Obsah / Content

Články / Articles

Recenze / Reviews

Zprávy / News

Články / Articles

Reprodukce řeči/myšlení v mluvených projevech jako předmět korpusového výzkumu

Jana Hoffmannová, Zuzana Komrsková, Petra Poukarová
Ústav pro jazyk český AV ČR, v. v. i., Praha
jana.hoffmannova@ujc.cas.cz; zuzana.komrskova@ff.cuni.cz; petra.poukarova@ff.cuni.cz

Reported Speech/Thought in Spoken Czech as an Object of Corpus Research

ABSTRACT: The article explores reported speech/thought in spoken Czech, especially reproductions introduced with various forms of *říct/říkat* (to say), with data provided by the Czech National Corpus. Most reproductions were introduced by the imperfective verb *říkat* (past and present tenses, first and third persons). By contrast, reproductions of thought were much less numerous and almost invariably involved the first person. We found twice as many examples of direct speech than indirect speech, and interesting transitional forms, some of which can be described as free indirect speech. Pauses separating introductory constructions from reproductions appear to be more typical of direct than indirect speech, but are generally infrequent, suggesting a lower degree of segmentation of spoken language. Sometimes, reproductions of the speech of others were signalled with reduced introductory constructions, with verba dicendi substituted by signals other than verbs, whereas reproductions of one's own speech were normally introduced with a verbum dicendi.

KEY WORDS: reported speech, reported thought, introductory construction, direct/indirect/free indirect speech/thought, reproduction of one's own speech, reproduction of the speech of others

KLÍČOVÁ SLOVA: reprodukce řeči, reprodukce myšlení, rámcový segment, přímá/nepřímá/polopřímá řeč/myšlení, reprodukce řeči vlastní, reprodukce řeči cizí

Deskripce, explanace, reprezentativnost: odpověď Františku Štíchovi

Jan Chromý, Radek Čech
Filozofická fakulta Univerzity Karlovy
jan.chromy@ff.cuni.cz
Filozofická fakulta Ostravské univerzity
radek.cech@osu.cz

Description, explanation, representativeness: A reply to František Štícha

ABSTRACT: The present paper is a reply to the article *Perspektivy korpusové lingvistiky: deskripce, nebo explanace* by František Štícha (2015) which is a critique of recent studies by Radek Čech (2014) and Jan Chromý (2014). It is shown that Štícha's argumentation is based on an inaccurate reading of the two criticized studies. Also, Štícha's conception of corpus linguistics as a discipline which aims to capture the morphological and syntactical norm of well-educated people is rather limited. This narrow-minded view seems to be another reason of Štícha's misunderstanding of the criticized papers.

KEY WORDS: corpus linguistics, explanation, description, representativeness, norm

KLÍČOVÁ SLOVA: korpusová lingvistika, explanace, deskripce, reprezentativnost, norma

Tak v současných mediálních mluvených rozhovorech: *ty zákony co se vztahujou teďkon na nás soukromníky tak to asi budu muset zavřít* (2. část)

Lucie Jílková
Ústav pro jazyk český AV ČR, v. v. i., Praha
jilkova@ujc.cas.cz

The word *tak* in contemporary spoken mass media dialogues

ABSTRACT: The paper tried to answer the following research questions: What are meanings and functions of the Czech word *tak* [*so* in English] in spoken texts used in contemporary media? How is the word *tak* incorporated into a syntactic structure of a turn? Two TV programmes included in the Dialog Corpus were chosen to be analysed: *Otázky Václava Moravce* (a very formal discussion with politicians hosted by Václav Moravec) and *Uvolněte se, prosím* (a very informal, spontaneous talk show). The chosen programmes were intentionally quite different; they noticeably differed even in the total numbers of the *tak* occurrences caught in them: 74 occurrences in *Otázky Václava Moravce* vs. 149 occurrences in the talk show *Uvolněte se, prosím*. In the talk show, we identified an extremely high occurrence of the word *tak* as the preparative particle as it is called, and even a higher number of this word was found as a simple connector. The connector *tak* placed at the end of a line appeared only in the talk show. In both programmes, the word *tak* was commonly used to express various syntactic relations, most often expressing conditions and consequences. The examples often indicated that the word *tak* might be the only word expressing such relations.

KEY WORDS: word *tak*, DIALOG corpus, mass media dialogue, spoken language, meanings and functions of the Czech word *tak*

KLÍČOVÁ SLOVA: slovo *tak*, korpus DIALOG, mediální dialog, mluvený jazyk, významy a funkce slova *tak*

Korpus ORAL: sestavení, lemmatizace a morfologické značkování

Marie Kopřivová, Zuzana Komrsková, David Lukeš, Petra Poukarová
Ústav Českého národního korpusu, FF UK Praha
marie.koprivova@ff.cuni.cz, zuzana.komrskova@ff.cuni.cz,
david.lukes@ff.cuni.cz, petra.poukarova@ff.cuni.cz

The ORAL corpus: construction, lemmatization and morphological tagging

ABSTRACT: The goal of this paper is to provide an overview of the structure and contents of the soon-to-be available ORAL corpus, which combines previously published corpora (ORAL2006, ORAL2008 and ORAL2013) with newly transcribed material into a single conveniently accessible and more richly annotated resource, about 6 million running words in length. The recordings and corresponding transcripts span a decade between 2002 and 2011; most of them capture interactions of mutually well-acquainted speakers, in informal situations and natural settings. The corpus is complemented by a marginal portion of more formal data, mostly public talks. It is tagged and lemmatized, and an effort was made to adapt existing tools (targeted at written language) to yield better results on spoken data. We hope the availability of such a resource will spawn further discussions on the morphological and syntactic analysis of spoken language, perhaps resulting in more radical departures in the future from the part-of-speech classification inherited from the linguistic analysis of written language.

KEY WORDS: spoken Czech, spoken language corpora, lemmatization, tagging, morphological analysis

KLÍČOVÁ SLOVA: mluvená čeština, korpusy mluveného jazyka, lemmatizace, tagování, morfologická analýza

Vladimír Petkevič: Morfologická homonymie v současné češtině

Praha: Nakladatelství Lidové noviny, 2014, 588 stran

Vojtěch Veselý
Ústav pro jazyk český AV ČR, v. v. i., Praha
vvesely@ujc.cas.cz

V knize Vladimíra Petkeviče Morfologická homonymie v současné češtině je prezentován rozsáhlý jazykový materiál: většinu z jejích téměř šesti set stran zaujímá klasifikovaný (a okomentovaný) soupis homonymních slovních tvarů (kap. 3). Této nejrozsáhlejší kapitole, zaujímající přes pět set stran textu, předchází krátká kapitola (kap. 2), v níž jsou na pozadí dostupné literatury vymezeny základní typy homonymie: homonymie lexikální, homonymie slovotvorných prostředků, homonymie morfologická, homonymie syntaktických konstrukcí aj. V poslední části knihy se autor věnuje některým tématům speciálním: morfologické disambiguaci (kap. 4), morfologické homonymii v jiných jazycích (kap. 5) a mezijazykové homonymii (kap. 6).

Jako materiálovou základnu pro výzkum homonymie autor využil především korpusy Českého národního korpusu: SYN2010, SYN2013PUB, SYN. Data z těchto korpusů analyzoval pomocí morfologického analyzátoru a morfologického slovníku (srov. Hajič, 2004), neprovedl však jejich disambiguaci. Získal tak soubor tvarů, kterým byly přiřazeny všechny údaje o lemmatech a morfologických vlastnostech.

Klasifikovaný soupis tvarů zahrnuje převážně jen homonyma částečná, morfologická. Jde o případy, kdy se dvě slova (nebo výjimečně i více než dvě slova) formálně shodují v některých svých tvarech. Základní tvar může, ale nemusí být součástí tohoto souboru shodných tvarů. Autor rozlišuje morfologickou homonymii vlastní, kdy slovní tvar má v rámci tvarového systému lexikální jednotky více interpretací (např. *město* je nominativ, akuzativ nebo vokativ singuláru), a homonymii nevlastní, kdy slovní tvar patří k různým tvarovým systémům jedné lexikální jednotky (např. adjektivní tvar *mladém* lze interpretovat jako lokál singuláru maskulina životného či neživotného nebo jako lokál neutra) nebo k různým lexikálním jednotkám příslušejícím k témuž slovnímu druhu (např. tvar *negativ* je nominativ nebo akuzativ singuláru neživotného maskulina *negativ* nebo genitiv neutra *negativum*; tvar *bělí* náleží ke slovesu *bělet* nebo *bělit*) nebo k odlišným slovním druhům (*loudal* je životné maskulinum nebo činné příčestí slovesa).

Soupis homonymních tvarů je v knize organizován systematicky, v souladu s vymezenou typologií. V rámci slovního druhu substantiv nejprve autor uvádí

(a hojnými příklady dokládá) typy homonymie vlastní, posléze představuje typy homonymie nevlastní; nevlastní homonyma mohou mít stejný, anebo odlišný jmenný rod. Uvnitř slovního druhu adjektiv autor opět rozlišuje homonymii vlastní a homonymii nevlastní, danou odlišným jmenným rodem nebo příslušností k různým druhům adjektiv. Podobně je tomu u dalších slovních druhů: např. u sloves je rozlišena homonymie vlastní a homonymie nevlastní uvnitř jednoho tvarového systému (příčestí *vyvážen* se tvoří od slovesa *vyvážit* nebo od slovesa *vyvážet*) nebo mezi různými tvarovými systémy (*vyjme* je prézens 3. osoby singuláru – s významem futurálním – nebo imperativ 1. osoby plurálu). Poslední část třetí kapitoly je věnována homonymii mezi slovními druhy.

Za každým homonymním tvarem v seznamech jsou uvedena (přinejmenším dvě) lemmata, k nimž daný tvar náleží, např.: *kachně* (*kachna*/*kachně*). Tvary v seznamech jsou řazeny abecedně; autor bohužel neuvádí, zda je pořadí lemmat v závorkách arbitrární, nebo je určeno nějakým pravidlem. Význam lexikálních jednotek zpravidla není okomentován (výjimečně se takový komentář objevuje v poznámce pod čarou), což je vzhledem k rozsahu soupisu homonymních tvarů pochopitelné. V některých případech čtenář pravděpodobně bude muset nahlédnout do výkladového slovníku. Např. *bouře* je tvarem feminina *boura* nebo feminina *bouře* (s. 81). Zatímco *bouře* náleží k jádru české slovní zásoby, *boura* ve významu ‚zevní vrstva zámotku bource morušového' (Slovník spisovného jazyka českého) je termín užívaný v textilnictví.

Vymezení morfologické homonymie autor přejímá od M. Těšitelové (1966, s. 8): „Morfologickou homonymií rozumíme **jakoukoli totožnost** formy (slovoformy, tvarů slov) [při funkčním rozrůznění, doplnil V. P.] v rámci jednoho, popř. dvou i více tvarových systémů." (s. 28) Uvedené vymezení zahrnuje jak případy, kdy shoda mezi formami je náhodná, tak případy, kdy je mezi formami (nenáhodný) významový vztah, tj. jde o polysémii. Lze přitom rozlišit čtyři typy vztahu mezi homonymními formami a jejich lexikálním významem: homonymní formy mají 1. týž lexikální význam, tj. náleží ke stejné lexikální jednotce, přičemž jejich význam morfologický je a) různý (*pánovi* je tvar dativu nebo lokálu singuláru), nebo b) stejný, tvary se liší jen formálněmorfologicky (výrazy *esej*, *prestiž* jsou neživotná maskulina nebo feminina; autor takové případy označuje jako formální homonymii), 2. lexikální významy, jež spolu nesouvisejí (*tancích* je tvar substantiva *tank* nebo tvar substantiva *tanec*), 3. lexikální významy navzájem související (*zřejmě* je příslovce nebo částice; oba významy spolu souvisejí, podle autora jde tedy o polysémii). Funkční rozrůznění ve výše uvedené definici by bylo vhodné specifikovat, nejspíše jako rozrůznění morfologických významů. Jinak by totiž mohlo jít také o různost významů lexikálních.[1] Jako morfologicky homonymní by pak byly

[1] O tom, že k funkcím jazykového znaku autor počítá i lexikální význam, svědčí, zdá se mi, výklad na s. 12. Zde je citována obecná definice homonymie M. Těšitelové: „Homonymie je totožnost formy při funkčním rozrůznění." (Těšitelová, 1966, s. 5) Autor k tomu poznamenává: „Z hlediska vztahů forem a funkcí v jazyce to znamená, že jistá forma má v jazykovém systému více funkcí, významů." Následně jsou citovány dvě definice polysémie (především) lexikálních jednotek.

hodnoceny i tvary, jejichž morfologický význam je stejný, avšak náleží k různým lexikálním jednotkám. Mohlo by jít o totožnost formy v rámci jednoho tvarového systému (např. *list* ve významu ‚část rostliny' i *list* ve významu ‚kus papíru' mají stejné soubory tvarů) i v rámci různých tvarových systémů (např. substantivum *neděle* má ve významu ‚sedmý den v týdnu' v genitivu plurálu dubletu *neděl*/*nedělí*, ve významu ‚týdny' je dané substantivum pomnožné, v genitivu se užívá jen kratší tvar *neděl*).

Problematickou skupinu představují maskulina, která mají tvary životné i neživotné. Na s. 62 autor rozlišuje čtyři typy těchto maskulin: 1. neživotná maskulina, která mohou mít v nominativu a vokativu plurálu životnou koncovku *-ové*: *dnové*, *jazykové*, *národové* (uvedené tvary jsou stylově příznakové), 2. neživotná maskulina mající ve svém paradigmatu životné i neživotné tvary: *ledoborci*/*ledoborce*, *slanečci*/*slanečky*, *křemenáči*/*křemenáče*, 3. maskulina, která jsou životná i neživotná: *činitele*/*činitelé* (autor odlišuje maskulinum životné *činitel* od maskulina neživotného *činitel*), *bacili*/*bacily*, 4. neživotná maskulina mající v akuzativu singuláru tvar rovnající se tvaru genitivu singuláru: *hřiba*, *klouzka*, *vira*. Zvláštním případem jsou 5. substantiva, u nichž kategorie životnosti rozlišuje význam: *agenti* vs. *agenty*, *manažeři* vs. *manažery*, *veteráni* vs. *veterány*.[2] Kritéria, na základě kterých lze substantiva zařadit do skupin č. 2, 3 nebo 4, autor bohužel neuvádí. Mohlo by se zdát, že do skupiny č. 4 náleží substantiva, která kolísají jen v akuzativu singuláru; proč by pak ale jedním z příkladů bylo substantivum *klouzek*, které je uvedeno rovněž ve skupině č. 2 (kolísá i v nominativu a vokativu plurálu)? Především není jasné, jak lze rozlišit případy, kdy je k neživotnému substantivu připojena životná koncovka, od případů, kdy si životné a neživotné maskulinum konkurují jakožto lexikální varianty (srov. typy č. 2 a 3). Pro rozlišení by se dalo využít kritérium formálněsyntaktické, resp. pravopisné: o variantnost lexikální (nikoli jen morfologickou) by mohlo jít v těch případech, kdy substantivum kolísá v nominativu plurálu, tj. tehdy, kdy se rozdíl v životnosti projevuje na tvaru predikátového slovesa.

Komentář zasluhuje rovněž homonymie mezi souborovými a druhovými číslovkami. Tvary těchto číslovek jsou zpravidla homonymní jen v nepřímých pádech, tvary v přímých pádech jsou heteronymní, srov.: *koupil si dvoje zápalky* (souborová číslovka) vs. *v obchodě se prodávaly dvojí koláče, makové a tvarohové* (druhová číslovka); *bez dvojích zápalek* vs. *bez dvojích koláčů*. Výjimkou je číslovka *jeden*, která vyjadřuje význam souborový i druhový: *koupil si jedny zápalky* vs. *v obchodě se prodávaly jen jedny koláče, makové*. V příkladu *Měl doma pouze jedny boty* (s. 259) číslovka *jeden* podle autora vyjadřuje buď prostý počet párů bot, nebo počet jejich souborů, nebo počet jejich druhů. Je známo, že souborové číslovky v kombinaci se jmény pomnožnými plní funkci číslovek základních, tj. vyjadřují prostý počet jevů. V literatuře lze najít diskusi o tom, zda je v syntag-

[2] Zatímco *veterán* ve významu ‚vojenský vysloužilec' je maskulinum životné, *veterán* ve významu ‚automobil starého typu' kolísá mezi životností a neživotností (obojí užití je doloženo v korpusu SYN).

matu *dvoje boty* tvar *boty* užitím substantiva *bota* a číslovka *dvoje* zde má význam souborový, nebo zda tvar *boty* náleží k pomnožnému substantivu *boty* a forma *dvoje* má význam číslovky základní (srov. Michalec – Veselý, 2016). Rozlišuje-li autor ve větě *Měl doma pouze jedny boty* význam prostého počtu (párů bot) a význam počtu souborů, zřejmě soudí, že jde o různé potenciální významy dané číslovky, nikoli o dvojí interpretaci významu jediného. Není však jasné, o jaký soubor bot by mělo jít, pokud ne o pár. Lze si představit, že by syntagma *dvoje boty* označovalo např. dvě krabice naplněné botami? Vzhledem k tomu, že dané spojení uzuálně vyjadřuje význam ‚dva páry bot', je taková interpretace obtížně představitelná. K příkladům 40a–f na s. 260 by bylo vhodné doplnit kontext, který by jednoznačně rozlišil souborové a druhové významy číslovek. Jako 40a i 40b je kupříkladu označena věta *Hovořili u ševce o celkem trojích botách*; v příkladu 40a jde o užití syntagmatu *troje boty* (*troje* = číslovka souborová), v příkladu 40b o užití syntagmatu *trojí boty* (*trojí* = číslovka druhová).

Zajímavou skupinu představují homonyma vzniklá „užitím homonymních stavebních prvků" (s. 16). Autor uvádí následující příklady: *travička* (‚žena, jež někoho tráví'; zdrobnělina od *tráva*), *lištička*, *dlaždička*. Stojí myslím za upozornění, že uvedené příklady reprezentují různé typy slovotvorné homonymie. Výrazy *lištička* (zdrobnělina od *liška*) a *lištička* (zdrobnělina od *lišta*) jsou homonymní v důsledku připojení téhož sufixu k homonymním základům, kdežto homonymie substantiv *dlaždička* (zdrobnělina od *dlaždice*) a *dlaždička* (ženský protějšek k *dlaždič*) je výsledkem spojení homonymních základů s homonymními sufixy.[3] Příklad *travička* je specifický: jde o homonymii výrazů, které vznikly spojením formálně různých základů s formálně různými sufixy (*travička* ← *travič* + *-k(a)*; *travička* ← *tráv(a)* + *-ičk(a)*).

Na s. 295 autor tvrdí (rozumím-li jeho výkladu správně), že se pasivní tvary na *-n*, *-t* tvoří jen od sloves přechodných. Trpné příčestí však lze utvořit (a užít) rovněž od některých sloves nepřechodných; příčestí pak má inkongruentní formu, např. *porozumět – porozuměno*. S tvořením příčestí souvisí také následující poznámka: Na s. 25 autor uvádí sloveso *topit* jako příklad úplné homonymie, ve skutečnosti však jde o homonymii částečnou. Sloveso *topit* ve významu ‚nořením do vody zabíjet' je tranzitivní, kdežto formálně stejné sloveso ve významu ‚udržovat oheň' je nepředmětové; v prvním případě se trpné příčestí tvoří bez omezení, v druhém má jen formu inkongruentní (*v kamnech bylo topeno*).

Některé příklady slovnědruhové homonymie jsou obtížně pochopitelné, jelikož tvary zahrnuté v seznamech nejsou užity v kontextu (navíc přesně nevíme, jaká autor při rozlišování slovních druhů uplatňuje kritéria). Na s. 391 je uvedena homonymie typu substantivum – číslovka – příslovce: *málo* (*málo*/*málo*/*málo*), *bezpočtu* (*bezpočet*/*bezpočtu*/*bezpočtu*) aj. Lze říci, že *málo* je substantivem, pokud nedeterminuje žádný další výraz (*vystačit s málem*), číslovkou, pokud determinuje počitatelné substantivum (*málo lidí*), a příslovcem, jestliže determinuje

[3] V obou případech jsou slovotvorné základy homonymní teprve po alternaci finálního konsonantu.

výraz nepočitatelný (*málo vody*, *málo se snažit*)? Výraz *bezpočtu* lze kombinovat zřejmě jen s počitatelným substantivem (**bezpočtu vody*, **bezpočtu se snažit*), v jakém kontextu by tedy plnil funkci příslovce? Málo jasná je rovněž slovnědruhová homonymie typu číslovka – příslovce – částice: např. *méně* je podle autora komparativem číslovky nebo příslovce, anebo částicí (s. 505). Na nesnadnost odlišení příslovcí od částic autor upozorňuje v pozn. 326 na s. 522. V případě výrazů jako *kdykoli*, *přinejhorším*, *zejména* aj. (s. 524–525) není zřejmé, v jakém kontextu by se mělo jednat o příslovce a v jakém kontextu o částici.

Diskutabilní je autorův předpoklad (s. 520), že výrazy *kromě*, *namísto* reprezentují homonymii typu předložka – příslovce, srov. např. *Namísto Petra nás navštívil Pavel* (předložka) vs. *Namísto v lese pracoval na stavbě* (příslovce).[4] Výraz *namísto* v druhé větě autor nehodnotí jako předložku nejspíše proto, že neprojevuje rekci. Jak by však týž výraz interpretoval v souvětí *Namísto toho, aby pracoval v lese, pracoval na stavbě*? Obě konstrukce jsou v zásadě synonymní. Srov. také následující příklady: a) *jsme doma vždy kromě pondělí* (předložka), b) *jsme doma vždy kromě v pondělí* (příslovce?), c) *jsme doma vždy, ale v pondělí ne* (spojka), d) *jsme doma, jen v pondělí ne* (částice).[5] Při uplatnění formálních kritérií lze výrazy *kromě*, *ale*, *jen* slovnědruhově zařadit tak, jak je to uvedeno v závorkách. Významově jsou si ovšem blízké: vyjadřují relaci mezi totalizátorem *vždy* a časovým určením (*v*) *pondělí* (v příkladech c) a d) ve spojení se zápornou částicí *ne*); věty a) až d) jsou přibližně synonymní.[6]

Množství typů morfologické homonymie, které V. Petkevič ve své knize vymezuje a bohatě exemplifikuje, je vskutku pozoruhodné. Jednotlivé typy jsou tím zajímavější (překvapivější), čím jsou příslušná homonyma od sebe významově a gramaticky vzdálenější. Soupisy apelativních tvarů jsou zpravidla „úplné“, tj. obsahují všechny tvary, které autor nalezl v korpusech. Žádné jiné dílo podobného druhu pro češtinu (a možná ani pro jiné jazyky) doposud nevzniklo. Bohatství předloženého materiálu je však zároveň atributem, který ztěžuje čtenářskou recepci knihy. Autor ale zřejmě nepředpokládal, že by čtenář seznamy tvarů četl jako souvislý text. Díla slovníkového typu se rovněž nečtou „stránku po stránce“, čtenáři v nich cíleně hledají jisté informace. Na rozdíl od alfabeticky řazených slovníků v knize nelze vyhledávat jednotlivé položky (homonymní tvary): ty jsou primárně tříděny podle morfologických typů, teprve uvnitř nich jsou řazeny alfabeticky. V knize se lze naštěstí dobře orientovat podle obsahu (s. 7–9): pro jednotlivé typy morfologické homonymie jsou vyhrazeny samostatné kapitoly.

Klasifikovaný soupis homonymních tvarů, který tvoří jádro knihy, je nepochybně lingvisticky velmi cenný. Možnosti jeho praktického využití jsou však bohužel v knize jen naznačeny. V úvodu (s. 10) autor vyjadřuje naději, že zjištěné

---

[4] Analogicky je výraz *kromě* zpracován ve Slovníku spisovné češtiny.

[5] Za uvedenou čtveřici příkladů vděčím kolegyni B. Štěpánkové.

[6] Je pochopitelné, že pro korpusovou lingvistiku jsou formální kritéria v jistém ohledu atraktivnější než kritéria významová: poskytují pevnější oporu pro přidělování jednoznačných morfologických značek.

výsledky přispějí „a) k obohacení teoretického popisu češtiny o poznatky související s jevem homonymie, jimž se dosud věnovala malá pozornost, b) ke zkvalitnění softwarových nástrojů pro automatickou morfologickou disambiguaci a úkoly s ní spjaté: je to zvláště gramatické a sémantické značkování korpusů a jejich analýza syntaktická a sémantická, strojový překlad, extrakce informací z textů a vůbec všechny aplikace vyžadující rozpoznání významu homonymních forem v textu“. Není zcela jasné, jakým jevům z oblasti homonymie se dosud věnovala malá pozornost a jak by zjištěné poznatky mohly přispět k teoretickému popisu češtiny (složka materiálová v knize výrazně převažuje nad složkou teoretickou, interpretační). Morfologické disambiguaci je věnována čtvrtá kapitola knihy. Autor v ní podrobně charakterizuje disambiguační metody užívané při počítačovém zpracování jazyka, nevysvětluje však, jak konkrétně by klasifikovaný soupis homonymních tvarů mohl být využit při vývoji softwaru pro morfologickou disambiguaci.

V souvislosti s morfologickou disambiguací se nabízejí např. tyto otázky: Existuje vztah mezi typem morfologické homonymie (tím, zda jde o homonymii vlastní či nevlastní, homonymii v rámci slovního druhu či mezi slovními druhy aj.) a pravděpodobností, že bezprostřední kontext tvaru odstraní jeho homonymii? Uvažme situaci, kdy se v bezprostředním kontextu tvaru A nachází homonymní tvar B. Je homonymie tvaru B překážkou pro disambiguaci tvaru A? Je pro zodpovězení této otázky relevantní typ homonymie tvaru A? Srov. např. syntagmata *večerní pečení* a *večerní pečeně*: Tvar *pečení* vykazuje vlastní i nevlastní homonymii (náleží k substantivu *pečení* nebo k substantivu *pečeně*); homonymní adjektivum *večerní* tuto homonymii neodstraňuje. Tvar *pečeně* vykazuje vlastní homonymii i nevlastní homonymii mezi slovními druhy (náleží k substantivu *pečeně* nebo k adverbiu *pečeně*); vlastní homonymii substantiva *pečeně* adjektivum *večerní* neodstraňuje, vylučuje však, že by tvar *pečeně* mohl být adverbiem. Lze vyslovit domněnku, že bezprostřední kontext umožňuje disambiguaci tvaru tím spíše, čím jsou od sebe příslušná homonyma gramaticky vzdálenější.

Recenzovaná kniha je přínosná zejména tím, že – na základě analýzy rozsáhlého jazykového materiálu – představuje úplný (nebo relativně úplný) přehled typů morfologické homonymie v češtině. Doufejme, že poslouží jako inspirace i zdroj dat pro další studie věnované homonymii v jazyce.

LITERATURA

HAJIČ, J. (2004): *Disambiguation of Rich Inflection (Computational Morphology of Czech).* Praha: Univerzita Karlova v Praze, Nakladatelství Karolinum.

MICHALEC, V. – VESELÝ, V. (2016): K významu substantiv s převahou plurálových tvarů. *Slovo a slovesnost*, 163–184.

Slovník spisovné češtiny pro školu a veřejnost. (1994). 2., upr. vyd. Praha: Academia.

TĚŠITELOVÁ, M. (1966): *O morfologické homonymii v češtině*. Praha: Academia.

Co je v ČNK nového VIII
(Zprávy z Českého národního korpusu)

Michal Škrabal
Ústav Českého národního korpusu Filozofické fakulty UK
michal.skrabal@ff.cuni.cz

V tomto pokračování seriálu o novinkách v projektu Český národní korpus bychom se rádi zaměřili především na novou verzi databáze překladových ekvivalentů ***Treq***,[1] zmiňované kol. A. Rosenem v předminulém díle Zpráv z ČNK (*KGA* 11/2015). Ta čerpá z dat paralelního korpusu ***InterCorp***, vždy z jeho poslední verze, aktuálně tedy z verze 9, uveřejněné vloni v září (chystá se nicméně verze 10, která bude zpřístupněna na jaře 2017). Do ní přibyl nový jazyk: romština, počet cizích jazyků tak vzrostl na 39 (vedle tzv. *pivotu* – češtiny, vůči níž jsou všechny texty v korpusu zarovnávány). Nově morfologicky označkovány a lemmatizovány byly tři další jazyky: srbština, chorvatština a lotyština.[2] Celkový počet slov v cizojazyčných textech se momentálně blíží půldruhé miliardě (1,46 mld.), z toho 232 milionů tvoří tzv. *jádro*, tedy ručně korigované beletristické texty s kontrolovaným zarovnáním, zbylých 1,229 mld. slov spadá do tzv. *kolekcí*,[3] jež obsahují texty zpracovávané nikoliv manuálně, ale automaticky; žánrově jde o různé typy textů od legislativních přes publicistické a zpravodajské až po záznamy jednání z europarlamentu či filmové titulky.[4]

Nástroj Treq si navzdory své krátké existenci (prvotní verze pochází ze září 2014) stačil získat oblibu mezi uživateli zejména pro svou jednoduchost a přímočarost.[5] Bezprostředně po zadání hledaného výrazu ve zdrojovém jazyce dostaneme soupis kandidátů na jeho protějšky v jazyce cílovém i s frekvencí těchto ekvivalencí (absolutní i vyjádřenou procentuálně); často přitom platí, že čím je tato frekvence vyšší, tím pravděpodobněji je daný kandidát funkčním ekvivalentem.

---

[1] https://treq.korpus.cz.

[2] Celkový počet otagovaných jazyků tak aktuálně činí 23, lemmatizovaných jazyků je 20.

[3] Česká složka InterCorpu přesahuje aktuálně 186 milionů slovních tvarů, z toho 97 milionů tvoří jádro, zbylými téměř 90 miliony jsou zastoupeny jednotlivé kolekce.

[4] Podrobnější informace o korpusu InterCorp včetně jeho poslední verze najdete na naší wikipedii na http://wiki.korpus.cz/doku.php/cnk:intercorp.

[5] Za rok 2016 bylo na portálu www.korpus.cz evidováno přes 719 tisíc dotazů; nejhojněji využívaným nástrojem je KonText (s více než 85 %), následovaný právě databází Treq (více než 70 tisíc dotazů, tj. bezmála 200 denně, což představuje necelých 10 % z celkového počtu zadaných dotazů).

Zásadní výhodou je možnost ověřit si jednotlivé realizace kteréhokoliv z nich přímo v KonTextu pomocí hypertextového odkazu – a snáze tak odlišit relevantní kandidáty od těch zavádějících. K tomu je ovšemže zapotřebí jistá obeznámenost (byť jen elementární) s daným zdrojovým a/nebo cílovým jazykem, případně též znalost příslušného diskursu. V tomto ohledu je databáze Treq nástrojem sloužícím často k aktivaci pasivních znalostí či k potvrzení uživatelových domněnek; může posloužit ale i k vyhledání inovativního překladu motivovaného stylistickými potřebami. Vzhledem k tomu, že výsledné soupisy slov vznikají na základě automatického zarovnání pomocí nástroje GIZA++[6] a nejsou dále nijak revidovány, je databázi Treq potřeba brát s příslovečným zrnkem soli a v žádném případě ji nelze zaměňovat za regulérní slovník. Měla by sloužit spíše jako jeho doplněk, který kýžené překladatelské řešení nenabízí automaticky, může však uživatele alespoň navést správným směrem. Oproti klasickým slovníkům odráží ve větší míře specifičnost některých případů mezijazykové ekvivalence, kromě frazeologičnosti např. též odlišnou slovnědruhovou platnost výrazu a jeho cizojazyčného ekvivalentu či jejich neparitní zarovnání (tj. v poměru 1:n / n:1). Zároveň nabízí nesrovnatelně bohatší exemplifikaci, všechny příklady nadto pocházejí z autentických, dále neupravovaných zdrojů.

Verze 2.0 databáze Treq přináší četná vylepšení: kromě uživatelsky přívětivějšího a přehlednějšího rozhraní (intuitivnější políčka Výchozí/Cílový jazyk, možnost omezit hledání na konkrétní kolekce, snazší přepínání mezi volbami aj.) přibyla především možnost zadávat do dotazovacího řádku víceslovné jednotky – a to v obou směrech – a očekávat výsledky jak jednoslovné, tak víceslovné. Uživatelé mají na výběr mezi oběma možnostmi (spolu s jinými, už prve dostupnými eventualitami, např. nerozlišováním mezi velkými a malými písmeny). Potenciál Trequ se tím podstatně rozšiřuje, např. pro kombinaci angličtina-čeština lze nyní vyhledávat mnohé třídy slov: frázová slovesa, diskursní částice (*discourse markers*), fráze v obecném smyslu aj., v opačném směru např. reflexivní slovesa. Mimoto nynější výsledky věrněji odrážejí reálný jazykový stav: ekvivalenci lexémů ve zdrojovém a cílovém jazyce nelze pochopitelně omezovat na „ideální" poměr 1:1. Užitečnost rozšíření funkcionality Trequ o víceslovné jednotky se prokázala při četných zkusmých sondách, a to i navzdory vyššímu počtu zavádějících kandidátů. Ti se nicméně na čelných pozicích frekvenčního seznamu objevují jen zřídkakdy; navíc je lze v případě potřeby snadno vypnout a přejít k ekvivalentům jednoslovným. Nadále platí, že zdaleka ne všechny nabízené ekvivalenty, včetně těch nejčastějších, jsou vhodné pro ten který kontext. Frekvenční hledisko, jakkoliv je pro uživatele Trequ primárním vodítkem, nelze absolutizovat – mnohá inspirativní řešení je možné najít i mezi ekvivalenty doloženými pouze jednou (jakkoliv právě v této kategorii převažují ty klamné).

---

[6] OCH, F. J. – NEY, H. (2003): A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29 (1), 19–51. Přesný postup extrakce slovníkové databáze Treq je popsán v článku M. Škrabala a M. Vavřína „Databáze překladových ekvivalentů Treq" v posledním čísle *Časopisu pro moderní filologii* (2/2017). Jeho součástí je též případová studie prověřující potenciál Trequ při překladu krátkého anglického textu do češtiny.

S víceslovnými výrazy se prohloubila potřeba zakomponovat také dotazovací jazyk, který by umožňoval dotazy pomocí regulárních (zástupných) výrazů; dosud totiž Treq vyhledával pouze přesné řetězce znaků. Kromě toho byl pro verzi 2 rozšířen primární jazyk slovníků (doposud jen čeština) o angličtinu: vedle oboustranných česko-cizojazyčných slovníků tak byly z dat v InterCorpu automaticky vygenerovány oboustranné slovníky anglicko-cizojazyčné. Možnost využívat Treq se tedy otevírá mnohem širšímu publiku než dosud, uživatelé už nejsou limitováni nutností ovládat češtinu.

Práce na vývoji databáze Treq pochopitelně nekončí. Další zlepšení poskytovaných výsledků lze očekávat úměrně s narůstajícím objemem dat,[7] větší žánrovou pestrostí textů a také s postupným zlepšováním nástrojů na automatické zarovnávání slov. Ty však nikdy nebudou co do přesnosti srovnatelné se zarovnáním ručním: jistá míra chybovosti je tudíž nevyhnutelná a je třeba ji brát jako nutnou daň za možnost efektivně zkoumat data v objemu manuálně nezpracovatelném.

[7] V tomto ohledu bude logicky rozdílná velikost a kvalita slovníku např. česko-anglického / anglicko-českého (anglická složka je v rámci InterCorpu vůbec největší, činíc bezmála 120 milionů slov, přičemž zastoupeny jsou všechny žánrové typy) oproti česko-vietnamskému / vietnamsko-českému (vietnamština je v InterCorpu reprezentována pouze filmovými titulky čítajícími méně než půldruhý milion slov).