CAPTURING NUMERALS AND PRONOUNS AT THE MORPHOLOGICAL LAYER IN THE PRAGUE DEPENDENCY TREEBANKS OF CZECH

BARBORA ŠTĚPÁNKOVÁ – MARIE MIKULOVÁ

Charles University, Prague, Czech Republic

Abstract: The paper presents a novel and unified morphological description of numerals and pronouns, as compiled for the newest edition of the Prague Dependency Treebank (Prague Dependency Treebank – Consolidated 1.0) and its integral part the morphological dictionary MorfFlex. On the basis of considerable experience with real data annotation and the use of the morphological dictionary, particular changes were proposed. For both of the parts of speech a new set of subtypes was proposed, based mainly on the morphological criterion and its combination with semantic properties and other relevant features, such as definiteness in numerals and possessivity, reflexivity, and clitichood in pronouns. Each subtype has a specific value at the 2nd position of the morphological tag, which serves also as an indicator of the applicability of other tag categories.

Keywords: numerals, pronouns, morphology, treebank, annotation, Czech

1   INTRODUCTION

The paper is focused on numerals and pronouns and their morphological description in the Prague Dependency Treebank. Although these word categories are considered to be a traditional part of the part of speech (POS) set in Czech linguistics, their morphological (as well as some other) properties are quite specific and different from the rest of the inflected words. For the newest edition of the Prague Dependency Treebank (Prague Dependency Treebank – Consolidated 1.0 [1], PDT-C in the sequel) we compiled a novel and unified description of these two POS and proposed its realization in the morphological tag. The PDT-C 1.0 release is enhanced with a manual linguistic annotation at the morphological layer: all tokens of the sentence are tagged and lemmatized. A key element to ensuring annotation consistency is the morphological dictionary MorfFlex [2]. Based on the long-time experience with the use of the dictionary and manual annotation of real data, some phenomena were proposed to be captured differently in the dictionary in order to achieve better consistency within the dictionary as well as between the dictionary and the annotated data. The changes concern several complicated morphological features of Czech, including some relating to numerals and pronouns.

2   NUMERALS AND PRONOUNS IN CZECH GRAMMARS

In grammars, a POS is usually defined on the basis of a combination of semantic, morphological, and syntactic criteria. Numerals and pronouns are mainly delineated by appeal to the semantic criterion, while the other two criteria are rather problematic with these two classes.

Numerals semantically indicate number or quantity (e.g., *pět* 'five', *třetí* 'third', *několik* 'several'). However, in various approaches the set of words included as numerals can differ, e.g., *pětina* 'a fifth' is often considered a noun, despite its numeric meaning and digit representation.[1] The overarching semantic feature of pronouns lies in their ability to substitute for nouns or adjectives (e.g., *on* 'he', *tento* 'this', *nějaký* 'some').

From the morphological point of view, and in many respects also from the syntactic point of view, both numerals and pronouns can also be considered as different POS (e.g., *tisíc* 'thousand' – numeral/noun; *který* 'which' – pronoun/adjective; *druhý* 'second' – numeral/adjective; *mnohokrát* 'many times' – numeral/adverbial). On the other hand, words like *někde* 'somewhere' and *kdykoliv* 'anytime' are usually regarded as pronominal adverbs.[2] For these reasons, the most recent Czech grammar [5] does not distinguish pronouns and numerals as individual POS and assigns them among other POS.

3   NUMERALS AND PRONOUNS IN PDT-C

Within the morphological annotation in PDT-C, forms are organized into entries (paradigms) according to their formal morphological behavior. The paradigm (set of forms) is identified by a unique lemma. For each form, full inflectional information is coded in 15 tag positions. The first two tag positions encode the part of speech. The traditional POS is captured in the 1st position. The 2nd position of the tag specifies the detailed subtype of the POS and serves as an indicator of the (non-)applicability of the other categories encoded in the tag [7, p. 17]. The other positions of the tag capture morphological properties, of which gender (3rd position), number (4th), case (5th), possgender (6th), possnumber (7th) and variant (15th) (explained in 5.3 and 5.4) are relevant to the description of pronouns and numerals (cf. [6] and [7]).

The PDT approach to numerals and pronouns follows the traditional classification, i.e., numerals and pronouns are considered separate POS (the 1st tag position has the value P in the case of pronouns and C in the case of numerals). In the next part of the article, we concentrate on the description of the 2nd tag position. The main criteria for a more detailed classification of numerals and pronouns are semantics and morphological behavior. We include the semantics as a criterion for traditional reasons. In linguistics, numerals and pronouns are traditionally classified according to their

---

[1] In the NovaMorph morphological description [3] words of this type are included in a group of fractional numerals.

[2] Komárek united these semantically similar words into one group of "deictic words" [4].

purpose (as e.g., cardinal, ordinal, multiplicative, in the case of the former, and as e.g., personal, demonstrative, relative, in the case of latter).

However, the crucial principle of the classification at the 2nd tag position is the morphological one. As already mentioned, the value of the 2nd position serves as an indicator of the applicability of the other tagged categories (such as gender, number, and case) and this principle is decisive for the system of pronouns and numerals in PDT-C.[3] This means that, for example, a numeral expressing agreement in grammatical gender (e.g., *jeden, jedna, jedno* 'one') cannot be in the same group as a numeral that behaves as a noun in that it has a single, fixed gender (e.g., *sto* 'hundred') or does not express any gender (e.g., *tři* 'three') although according to the semantic criterion, such numerals may belong to the same category (the numerals *jeden*, *sto* and *tři* are all cardinal numerals; cf. Tab. 1).

In each POS group, the two main criteria (semantic and morphological) are supplemented by other criteria, following from the properties of the group. Within the class of numerals, there is the criterion of definiteness, and within the group of pronouns, possessivity, reflexivity, and clitichood are considered.[4]

4   SUBTYPES OF NUMERALS

Numerals (except for the numbers written down with Arabic or Roman numerals) are classified into several subtypes according to the various combinations of the two basic features (cf. Tab. 1):
   •   morphological behavior,
   •   semantics including definiteness.

Semantics & Morphological behavior		Adjectival		Nominal		Non-gendered	
Type	Subtype	Tag	Example	Tag	Example	Tag	Example
Cardinal	definite	**Cn**	*jeden*	**Cz**	*sto*	**Cl**	*tři*
	indefinite	**Cy**	*nejeden*	-	-	**Ca**	*málo*

---

[3] In the previous proposal [6], this principle was violated in several cases. The new proposal eliminates these violations.

[4] By comparison, some other West-Slavonic languages corpora approach description of numerals and pronouns differently. The description in the Slovak National Corpus [8] is based on a morphological criterion, the primary classification corresponds to the declension type (nominal, adjectival, mixed, etc.) that indicates other tag positions. The semantic criterion is present by inclusion of deictic words among pronouns (cf. [9]). The National Corpus of Polish takes into account mainly the morphological and syntactic criterion (strong vs. non-strong position, post-prepositional vs. non-post-prepositional position). The semantic criterion is noticeable only in a few tag positions (e.g., different values for main vs. collective numerals, personal vs. personal reflexive pronouns; cf. [10] and [11]).

Semantics & Morphological behavior		Adjectival		Nominal		Non-gendered	
Ordinal	definite	**Cr**	*první*	-	-	-	-
	indefinite	**Cw**	*kolikátý*	-	-	-	-
Multiplicative	definite	-	-	-	-	**Cv**	*třikrát*
	indefinite	-	-	-	-	**Co**	*tolikrát*
Collective	definite	**Cd**	*dvoje*	**Cj**	*patero*	-	-
	indefinite	**Ch**	*kolikerý*	**Ck**	*kolikero*	-	-
Arabic	-					**C=**	*1,25*
Roman	-					**C}**	*MDX*

Tab. 1. Subtypes of numerals

4.1 Morphological behavior

The morphological criterion is fundamental. We distinguish between adjectival, nominal and non-gendered numerals (cf. Tab. 1).

Form	Lemma	Tag
jeden muž	jeden`1	`CnYS1----------`
jedno dítě	jeden`1	`CnNS1----------`
několikátý problém	několikátý	`CwYS1----------`
několikáté problémy	několikátý	`CwIP1----------`

Tab. 2. Examples of adjectival numerals

The **adjectival numerals** express the same gender (and also the number) as that of the governing noun (e.g., *jedna žena* 'one woman' (fem. sg.), *jedno dítě* 'one child' (neut. sg.) or *několikátý problém* 'umpteenth problem' (masc. sg.), *několikáté problémy* 'multiple problems' (masc. pl.)). All forms are represented by one lemma; similarly to adjectives. However, in contrast to adjectives there is no comparative and superlative form in the adjectival declension of numerals, and some paradigms have only singular (e.g., *jeden* 'one'), or only plural forms (e.g., *dva* 'two') depending on their meaning. In the tag, the gender, number, and case are specified (see Tab. 2).

The subtype of numerals with **nominal declension** consists of numerals, whose morphological behavior is similar to that of nouns (e.g., *sto* 'hundred', *nula* 'zero', *patero* 'five-kinds-of'). They express grammatical gender, thus their tag position for gender is filled, with the whole paradigm sharing the same value (see Tab. 3).

Form	Lemma	Tag
sto lidí	sto-1`100	`CzNS1----------`
napsal dvojku a tři *nuly*	nula	`CzFP4----------`
z *patera* přikázání	patero`5	`CjNS2----------`

Tab. 3. Examples of nominal numerals

Other numerals express **no gender** and they are quite specific and diverse from the morphological point of view; they can be inflected (e.g., *tři domy* (Nom.), *tří domů* (Gen.) 'three houses'; *mnoho domů* (Nom.), *mnoha domů* (Gen.) 'many houses'), or they are uninflected, sometimes with a variant form (e.g., *několikrát, několikráte* 'several-times').[5] The tag position for gender is not filled. Other tag positions (case, number) are filled if the subtype expresses them (cf. Tab. 4).

Only the group of definite cardinal numerals covers all morphological subtypes, i.e., it expresses agreement in grammatical gender (then the tag begins with Cn; e.g., *jeden* 'one'), only one (lexical) gender (then the tag begins with Cz; e.g., *sto* 'hundred') or it expresses no gender (then the tag is Cl; e.g., *tři* 'three'); cf. Tab. 1.

Form	Lemma	Tag
tři domy	tři`3	`Cl-P1----------`
do *mnoha* zemí	mnoho	`Ca--2----------`
několikrát zazvonil	několikrát	`Co-------------`
několikráte zazvonil	několikrát	`Co------------1`

Tab. 4. Examples of non-gendered numerals

In certain contexts, numerals do not mark grammatical relations with other words in a sentence by means of inflection, they are used in their uninflected form. In the following examples, despite the formal differences (and some stylistic nuances) the meaning of these variant forms is broadly equal, i.e., the inflected forms of *sto* 'hundred' in *ke čtyřem stům dětem* 'up to four hundred children' or *do sta lidí* 'up to a hundred people' can also be expressed with the uninflected form, as in *ke čtyři sta dětem* 'up to four hundred children' and *do sto lidí* 'up to a hundred people'. In particular combinations, uninflected forms are quite usual, e.g., *až po stovky tisíc let* 'up to hundreds of thousands of years'; *sedmdesát jedna občanů* 'seventy-one citizens'; *před tři čtvrtě rokem* 'three quarters of a year ago'; *o pár stech tisících* 'about a few hundreds of thousands'. Thus, for most cardinal numerals we introduced the subspecified value X for any gender, number, or case; cf. Tab. 5.

---

[5] Variant forms are distinguished at 15th tag position by a numerical index (cf. Tab. 4).

Form	Lemma	Tag
ke *čtyřem* stům dětem	čtyři`4	`Cl-P3----------`
ke *čtyři* sta dětem	čtyři`4	`Cl-XX----------`
ke čtyři *sta* dětem	sto-1`100	`CzNXX---------1`
do *sto* lidí	sto-1`100	`CzNXX----------`
sedmdesát *jedna* občanů	jeden`1	`CnXXX----------`
před tři *čtvrtě* rokem	čtvrt	`CzFXX---------1`

Tab. 5. Examples of uninflected forms of numerals

4.2 Semantics including definiteness

In accordance with semantic features, we classify numerals into the following subtypes. Within the each semantic subtype, the numerals are further classified according to their definiteness; cf. Tab. 1.

Cardinal – express quantity (e.g., *oba* 'both', *kolik* 'how much', *málo* 'a little', *milion* 'million').

Ordinal – express position in a sequential order (e.g., *třetí* 'third', *několikátý* 'umpteenth').

Multiplicative – express how many times something occurred (e.g., *sedmkrát* 'seven-times', *mnohokrát* 'many-times').

Collective – express the number of kinds, types (e.g., *dvojí* 'two-kinds-of', *několikerý* 'several-kinds-of', *desatero* 'ten-kinds-of').

In contrast to Czech grammars, we do not distinguish interrogative numerals as a separate type: interrogative numerals are included in the corresponding types of indefinite numerals; e.g., *kolik* 'how-many' is included in the cardinal indefinite type, *kolikátý* 'at-what-position-in-a-sequence' is included in the ordinal indefinite type, or *kolikrát* 'how-many-times' is included in the multiplicative indefinite type.

5   SUBTYPES OF PRONOUNS

Pronouns form a more complicated part of speech than numerals. More criteria need to be considered for their adequate classification. As with the category of numerals, we chose the morphological behavior and semantics as the main criteria for pronouns.

Then we identified several other features that can be used to divide pronouns into various subtypes (cf. Tab. 6):

- possession,
- reflexivity,
- clitichood.

5.1 Morphological behavior

The morphological criterion divides pronouns into two groups: gendered pronouns and non-gendered pronouns.

Semantics & morphological behavior		Gendered		Non-gendered	
Type	Subtype	Tag	Example	Tag	Example
Personal	-	**PE**	*on, něj*	**PP**	*já, ty, vy*
	Clitic	**P5**	*mu*	**PH**	*mi*
	Reflexive	-	-	**P6**	*sebe*
	Reflexive Clitic	-	-	**P7**	*se, si*
	Possessive	**PS**	*můj, náš*	-	-
	Possess. 3ʳᵈ Pers	**P9**	*jeho, jejich*	-	-
	Reflex. Possess.	**P8**	*svůj*	-	-
Relative	-	**P4**	*který, čí, jenž*	**PQ**	*kdo, copak*
	Possessive	**P1**	*jehož, jejichž*		
Indefinite	-	**PZ**	*nějaký, čísi*	**PK**	*kdosi, nevímco*
Negative	-	**PW**	*nijaký, žádný*	**PY**	*nikdo, nic*
Demonstrative	-	**PD**	*ten, takový*	-	-
Delimitative	-	**PL**	*všechen*	-	-

Tab. 6. Subtypes of pronouns

Gendered pronouns express different values of the gender (and also number):

(a) depending on the grammatical gender and number of the governing noun; cf. *žádný dům* 'no house' (masc. inanim. sg.), *žádná žena* 'no woman' (fem. sg.), *žádní muži* 'no men' (masc. anim. pl.). These pronouns behave as syntactic adjectives in sentences;

(b) according to the gender, animacy or number of the referent they substitute (e.g., *on* 'he', *ona* 'she', *ono* 'it', *oni* 'they'), they behave syntactically as nouns.

All forms of both of the types are represented by one lemma (Nom. sg. masc. anim.), similarly to adjectives: e.g., *žádný* 'no' for *žádná* (fem.), *žádné* (neut.), and *on* 'he' for *ona* 'she' and *oni* 'they', etc. The tag positions for gender and number are filled; cf. Tab. 7.

Non-gendered pronouns are pronouns that express no gender and number (e.g., *ty* 'you'; *nikdo* 'nobody', *cosi* 'something'; Tab. 8). The gender and number tag positions are not filled. These pronouns behave syntactically as nouns.

Form	Lemma	Tag
on	on-1	`PEYS1--3-------`
oni	on-1	`PEMP1--3-------`
ono	on-1	`PENS1--3-------`
žádná žena	žádný	`PWFS1----------`
žádní muži	žádný	`PWMP1----------`

Tab. 7. Examples of gendered pronouns

Form	Lemma	Tag
nikdo	nikdo	`PY--1----------`
nikoho	nikdo	`PY--2----------`
cosi	cosi	`PK--1----------`
čehosi	cosi	`PK--2----------`

Tab. 8. Examples of non-gender pronouns

We are aware that the pronoun *kdo* 'who' (and other various personal pronouns) could be classified as masculine, and the pronoun *co* 'what' (and other various non-personal pronouns) could be classified as neuter. However, there are many uses of these pronouns where the gender and number category seems to be questionable (e.g., *kdo jste tam byli* (masc. anim. pl.) 'who of you were there', *každá* (fem. sg.)*, kdo jste přišla* (fem. sg.) 'each of you who came', *nikdo nejsme* (1st pers. pl. – 'we') *dokonalý* 'none of us are perfect'; cf. also examples and the discussion in [12]).

5.2 Semantics

We identified six main semantic groups of pronouns, largely following the Czech grammar tradition. In each group, we use the unique tag value to distinguish between gendered and non-gendered pronouns:

Personal: substitute a particular word referring to a person, thing, and the like, including pronouns indicating clitichood, possession and reflexivity (see below); e.g., *já* 'I', *on* 'he', *ní* 'her', *náš* 'our', *jeho* 'his', *svůj* 'self'.

Relative/Interrogative: in relative clauses, they are used as connecting words referring back to their antecedents; in questions, they serve as interrogative words; e.g., *jaký* 'what', *který* 'which', *čí* 'whose', *co* 'what', *kdož* 'who'.

Indefinite: refer to one or more unspecified persons or things; e.g., *nějaký* 'some', *čísi* 'somebody's', *sotvakterý* 'hardly-some', *někdo* 'somebody', *kdokoliv* 'whoever'.

Negative: refer to nonexistence of persons, things or their properties; e.g., *ničí* 'nobody's', *žádný* 'no/none', *nic* 'nothing', *nikdo* 'nobody'.

Demonstrative: point to a specific person or thing; e.g., *ten* 'this', *tamten* 'that', *onen* 'that-over-there', *tentýž* 'same', *takový* 'such'.

Delimiting: (sometimes included in the indefinite group [13, p. 224]) indicate the universality or totality; e.g., *všechnen* 'all', *sám* 'alone', *veškerý* 'whole'.

In contrast to Czech grammars (e.g., [13, pp. 221–222]), we do not distinguish interrogative pronouns as a separate subtype because of their unclear distinction from relative pronouns.

5.3 Possessivity and reflexivity

Several subtypes of the pronouns are introduced based on the feature of reflexivity and possession. These features are characteristic of particular personal and relative pronouns.

Besides gender and number of an object, **possessive pronouns** (for the 3rd person) also express the number and gender of the possessor; e.g., *jeho chalupy* 'his cottages' (fem. pl., possessor: masc. sg.), *z jejíhož domu* 'from whose house' (masc. sg., possessor: fem. sg.). This is why the 6th (possgender) and the 7th (possnumber) tag positions are also filled; see the comparison with the other possessive pronouns which express only the number of a possessor (e.g. *můj dům* 'my house' (possesor: sg.), *náš dům* 'our house' (possesor: pl.)) or nothing (e.g, *hájí svoji pravdu* 'defends his/their truth') in Tab. 9.

Reflexive pronouns express only a limited number of morphological categories. Possessive reflexive pronouns only express agreement in gender, number, and case. Gender and number of the possessor are not distinguished. Personal reflexive pronouns express only case (e.g., *mluví o sobě* (Loc.) 'he talks about himself'; *zatleskejte si* (Dat.) 'give yourselves a clap'), cf. Tab. 9.

Form	Lemma	Tag
můj dům	můj	PSYS1-S1-------
náš dům	náš	PSYS1-P1-------
jeho chalupy	jeho	P9XXXZS3-------
z *jejíhož* domu	jehož	P1ZS2FS3-------

Form	Lemma	Tag
hájí *svoji* pravdu	svůj	`P8FS4----------`
mluví o *sobě*	se	`P6--3----------`
zatleskejte *si*	se	`P7--3----------`

Tab. 9. Examples of possessive and reflexive pronouns

5.4 Clitichood

Some of the personal pronouns have a very wide set of forms. In addition to the basic variants, they often have clitic (weak) variants used in specific syntactic contexts (e.g., *zná ho dobře* 'he knows him well'; *dej mi to* 'give me that', and *zatleskejte si* 'give yourselves a clap') which have a special value of the 2nd tag position; cf. Tab. 9 and 10. Distinguishing clitic forms at the 2nd tag position violates the principle that the 2nd position is the same for the whole paradigm [7, p. 7]. This exception follows from the historical development of the MorfFlex dictionary.

Form	Lemma	Tag
zná jen *jeho*	on	`PEYS4--3-------`
zná *ho* dobře	on	`P5ZS4--3-------`
pro *něho*	on	`PEZS4--3------1`
proň	on	`PEZS4--3-----p-`
dej *mi* to	já	`PH-S3--1-------`
dej *mně* to	já	`PP-S3--1-------`

Tab. 10. Examples of personal pronouns

Furthermore, several pronouns (e.g., *on* 'he', *jenž* 'who/what/which') have a special form when following a preposition (cf. *zná jen jeho* (Accus. sg.) 'he knows only him' vs. *pro něho* (Accus. sg.) 'for him', or its rarely used form *proň*[6]). These forms are marked at the 15th position of the tag (cf. Tab. 10).

6 CONCLUSION

Even in Czech linguistics, there are many approaches to these POS types, each with its pros and cons. We have proposed one that primarily takes into account the morphological aspect. In the multi-layer concept of language [15] applied to PDT-C, syntactic and semantic properties are captured in higher layers. In the proposal under

---

[6] The form *proň* belongs to the so called "aggregates", forms created by joining two or more forms. Note the letter *p* (for the preposition *pro*) at the 14th tag position (in Tab. 10). For more details see [14].

discussion, numerals and pronouns are sorted at the 2ⁿᵈ tag position according to their morphological behavior combined with the traditional semantic classification. The value of the 2ⁿᵈ tag position determines the type of declension (adjectival, nominal, etc.) or the uninflected character of word, as well as what other morphological categories are expressed in the given subtype.

ACKNOWLEDGEMENTS

References

[1] Hajič, J. et al. (2020). Prague Dependency Treebank – Consolidated 1.0 (PDT-C 1.0). LINDAT/CLARIAH-CZ, Prague. Accessible at: http://hdl.handle.net/11234/1-3185.
[2] Hajič, J., Hlaváčová, J., Mikulová, M., Straka, M., and Štěpánková, B. (2020). MorfFlex CZ 2.0. LINDAT/CLARIAH-CZ, Prague. Accessible at: http: //hdl.handle.net/11234/1-3186.
[3] Petkevič, V., Hlaváčová, J., Osolsobě, K., Svášek, M., and Šimandl, J. (2019). Parts of Speech in NovaMorf, a New Morphological Annotation of Czech. Jazykovedný časopis 70(2), pages 358–369.
[4] Komárek, M. et al. (1986). Mluvnice češtiny 2. Academia, Prague.
[5] Štícha, F. et al. (2018). Velká akademická gramatika spisovné češtiny. Academia, Prague.
[6] Hajič. J. (2004). Disambiguation of Rich Inflection (Computational Morphology of Czech). Karolinum, Prague.
[7] Mikulová, M. et al. (2020). Manual for Morphological Annotation, Revision for the Prague Dependency Treebank – Consolidated 2020 release. Technical report, 2020/TR-2020–64, Charles University, Prague.
[8] Slovenský národný korpus – prim-6.1-public-sane. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2013. Accessible at: https://korpus.juls.savba.sk/.
[9] Morphological annotation of texts in the Slovak National Corpus – Numerals and Pronouns. Accesible at: https://korpus.sk/num.html; https://korpus.sk/pronom.html.
[10] https://www.sketchengine.eu/polish-nkjp-part-of-speech-tagset/.
[11] Prepiórkovski, A. (2009). A comparison of two morphosyntactic tagsets of Polish. In Representing Semantics in Digital Lexicography: Proceedings of MONDILEX Fourth Open Workshop, Warsaw, pages 138–144.
[12] Petkevič, V. (2010). L'accord en tchèque: le centre et la périphérie. Écho des études romanes, 6(1–2), pages 143–160.
[13] Cvrček, V. et al. (2015). Mluvnice současné češtiny. Karolinum, Prague.
[14] Hlaváčová, J., Mikulová, M., Štěpánková, B., and Hajič, J. (2019). Modifications of the Czech morphological dictionary for consistent corpus annotation. Jazykovedný časopis 70(2), pages 380–389.
[15] Sgall, P., Hajičová, E., and Panevová, J. (1986). The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Academia, Prague.