# **Derivation in the Czech National Corpus**

## Jana Klímová\*, Jan Kocek†

\*Institute of the Czech Language
Academy of Sciences of the Czech republic
Letenská 4, 118 51 Praha 1, Czech republic
jana.klimova@ff.cuni.cz

†Institute of the Czech National Corpus
Charles University, Faculty of Philosophy
nám. J. Palacha 2, 116 38 Praha 1, Czech republic
jan.kocek@ff.cuni.cz

#### Abstract

The aim of this paper is to describe one the main means of Czech word formation - derivation. New Czech words are created by composition or by derivation (by using prefixes or suffixes).

The suffixes which are added to the stem are used much more frequently than prefixes standing before the stem.

The most frequent suffixes will be classified according to the paradigmatic and semantic properties and according to the changes they cause in the stem.

The research is done on the Czech national corpus (CNC), the frequencies of the investigated suffixes illustrate their productivity in present day Czech language.

This research is of a particular value for a highly inflected language such as Czech. Possible applications of this system are various NLP systems, e.g. spelling checkers and machine translation systems. The results of this work serve for the computational processing of Czech word formation and in future for the creation of the Czech derivational dictionary.

## 1. Classification of suffixes

Derivation by suffixes is the most frequent way of formation new Czech words. The suffixes can cause the change of POS or gender of a noun. During the process of word formation different changes in the stem (alternations of characters) are caused by the suffixes.

The most productive suffixes are classified according to:

### 1.1 Paradigmatic properties

The different classes of suffixes represent the parts of speech<sup>1</sup> of the basic and derived words. The suffixes belonging to the NN class are further classified into different classes according to the gendres (e.g. the diminutive suffixes do not change the gender of the basic word but on the other hand there are suffixes which change the masculine gender of basic words into feminine gender (e.g. učitel (male teacher) - učitelka (female

teacher)).

The list of VN suffixes (serving for the creation of nouns from verbs):

-tel (names of agents of intellectual activities), e.g. ředitel (director), cestovatel (traveller), buditel (awakener), školitel (tutor), pokračovatel (continuator), sběratel(collector)

frequency in CNC: 366933

-č (names of agents of physical activities), e.g. řidič (driver), trubač (bugleman), lamač (breaker), sazeč (printer), hlídač (watcher)

frequency in CNC: 149761

**-dlo** (names of means), e.g. mýdlo (soap), žrádlo (food), chodidlo(sole), sedadlo (seat), letadlo (plane), prostěradlo (sheeting), kloktadlo (gargle)

frequency in CNC: 126009

-tko (names of means), e.g. sluchátko (receiver), kukátko (spy hole), plivátko (spittoon), trsátko (plectron), párátko (toothpick), šoupátko (slider)

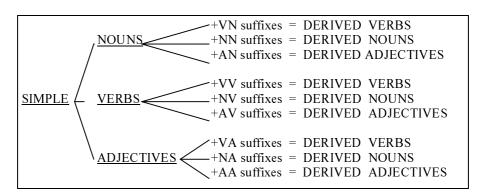


Figure 1: Relations between basic and derived words and different types of suffixes

frequency in CNC: 15807

-ák (names of means), e.g. šroubovák (screwer), piják (blotch), věšák (stand)

frequency in CNC: 124079

<sup>&</sup>lt;sup>1</sup> A - adjective, N - noun, V - verb, e.g. VN suffixes serve for the creation of nouns from verbs (e.g. -tel, -č), NN suffixes serve for the creation of nouns from nouns (e.g. diminutive suffixes), etc.

**-ivo**, e.g. topivo (fuel), stelivo (roughage), hnojivo (fertilizer), pojivo (binder)

frequency in CNC: 24713

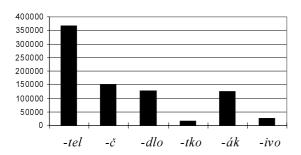


Figure 2: Frequencies of VN suffixes in CNC

#### The list of NN suffixes

(serving for the creation of nouns from nouns):

-ář (names of agents), e.g. kovář (smith), mlynář (miller), nástrojář (toolmaker), rybář (fisherman)

frequency in CNC: 162191

-ník (names of agents), e.g. školník (school servant), hrobník (gravedigger), lodník (sailor)

frequency in CNC: 361643

-ista (names of agents), e.g. traktorista (tractor driver), houslista (violinist), klavírista (piano player)

frequency in CNC: 115662

-rna (names of locations), e.g. cukrárna (sweetshop), vinárna (wineroom), elektrárna (power plant), cementárna (cement works), kavárna (coffee house), konírna (stable)

frequency in CNC: 54123

-ovna, e.g. (names of locations), e.g. knihovna (library), strojovna (machine room)

frequency in CNC: 63825

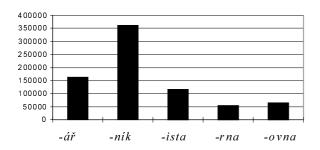


Figure 3: Frequencies of NN suffixes in CNC

#### The list of AN suffixes

(serving for the creation of nouns from adjectives):

-ost (names of properties), e.g. hloupost (stupidity), soudnost (judiciousness), přímost (straightness), velikost (greatness)

frequency in CNC: 221574

-ec, e.g. krasavec (good looker), milenec (lover)

frequency in CNC: 235688

#### The list of NA suffixes

(serving for the creation of adjectives from nouns):

-ový, e.g. jahodový (strawberry), jablkový (apple), tiskový (printing), listový (foliar)

frequency in CNC: 1067288

-natý, e.g. dřevnatý (ligneous), kolenatý (knee), ramenatý (square shouldered), dužnatý (pulpy)

frequency in CNC: 6657

-ovitý, e.g. houbovitý (spongy), šlachovitý (stringy), dřevitý (woody), zemitý (earthy)

frequency in CNC: 7540

-ský, e.g. rajský (ambrosial), horský (mountain), chlapský (manful)

frequency in CNC: 1597518

-cký, e.g. otrocký (slavish), světácký (worldly), pijácký (drinky)

frequency in CNC: 837379

-ný, e.g. dřevný (wood), senný (hay), vonný (aromatic)

frequency in CNC: 2086582

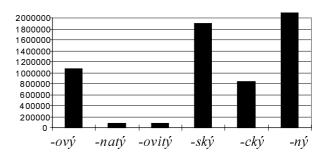


Figure 4: Frequencies of NA suffixes in CNC

#### The list of AA suffixes

(serving for the creation of adjectives from adjectives): -ejší/ější, e.g. světlejší (lighter), dolejší (lower), krásnější (more pretty)

frequency in CNC: 192158

-avý, e.g. bělavý (whitish), modravý (bluish), zelenavý (greenish)

frequency in CNC: 82929

-čký, e.g. maličký (tiny), mladičký (youthful), stařičký (aged), maloučký (tiny)

frequency in CNC: 3417

-nký, e.g. malinký (tiny), prostinký (very simple), lehoulinký (feathery), malounký (tiny), slaďounký (very small)

frequency in CNC: 3030

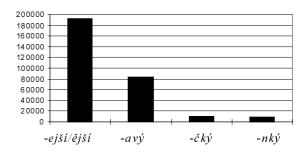


Figure 5: Frequencies of AA suffixes in CNC

### The list of VA suffixes

(serving for the creation of adjectives from verbs):

-ici, e.g. honici pes (hunting dog), vodici (leading), pěnici (foaming)

frequency in CNC: 4894

-ivý, e.g. vodivý (conducting), perlivý (sparkling), horlivý (zealous)

frequency in CNC: 119887

### The list of NV suffixes

(serving for the creation of verbs from nouns):

-ět, e.g. zkrásnět (turn beautiful), vědět (know), vidět (see), letět (fly)

frequency in CNC: 256541

-it, e.g. trůnit (throne), strašit (frighten), ztvárnit (externalize), vyrobit (produce)

frequency in CNC: 128206

-at, e.g. vnímat (perceive), drbat (scratch), chovat (raise), skákat (jump)

frequency in CNC: 371637

-ovat, e.g. startovat (start), pochodovat (march), válcovat (roll)

frequency in CNC: 142

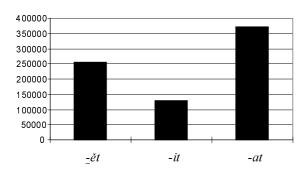


Figure 6: Frequencies of NV suffixes in CNC

#### The list of VV suffixes

(serving for the creation of verbs from verbs):

**-nout,** e.g. vyvinout (develop), kousnout (bite), rýpnout (dig)

frequency in CNC: 151907

-out, e.g. plout (sail), jmout (take upon one to), nazout

frequency in CNC: 169486

-it, e.g. vyrobit (produce), uhodit (hit), zlomit (break) frequency in CNC: 128206

-at, e.g. vstávat (get up), chodívat (walk), volávat (call)

frequency in CNC: 371779

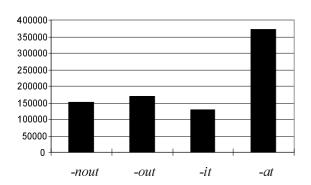


Figure 7: Frequencies of VV suffixes in CNC

## The list of AV suffixes

(serving for the creation of adjectives from verbs):

-ávat, e.g. zmodrávat (turn blue)

frequency in CNC: 113

#### 1.2 Alternations

The changes in the stem caused by the suffixes may be of the following kinds:

- (a) palatalization of final consonant e.g., hák háček, cukr cukřík
- (b) change of quality of vocal (shortening, e.g., žába žabka, prolongation e.g., hřib hříbek, or other change, e.g., kůl kolík, anděl andílek, brambor brambůrek)
- (c) and combinations thereof, e.g.,  $rak r\acute{a}\check{c}ek$ ,  $v\acute{t}tr v\check{e}t\check{r}\acute{t}k$ ,  $buk bou\check{c}ek$ ,  $o\check{r}ech o\check{r}\acute{t}\check{s}ek$ ,  $kniha kn\acute{t}\check{z}ka$ ,  $d\check{r}evo d\check{r}\acute{v}ko$

This classification serves for the definition of derivational rules which form the algorithm of the programming tool processing the Czech derivation (both analysis of derived words and generation of new words). Examples of derivation rules for creation of masculine diminutives are given in Table 1.

Rules masculine gender	Examples
aS - áSek - áSeček	vlas - vlásek
áS - áSek - áSeček	jestřáb - jestřábek
eS - eSík - eSíček	dřep - dřepík
eS - í/ýSek - í/ýSeček	jelen - jelínek
éS - e/éSík - éSíček	chléb - chlebík
ěS - íSek - íSeček	medvěd - medvídek
iS - íSek - íSeček	hřib - hříbek
íS - íSek - íSeček	díl - dílek
oS - oSík - oSíček	strom - stromek
oS - ůSek - ůSeček	hrob - hrůbek
uS - ouSek	holub - holoubek
ouS - ouSek - ouSeček	kloub - kloubek
yS - ySek - ySeček	záhyb - záhybek
ýS - ýSek - ýSeček	sýr - sýrek
rS - rSík - rSíček	prd - prdík
S1 - Slík - Slíček	štrůdl - štrůdlík
Sr - Sřík	zubr - zubřík

Table 1: Derivation relations for masculine diminutives

### 1.3 Semantic properties

The stem wears the same semantic properties in all its derivatives, the suffix gives its function (semantic properties) to the new word. It is quite usual that a suffix is being used freely by writers and speakers of the language, if one knows the meaning of the given suffix added to a known basic word, the meaning of the whole word could be understood even if this newly created word was not seen before.

The suffixes are classified into following classes regarding their semantic properties (Dokulil, 1962):

- (a) class of suffixes which do not change the lexical meaning of the basic words (syntactic derivation), e.g.,
- nouns (in English grammar gerund) created from verbs expressing an activity (spaní (sleeping), praní (washing), chození (walking), nakupování(shopping))
- nouns created from adjectives expressing a property (rychlost (speed), zvědavost (curiosity) dobrota (goodness), citlivost (sensitivity))
- (b) class of suffixes which modify the lexical meaning of basic words *e.g.*,
  - diminutives
  - nouns forming sex-opposites
- (c) class of suffixes which change basically the lexical meaning of basic words, e.g.,

- names of agents (učitel (teacher), řidič (driver), kovář (smith), dělník (worker))
- names of means (mýdlo (soap), šroubovák (screwdriver))
- names of places (divadlo (theatre), kovárna (smithy), studovna (studying room), ložnice (sleeping room))

The ambiguity of some suffixes (some suffixes have different semantic properties combination with various stems) has to be taken into account and the different cases should be processed in a different way. This phenomenon can be illustrated by

- (a) the suffix -dlo which serves for the formation of
- (a1) nouns of means (mýdlo (soap))
- (a2) nouns of locations (divadlo (theatre))
- (b) the suffix -ka which serves for the derivation of
- (b1) feminine nouns from masculine nouns (učitelka (female teacher))
  - (b2) feminine diminutives (chatka (little cottage))

#### 2. Sources of data

Corpora are excellent sources of information about the existence of derived words and about the frequency of their use. In the course of the work, the lemmatised and tagged Czech National Corpus (CNC) with 100 million current words covering about 50 different genres was used. CNC therefore serves for illustrating the productivity of suffixes in contemporary Czech language.

Backordered dictionaries (Slavíčková, 1975) provide material for the information about suffixes and about the possible alternations in stem caused by these suffixes.

## 3. Conclusion

This research is of a particular value for a highly inflected language such as Czech. Possible applications of this system are various NLP systems, e.g. spelling checkers and machine translation systems.

The results of this research serve for the definition of derivation relations forming the basis of the algorithm of the programming tool which is able to analyse derived words and generate new words. In future these findings will be used for the Czech derivational dictionary defining the most productive suffixes, their paradigmatic and semantic properties.

## 4. Acknowledgement

This research was supported by the GACR, Grant Nr. 405/96/K214.

#### 5. References

CNC: http://ucnk.ff.cuni.cz/cnc

Čermák, F., 1990. Syntagmatika a paradigmatika tvoření slov II, Morfologie a tvoření slov. Praha: Univerzita Karlova.

Dokulil, M., 1962. Tvoření slov v češtině. Praha: Academia.

Slavíčková, E., 1975. Retrográdní morfématický slovník češtiny. Praha: Academia.

Slovník spisovného jazyka českého 1-4, 1957-1971. Praha: Academia.