218, Budapest, Hungary, 2013. IEEE.

Kolářová, Veronika. Special valency behavior of Czech deverbal nouns. In Spevak, O., editor, *Noun Valency*, pages 19–60. John Benjamins Publishing Company, Amsterdam, The Netherlands, 2014. ISBN 9789027259233.

Kolářová, Veronika. Valence českých deverbativních substantiv reprezentujících vybrané sémantické třídy. *Prace Filologiczne*, in press. ISSN 0138-0567.

Leech, Geoffrey. Grammars of Spoken English: New Outcomes of Corpus-Oriented Research. *Language learning*, 50(4):675–724, 2000.

MacWhinney, Brian. The CHILDES project: Tools for analyzing talk. *Computational Linguistics*, 26(4):657–657, 2000.

Marcus, Mitchell, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. *Penn Treebank-3*. Linguistic Data Consortium, LDC99T42, University of Pennsylvania, 1999.

Mikulová, Marie. Annotation on the tectogrammatical level. Additions to annotation manual (with respect to PDTSC and PCEDT). Technical Report ÚFAL TR-2013-52, Prague, Czech Republic, ÚFAL MFF UK, 2014.

Mikulová, Marie and Jana Hoffmannová. *Korpusy mluvené češtiny a možnosti jejich využití pro poznání rozdílných "světů" mluvenosti a psanosti*, pages 78–92. Studie z korpusové lingvistiky. Lidové noviny, Praha, 2011. ISBN 978-80-7422-115-6.

Mikulová, Marie, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report 30, ÚFAL MFF UK, ÚFAL, Prague, Czech Republic, 2006.

Mikulová, Marie, Jan Štěpánek, and Zdeňka Urešová. Liší se mluvené a psané texty ve valenci? *Korpus – gramatika – axiologie*, 8:36–46, 2013. ISSN 1804-137X.

Mikulová, Marie, Anja Nedoluzhko, Jiří Mírovský, Jan Štěpánek, Petr Pajas, and Jan Hajič. *Prague Dependency Treebank of Spoken Czech 2.0*. Charles University, Prague Czech Republic, in press.

Müllerová, Olga. *Mluvený text a jeho syntaktická výstavba*. Academia, Praha, 1994.

Panevová, Jarmila. On verbal frames in functional generative description. Part I. *Prague Bulletin of Mathematical Linguistics*, 22:3–40, 1974.

Panevová, Jarmila. On verbal frames in functional generative description. Part II. *Prague Bulletin of Mathematical Linguistics*, 23:17–52, 1975.

Panevová, Jarmila. Ještě k teorii valence. *Slovo a slovesnost 59, č.1*, 1998.

Panevová, Jarmila. Valence a její univerzální a specifické projevy. In Hladká, Z.; Karlík, P., editor, *Čeština - univerzália a specifika. Sborník konference ve Šlapanicích u Brna 17.-18. 11. 1998*. Masarykova univerzita v Brně, Brno, 1999.

Panevová, Jarmila. K valenci substantiv (s ohledem na jejich derivaci). *Zbornik matice srpske za slavistiku*, 61:29–36, 2002.

Panevová, Jarmila. Contribution of Valency to the Analysis of Language. In Spevak, O., editor, *Noun Valency*, pages 1–18. John Benjamins Publishing Company, Amsterdam, The Netherlands, 2014. ISBN 9789027259233.

Piťha, Petr. Case frames of nouns. In Sgall, P., editor, *Contributions to functional syntax, semantics, and language comprehension,*, pages 91–99. John Benjamins, Amsterdam, Philadelphia, 1980.

Roberts, Celia and Brian Street. Spoken and written language. *The handbook of sociolinguistics*, pages 168–186, 1997.

Sagae, Kenji, Brian MacWhinney, and Alon Lavie. Adding Syntactic Annotations to Transcripts of Parent-Child Dialogs. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*. European Language Resources Association, Lisbon, Portugal, 2004.

Schuurman, Ineke, Machteld Schouppe, Heleen Hoekstra, and Ton Van der Wouden. CGN, an annotated corpus of spoken Dutch. In Abeillé, A., S. Hansen-Schirra, and H. Uszkoreit, editors, *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, pages 101–108. Budapest, Hungary, 2003.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. *The meaning of the sentence in its semantic and pragmatic aspects*. Reidel, Dordrecht, 1986. ISBN 90-277-1838-5.

Šonková, Jitka. *Morfologie mluvené češtiny: frekvenční analýza*. Lidové noviny, Praha, 2008.

Štěpánek, Jan and Petr Pajas. Querying Diverse Treebanks in a Uniform Way. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1828–1835. European Language Resources Association, Valletta, Malta, 2010. ISBN 2-9517408-6-7.

Těšitelová, Marie. *Psaná a mluvená odborná čeština z kvantitativního hlediska:(v rámci věcného stylu)*. Československá akademie věd, Ústav pro jazyk český, 1983.

Urešová, Zdeňka. Building the PDT-VALLEX valency lexicon. In *Proceedings of the fifth Corpus Linguistics Conference*, pages 1–18. University of Liverpool, Liverpool, UK, 2012.

Zikánová, Šárka, Eva Hajičová, Barbora Hladká, Pavlína Jínová, Jiří Mírovský, Anna Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, and Jan Václ. *Discourse and Coherence. From the Sentence Structure to Relations in Text*. Studies in Computational and Theoretical Linguistics. ÚFAL, Prague, Czech Republic, 2015. ISBN 978-80-904571-8-8.

Address for correspondence:
Veronika Kolářová
kolarova@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic