Revize morfologických gramatémů pro Pražský závislostní korpus 3.0

Jarmila Panevová
Magda Ševčíková

{panevova,sevcikova}@ufal.mff.cuni.cz

Ústav formální a aplikované lingvistiky
Matematicko-fyzikální fakulta
Univerzita Karlova v Praze

- Gramatémy ve FGP a PDT
- Anotace gramatémů v PDT 2.0
- Revize gramatémů pro PDT 3.0
  - Slovesné gramatémy
    - Gramatém gramatických diatezí
    - Gramatém skutečnostní modality
  - Substantivní gramatém čísla
- Závěrečné poznámky
- Reference

Gramatémy ve FGP a PDT

- atributy uzlů tektogramatického stromu
- zachycují významy morfologických kategorií, které jsou relevantní pro význam věty
  - číslo u substantiv, čas u sloves
  - nikoli pád u substantiv, číslo u adjektiv
- morfologické významy jako součást popisu větného významu
  - ve FGP od počátku této teorie (Sgall 1967)
  - teorie smysl-text („gramémy“, Meľčuk 1988)

Gramatémy v PDT 2.0

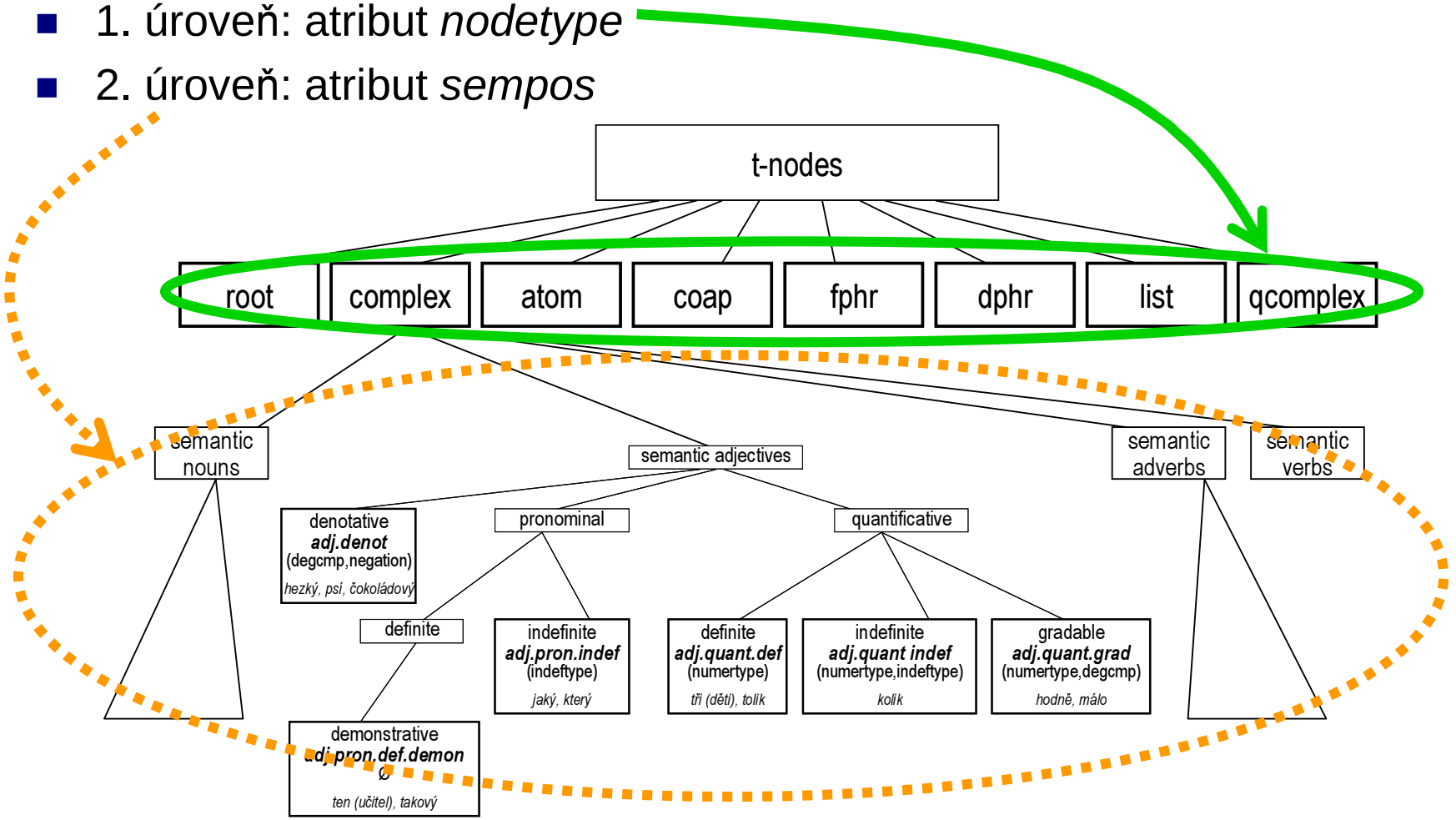
- 15 gramatémů
  - *number*
  - *gender*
  - *person*
  - *politeness*
  - *indeftype*
  - *numertype*
  - *negation*
  - *degcmp*
  - *tense*
  - *aspect*
  - *verbmod*
  - *deontmod*
  - *dispmod*
  - *resultative*
  - *iterativeness*

Přidělování gramatémů tektogramatickým uzlům v PDT 2.0

- ne všechny gramatémy jsou relevantní pro všechny uzly
  - čas u substantiv, stupeň u sloves apod.
- typování uzlů – dvouúrovňové:
  - 1. úroveň: 8 typů uzlů
    - atribut *nodetype*
    - gramatémy relevantní pouze pro jeden z nich: tzv. komplexní uzly (*nodetype=complex*)
  - 2. úroveň: komplexní uzly rozčleněny do 19 skupin (detailně členěné sémantické slovní druhy, Dokulil 1962)
    - atribut *sempos*

Atributy *nodetype* a *sempos*

- 1. úroveň: atribut *nodetype*
- 2. úroveň: atribut *sempos*



Hodnoty atributů *nodetype* a *sempos* v PDT 2.0

- hodnoty atributu *nodetype*:

complex	550947
root	49442
qcomplex	46015
coap	35747
atom	34035
fphr	4549
list	2512
dphr	1282

- hodnoty atributu *sempos*:

n.denot	236926
adj.denot	100877
v	88037
n.pron.def.pers	32903
adj.quant.def	19441
n.denot.neg	18831
n.pron.indef	11343
adv.denot.ngrad.nneg	8947
n.quant.def	7994
adj.pron.def.demon	5746
n.pron.def.demon	4759
adj.pron.indef	3383
adv.pron.indef	3107
adv.pron.def	2928
adj.quant.grad	1865
adv.denot.grad.neg	1315
adv.denot.grad.nneg	1139
adv.denot.ngrad.neg	751
adj.quant.indef	655

Hodnoty gramatémů v PDT 2.0

- hodnoty
  - vlastní (základní) hodnoty
    - př. hodnoty *sg, pl* pro gramatém *number*, hodnoty *ant*, *sim, post* pro *tense*
  - hodnota *nr*
    - možná u všech gramatémů
    - pokud nelze rozhodnout mezi vlastními hodnotami
    - př. v gramatému *number* u substantiva ve větě *Otevřel dveře*
  - hodnota *nil*
    - pouze u některých slovesných gramatémů
    - pokud není relevantní ani jedna z vlastních hodnot
    - př. v gramatému *tense* u uzlu reprezentujícího imperativ nebo infinitiv
  - hodnota *inher*
    - u koreferujících uzlů reprezentujících relativa nebo reflexiva
    - pokud uzel „dědí“ gramatémové hodnoty od svého antecedentu
    - př. v gramatému *number*, *gender* u zájmena ve větě *Muži, kteří přišli ...*

Počet gramatémových hodnot v PDT 2.0

- 1 594 333 hodnot gramatémů
  - u 550 947 komplexních uzlů
- z toho manuální anotace
  - 17 520 gramatémových hodnot
  - např. sémantické číslo u pomnožných substantiv
  - mezianotátorská shoda: 70 až 85 %

Revize gramatémů

- v rámci projektu GA ČR „*Komputační lingvistika: Explicitní popis jazyka a anotovaná data se zřetelem na češtinu*“
  - hl. řešitelka prof. J. Panevová
  - 2010–2013
  - náplň
    - revize tektogramatické anotace (**gramatémy**, koreference, ...)
    - zvětšení objemu anotovaných dat
  - revidovaná data + nová data jako **PDT 3.0**

Revize slovesných gramatémů

- v PDT 2.0: 7 slovesných gramatémů
- pro PDT 3.0
  - gramatémy *tense, iterativeness* a *deontmod* beze změny
  - pro gramatém *aspect* definována další hodnota
  - gramatémy *resultative*, *dispmod* a *verbmod* zrušeny
  - zavedeny 3 nové gramatémy
    - *diatgram, diatsynt, factmod*

Slovesné gramatémy
PDT 2.0 vs. PDT 3.0

Grammateme	Values and explanation
tense	ant, sim, post meanings of the morphological category of tense
iterativeness	it0, it1 whether an event is not/is presented as a repeated action
deontmod	deb, hrt, vol, poss, perm, fac, decl modal meanings (possibility, permision etc.) expressed by modal verbs
aspect	proc, cpl meanings of verbal aspect
resultative	res0, res1 whether an event is not/is presented as a result of the preceding action
dispmod	disp0, disp1 whether the attitude of the agent is not/is expressed by a special construction
verbmod	ind, cdn, imp direct counterpart of the morphological mood

Grammateme	Values
tense	ant, sim, post
iterativeness	it0, it1
deontmod	deb, hrt, vol, poss, perm, fac, decl
aspect	proc, cpl, perf
diatgram	act, pas, res1, res2, deagent, disp, recip
diatsynt	inferred from the tree structure
factmod	asserted, potential, irreal

Slovesné gramatémy zasažené revizí

- počet výskytů jednotlivých hodnot v datech PDT 2.0

Grammateme	Value	# of occurences
tense	ant	31217
	sim	40987
	post	8654
	nil	7166
iterativeness	it0	87919
	it1	105
deontmod	deb	1173
	hrt	3255
	vol	1016
	poss	2777
	perm	92
	fac	95
	decl	79616
aspect	proc	51900
	cpl	35839
	nr	285
resultative	res0	87669
	res1	355
dispmod	disp0	80824
	disp1	9
	nil	7191
verbmod	ind	77145
	cdn	3680
	imp	375
	nil	6824

Gramatém gramatických diatezí *diatgram*

- úzká spojitost s tradiční kategorií slovesného rodu
  - základní opozice aktivum vs. pasivum
  - do nového gramatému zahrnuty další gramatikalizované diateze
- obměna výchozí bezpříznakové diateze aktivní formálně charakterizovaná
  - změnou slovesné formy
    - pasivizací, reflexivizací, vytvořením složeného tvaru za účasti pomocných sloves *být, mít, dostat*
  - změnou v uspořádání aktantů (jiná hierarchizace)
- obměna významově charakterizovaná výběrem hodnoty gramatému *diatgram*
  - (i) aktivum (hodnota ***diat0***)
  - (ii) pasivum opisné *(**pas**)*
  - (iii) dispoziční diateze *(**disp**)*
  - (iv) rezultativní diateze se dvěma hodnotami *(**res1**, **res2**)*
  - (v) recipientní diateze (***recip***)
  - (vi) deagentní diateze *(**deagent**)*

Gramatém *diatgram* (pokr.)

- př.
  - (i) *Nakladatelství KAROLINUM **vydá** vybrané spisy Petra Sgalla.* (***diat0***)
  - (ii) *Nakladatelstvím KAROLINUM **budou vydány** vybrané spisy Petra Sgalla.* (***pas***)
  - (iii) ***Hrálo se** mi výborně, vůbec se mi nechtělo střídat.* (PDT 2.0) (***disp***)
  - (iv) *Polévka už **je uvařena**.* (***res1***)

    *Matka už **má** polévku **uvařenu**.* (***res2***)
  - (v) ***Dostal jsem** od otce **nařízeno** koupit noviny*.(***recip***)
  - (vi) *Ve Slezsku **se** často **mluví** polsky*. (***deagent***)

- Gramatická diateze je gramatikalizovaná (morfologická) kategorie, slovesa *mít* v (iv), (v), *dostat* v (v) jsou pomocná slovesa, *má uvařeno* v (iv) je složený tvar slovesa stejně jako *bude vydáno* v (ii)

Gramatém *diatgram* – omezení

- (iii) ***Hrálo se.disp*** *mi výborně.*
  - obligatorně přítomno určení způsobu (quale děje, snadnost, obtížnost provedení děje)
- (iv) *Polévka už* ***je uvařena.res1***, *Matka už* ***má*** *polévku* ***uvařenu.res2***,
  (v) ***Dostal jsem*** *od otce* ***nařízeno.recip*** *koupit noviny.*
  - netvoří se od všech sloves
  - zapsat jako slovníkový rys: „tvoří” *res, recip*
  - **(iv):** tvoří se od většiny tranzitivních, od některých intranzitivních (*má nakročeno, našlápnuto*, *namířeno, nahnáno*), zpravidla dokonavých, ale někdy i od nedokonavých
  - **(v):** sémantické skupiny sloves: *dostal zaplaceno, přiděleno; povoleno, nařízeno; nařezáno, vynadáno* – od dokonavých sloves
- (vi) *Ve Slezsku* ***se*** *často* ***mluví.deagent*** *polsky*.
  - přítomnost Gen.ACT

Gramatém *diatgram* – syntaktické důsledky

- (iii) ***disp**:* ACT (pokud není Gen) -> dativ Subjektu
- (iv) ***res2**:* ACT nebo ADDR -> Sb
  - př.

    (1) *Student.ADDR už **má** děkanátem*.ACT *nostrifikaci diplomu **uznánu**.*

    (2) *Ruce* (ACT) ***měla schovány** pod róbou*. (SYN2005)

    (3) *Také* (ACT) ***mám rozpracován** nejpozoruhodnější román, jaký jsem v životě napsal.* (SYN2005)

    (4) *Pacient*.ADDR ***měl zasaženy** vnitřní orgány*. (SYN2005)

    (5) *Ženu kriminalisté nalezli v jejím bytě,* (ADDR nebo ACT) ***měla** kolem krku **omotáno** vodítko na psa*. (SYN2006PUB)

    (6) *Dnes už* (ADDR nebo ACT) ***máme sepsánu** hospodářskou smlouvu, kde jsou specifikovány veškeré jakostní znaky.* (PDT 2.0)
- (v) ***recip***: ADDR / BEN / PAT (u dvouvalenčních) -> Sb
  - př.

    (7) *Očividně **dostal** dávno **odpuštěno.*** (SYN2005)

    (8) *Ta reagovala odpovědí, že **dostala zakázáno** mi vydat letenku.* (PDT 2.0)

    (9) *Od nynějška **mají** proto oficiálně **povolen** vstup do vlasti i ti členové, kteří …* (SYN2006PUB)

    (10) *Musil jsem se vyhýbat hlavním třídám a náměstím, která jsem **měl** v neděli **zakázána**.* (SYN2005)

Anotace gramatému *diatgram*

- ***res2***: 60 výskytů v PDT 2.0 (z toho 23 dvojznačných)
- ***recip****:* 0 výskytů v PDT 2.0
- ***pas***, ***disp***, ***deagent***: automaticky z údajů v PDT 2.0

- Do jaké míry lze při manuální anotaci PDT 3.0 požadovat rozlišení ***pas*** x ***res1*** x *stav*?
- ***res1***: nejasné hranice ***pas*** – ***res1*** – *(stav na ATS)*:
  - (11a) ***Bylo zavřeno/rozsvíceno*** – ***res1*** nebo ***pas***
  - (11b) *Obchod* ***je otevřen*** – ***res1*** / *Obchod* ***je otevřený***.PAT
  - (12a) *Již v 6 hodin* ***byla*** *světla* ***rozsvícena***. (Štícha, 1980)
  - (12b) *Již v 6 hodin* ***byla*** *světla* ***rozsvícena.res1***, *rozsvěcují je pravidelně v 17,45*.
  - (12c) *Již v 6 hodin* ***byla*** *světla* ***rozsvícena.pas*** pomocí automatického spínače.

- dva významy u aktivní formy dokonavé (týkají se vidu, ne ***diatgram***)
  - (13a) ***Roztrhla.*****cpl** ***si*** *šaty o hřebík*. (**≠** měla šaty roztržené o hřebík)
  - (13b) ***Roztrhla.*****perf** ***si*** *šaty, přesto v nich šla do divadla*. (= měla šaty roztrženy/é)

Anotace gramatému *diatgram* (pokr.)

- anotační experiment:
  - /1/ *Okno je otevřené.*PAT
    - v PDT 2.0 tak je, stejně jako *Tráva je zelená*
    - zůstává stejné (jde o stav, 2 uzly)
  - /2/ tag = Vs.*, rodič = *být*, na *být* závisí ACT (ale ne ACT dogenerovaný)
    - diatgram = ***pas***, sloučení uzlů *být* + Vs („s“ na 2. pozici tagu je part.pas.),
  - /3/ tag = Vs.*, afun = PNOM, vid = všechny hodnoty (včetně vid neurčen), rodič = *být*
    - diatgram = ***res1***, sloučení uzlů *být* + Vs

- I. výsledek: rozhodnutí mezi ***pas*** a ***res1*** se opírá o rozhodnutí anotátorů na ATS a na letmé sondě ve vyhledávkách
- II. novou hodnotu vidu (***perf***) do anotačního schématu nezavádět

Kombinace gramatémů gramatických diatezí a vidu

aspect / diatgram	proc	cpl	perf
act	*Bratr píše dopis.* lit: 'Brother writes / is writing a letter.'	*Bratr napsal dopis.* lit: 'Brother wrote a letter.'	*Bratr napsal dopis.* lit: 'Brother has written a letter.'
pas	*Dopis byl psán Napoleonem.* lit: 'The letter was (being) written by Napoleon.'	*Dopis byl napsán Napoleonem u Borodina.* lit: 'The letter was written by Napoleon near Borodino.'	*Dopis byl napsán, odešli ho.* lit: 'The letter has been written, send it away.'
res1	–	–	*Oběd je uvařen. / Dopis je napsán.* lit: 'The lunch is cooked. / The letter is written.'
res2	–	–	*Matka měla oběd uvařen.* lit.: 'Mother had the lunch prepared. / Mother had had the lunch prepared.'
deagent	*Dopisy se dnes píší na počítači.* lit: 'Today, the letters are being written on computers.'	*Citace se napíšou kurzivou.* lit: 'Quotations will be written in italics.'	*Bábovka se snědla celá.* lit: 'The cake has been eaten whole.'
disp	*Eseje se (mu) píšou snadno.* lit: 'Essays are easy (for him) to write.'	*Esej se (mu) napíše snadno.* lit: 'An essay will be easy (for him) to write.'	–
recip	*Bratr dostává (od otce) vynadáno.* 'lit: Brother gets a scolding (from his father).'	*Bratr dostal (od otce) vynadáno.* lit: 'Brother got a scolding (from his father).'	*Bratr dostal (od otce) vynadáno.* lit: 'Brother has got a scolding (from his father).'

Kombinace gramatémů času a vidu

aspect / tense	proc	cpl	perf
sim	*vaří* lit.: 'she is cooking / she cooks'	–	*má uvařeno / je uvařeno* *uvařila* lit.: 'she has (the meal) cooked / (it) is cooked' lit.: 'she has cooked (the meal)'
anter	*vařila* lit.: 'she was cooking / she cooked'	*uvařila* lit.: 'she cooked'	*měla uvařeno / je uvařeno* *uvařila* lit.: 'she had (the meal) cooked / (it) was cooked' lit.: 'she had cooked (the meal)'
poster	*bude vařit* lit.: 'she will be cooking'	*uvaří* lit.: 'she will cook'	*bude mít uvařeno / bude uvařeno* *uvaří* lit.: 'she will have (the meal) cooked / (it) will be cooked' lit.: 'she will cook (the meal)'

Gramatém skutečnostní modality *factmod*

- nahrazuje gramatém *verbmod* použitý v PDT 2.0
  - *verbmod* jako prozatímní řešení
    - hodnoty (*ind, cdn, imp)* odpovídaly morfologickým způsobům, vyplněny na základě morfologického tagu (nebo jejich kombinace – u analytických forem)
    - *verbmod* se hodnotou *imp* překrýval s gramatémem větné modality *sentmod*
- (indikativ a kondicionál) vs. imperativ
  - indikativ a kondicionál
    - tzv. skutečnostní modalita (factual modality) jako význam slovesa
  - imperativ
    - komunikační funkce věty
- slovesným gramatémem *factmod* budou zachyceny významy vyjadřované indikativem a kondicionálem

Gramatém *factmod* (pokr.)

- indikativ vs. kondicionál
  - kondicionál přítomný vs. kondicionál minulý
- pro *factmod* definovány tři hodnoty
  - ***asserted***
    - děje/stavy konstatované indikativní formou jako reálné
  - ***potential***
    - děje/stavy prezentované kondicionálem přítomným jako potenciální
  - ***irreal***
    - děje/stavy prezentované kondicionálem minulým jako ireálné
- př.
  - *Rekonstrukce bytu **stojí** milion*. (***asserted***)
  - *Rekonstrukce bytu **by stála** milion*. (***potential***)
  - *Rekonstrukce bytu **by byla stála** milion*. (***irreal***)

Anotace gramatému *factmod*

- automatický přepis hodnot gramatému *verbmod* na hodnoty *factmod*
  - hodnoty si přímo neodpovídají
  - zohlednění hodnot gramatému času *tense*
  - -> úpravy v gramatému *tense*
  - -> dílčí revize gramatému větné modality *sentmod*

- kondicionál v závislých klauzích
  - podmínkové klauze se spojkou *kdyby*, účelové klauze se spojkou *aby*
  - klauze s *aby* ve funkci aktantu
    - př. *Řekněte svým dětem, aby to nedělaly* (PDT 2.0) vs. ... *že se představení nekoná* vs. *..., jestli je na představení pustíte* vs. ...

Revize substantivního gramatému čísla

- základní významová opozice
  - jednotlivina vs. větší množství jednotlivin
  - prototypicky vyjádřena singulárem vs. plurálem
- plurálové formy mohou odkazovat k typickému páru/souboru (nikoli prostě k pouhému většímu množství jednotlivin)
  - *jedna bota* vs. *dvě boty* vs. *jedny boty, dvoje boty*
  - párový/souborový význam typický pro substantiva jako *oči*, *vlasy, sirky, rodiče*
  - souborový význam také: *troje stopy ...*
- začlenění významu souborovosti do formálního popisu
- zkušební anotace

Na závěr

- zrevidovaný scénář pro PDT 3.0
  - kategoriální zobecnění, vzájemná (ne)kompatibilita (***diatgram***)
  - odstranění duplicit v anotaci *(verbmod – sentmod)*
  - vyšší teoretická adekvátnost (***factmod****:* funkce kondicionálu přítomného, minulého, nikoli kopírování formy)

- F. Daneš: *Dostal jsem přidáno* a podobné pasívní konstrukce, *Naše řeč* 51, 1968, s. 269–290.
- M. Dokulil: *Tvoření slov v češtině.* Praha, Academia, 1962.
- M. Giger: *Resultativa im modernen Tschechischen*. Bern – Berlin – Frankfurt a. M. – New York – Oxford – Wien, Peter Lang. 2003.
- K. Hausenblas: Slovesná kategorie výsledného stavu v dnešní češtině, *Naše řeč* 46, 1963, s. 13–28.
- V. Mathesius: Slovesné časy typu perfektního v hovorové češtině, *Naše řeč* 9, 1925, s. 200–202.
- I. A. Mel'čuk: *Dependency Syntax: Theory and Practice*. New York, State University of New York Press, 1988.
- M. Načeva Marvanová: *Perfektum v současné češtině*. Praha, Nakl. Lidové noviny, 2010.
- V. P. Nedjalkov (red.): *Typology of Resultative Constructions*. J. Benjamins Publ. House, Amsterdam – Philadelphia, 1988 (viz i *Tipologija rezul'tativnyx konstrukcij*, Leningrad, Nauka, 1983).
- P. Sgall: *Generativní popis jazyka a česká deklinace.* Praha, Academia, 1967.
- H. Skoumalová: *Czech Syntactic Lexicon*. (Ph.D. Thesis). Praha, FF UK, 2001.
- V. Šmilauer: Slovesný čas. In *Druhé hovory o českém jazyce*. Praha, Kruh přátel českého jazyka, 1947, s. 149–165.
- F. Štícha: Konkurence krátkých a dlouhých variant participiálních tvarů v přísudku, *Naše řeč* 63, 1980, s. 1–14.

- J. Panevová: O rezultativnosti (zejména) v češtině. In *Zbornik Matice srpske*, v tisku.
- J. Panevová – M. Ševčíková: Annotation of Morphological Meanings of Verbs Revisited. In *Proceedings of LREC 2010*, pp. 1491–1498.
- J. Panevová – M. Ševčíková: Počítání substantiv v češtině (Poznámky ke kategorii čísla). *Slovo a slovesnost*, připr.
- M. Ševčíková: *Funkce slovesného způsobu z hlediska významové roviny*. Praha, ÚFAL 2009.
- M. Ševčíková – J. Panevová – Z. Žabokrtský: Grammatical number of nouns in Czech: linguistic theory and treebank annotation. In *Proceedings of TLT 2010*, in press.