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

1. ÚVOD

V KGA 12/2015 uveřejnil František Štícha článek *Perspektivy korpusové lingvistiky: deskripce, nebo explanace?* (Štícha, 2015), v němž reaguje na naše dva texty z korpusového dvojčísla Naší řeči. Jednak šlo o text *Jen popis s čísly? Perspektivy korpusové lingvistiky* (Čech, 2014), jednak o text *Korpus a reprezentativnost* (Chromý, 2014). Štíchova kritika byla ostrá, avšak nepřesná, založená spíše na neporozumění tomu, co jsme chtěli říci. Považujeme proto za potřebné na tuto kritiku reagovat.

Nejprve bychom však rádi upozornili na jednu obecnější věc. Štícha píše, že „se zdá […] jako by vypukala nová bitva v již mnohaleté, spíše latentní válce o pozice ‚korpusové lingvistiky' v komplexu lingvistických nauk" (s. 75) a také, že „někteří lingvisté, místo aby tento jedinečný nástroj poznávání jazyka [tj. korpus] vítali, dělají pravý opak – hledají všemožné záminky k tomu, jak lingvisty od této práce odradit" (s. 82). Podle nás se o žádnou bitvu či válku nejedná a naším cílem není kohokoliv od práce s korpusy odrazovat. Korpusy jsou podle našeho názoru důležité a užitečné nástroje lingvistického výzkumu. To ovšem neznamená, že nemá smysl – věcně a nikoliv emocionálně – debatovat o obecnějších otázkách, které se práce s nimi týkají. V prvé řadě má smysl se zaměřit na to, co nám korpusy vlastně umožňují popisovat, jaké jsou tedy možné cíle (korpusově) lingvistického výzkumu, a v kontrastu s tím se zamýšlet nad tím, co všechno by mohlo

být náplní korpusové lingvistiky a co už leží za hranicemi možností korpusově založeného výzkumu. Tyto otázky jsou pak nedílně spjaty s problematikou reprezentativnosti korpusů, tedy toho, nakolik jazykový vzorek (tj. korpus) odpovídá jazykové populaci (tj. jazykové realitě).

2. CÍLE KORPUSOVÉ LINGVISTIKY

Korpus je nástroj, který nám umožňuje nejrůznější analýzy, přičemž lze předpokládat, že spektrum těchto možností se bude do budoucna ještě zvětšovat (například v souvislosti s rozvojem multimodálních korpusů). Obecně lze říci, že korpusy můžeme využívat jak pro deduktivně, tak induktivně zaměřený výzkum (srov. zde analogickou distinkci corpus-based vs. corpus-driven, Cvrček – Kováříková, 2011; Tognini-Bonelli, 2001), můžeme jeho prostřednictvím analyzovat nejrůznější jazykové roviny, můžeme se zaměřovat na jazykovou změnu či na rozdíly v užívání určitých prostředků u různých typů mluvčích, na různých územích či v průběhu života atd. Multimodální korpusy nám pak umožňují sledovat užívání jazyka v souvislosti s nonverbálními jevy a v interakci.

Zdá se tedy, že korpus je univerzální nástroj sloužící téměř jakýmkoliv lingvistickým účelům. To by však podle našeho názoru nemělo vést k nekritickému postoji ke korpusům jako takovým a k pracím na nich založeným. V prvním ze Štíchou kritizovaných textů představuje Radek Čech domněnku, že korpusy jsou dosud využívány v omezené míře a že jsou to „nástroje, které umožňují mnohem více, než jen popisovat a třídit jazykové jevy" (s. 172). Popis ve smyslu klasifikace je pro Čecha „prvním nezbytným výzkumným krokem", neměl by však být „krokem finálním" (s. 175). Čech tvrdí, že tento krok představuje experimentální přístup, v kterém můžeme ověřovat teoreticky odůvodněné hypotézy (s. 180). To podle Čecha umožňuje explanaci jazykových jevů. Štíchova kritika této představy je vedena z bytostně deskriptivistických pozic a roli explanace apriorně snižuje, aniž by bylo vysvětleno, v čem je explanace nepotřebná či špatná. Štícha ve své argumentaci používá argumentační klamy – například důležitost explanace snižuje tím, že říká, že „když se lidstvo vymanilo z pout religiózní a filozofické explanace a začalo systematicky poznávat a popisovat přírodu, začaly se hromadit poznatky a vznikal jeden vědecký obor za druhým" (s. 82). Představu experimentálního přístupu, ve smyslu, jak je známá a běžně praktikovaná v řadě přírodovědných oborů, odsuzuje jako „naivní" (s. 76), aniž by uvedl proč. Jediným argumentem je odkaz na blíže nespecifikované důsledky experimentálního přístupu v lingvistice, které jsou prý popsány v jeho dvou recenzích – bez náležitého bibliografického odkazu ovšem těžko dohledatelných. Štíchova kritika je postavena na obraně deskripce, která se mu asi jeví být ohrožena Čechovým požadavkem pokusit se jevy nejen popsat a třídit, ale také vysvětlit jejich povahu. V tom je však pomýlená, protože ani jeden z nás ve svých textech nepovažuje deskripci za apriorně špatnou, nedůležitou či dokonce zbytečnou. Radek Čech se pouze snaží upozornit na to, že bychom neměli u deskripce zůstávat a že bychom se měli snažit také o něco více, než je induktivní hromadění poznatků. Dále, Štíchovo porozumění tomu, co explanace znamená, je přinejmenším nestandardní – o tom svědčí například jeho poukaz na povahu objevů, za kterou jsou udělovány Nobelovy ceny

(s. 77). Ty jsou totiž v oblasti přírodních věd udělovány vesměs za vysvětlení toho, jak se jevy chovají, nikoliv za jejich popis a třídění (jak vyplývá z jeho textu). Například A. Einstein dostal Nobelovu cenu za fyziku z r. 1921 „for his services to Theoretical Physics, and especially for his discovery of the law of the photoelectric effect" (viz <https://www.nobelprize.org/nobel_prizes/physics/laureates/>, cit. dne 25. 10. 2016), čili za objevení zákona, na jehož základě je daný jev vysvětlen. Podobná zdůvodnění lze najít snad u všech laureátů v oblasti fyziky a chemie, jak se o tom může každý přesvědčit na oficiálních stránkách nadace.

Odkud se bere tak sveřepá potřeba obhajovat a upřednostňovat deskripci a zpochybňovat důležitost explanace? Domníváme se, že vyplývá z příliš úzkého pojetí korpusové lingvistiky Františka Štíchy. Ve své kritice totiž přichází s jedním tvrzením, které je hodno pozornosti. Říká, že „hlavním smyslem korpusových statistik je, pokud jde o gramatiku, zjišťování morfologických a syntaktických norem: ať jde o normu při volbě koncovky, o volbu slovotvorného způsobu či sufixu, o volbu slovosledu" (s. 77). Takový přístup je jistě možný, považujeme však za sporné, nakolik je to tvrzení platné obecně. Navíc, pro Štíchu má smysl zkoumat (tj. v jeho pojetí zjišťovat normu) pouze některé texty, konktrétně texty většiny vzdělaných lidí (s. 78), které jediné jsou hodny toho, aby byly analyzovány, protože to je „rozhodující nejen pro národní kulturu, ale i pro nepřímé poznávání charakteru a činnosti vzdělaných a zdravých myslí" (s. 78). Tento přístup připomíná „lingvistickou eugeniku" – nositelem normy národního jazyka je zde privilegovaná vrstva lidí, jen jejich užívání jazyka se má zkoumat a následně odrážet v gramatikách, slovnících atp. Pokud bychom možnosti, které nám poskytuje korpus, takto omezovali, respektive pokud bychom obecně upřednostňovali právě Štíchou uváděné cíle, opomíjeli bychom spektrum dalších možností, které nám korpusy nabízí. Například i lidé s nízkým vzděláním a slabou úrovní psaného či mluveného projevu používají jazyk tak, že jsou komunikačně úspěšní. Musí tedy zřejmě existovat určité mechanismy (a zřejmě i jazykové zákony), které vedou ke komunikační úspěšnosti a kterými se řídí jak jazyk vzdělanců, tak i jazyk lidí se základním vzděláním. Objevit takové zákony by bylo bez jazykových korpusů zřejmě dost obtížné (ne-li nemožné).

Štíchova kritika je mnohdy založena na zásadním nepochopení, a někdy dokonce i dezinterpretaci toho, co tvrdíme. Jednoznačně je to vidět v případě jeho komentáře týkajícího se frekvenčních poměrů, kdy Štícha tvrdí: „Pokud jde o zjišťované frekvenční poměry, Radek Čech namítá: ‚Intuitivně víme, že 90% rozdíl ve výskytu s velkou pravděpodobností znamená rozdíl významný […]. Ale jak je to s jinými poměry, které nejsou na první pohled evidentní – 5 %, 10 %, […]' Ale copak 10 % není opakem 90 %? Jestliže např. zjišťujeme, že ze dvou konkurenčních koncovek […] nebo ze dvou slovosledných struktur […], jež nazveme koncovka a struktura A a koncovka a struktura B, má koncovka a struktura A 90 % výskytů, pak konkurenční koncovka a struktura B má 10 % výskytů" (s. 78). Štícha si bohužel nevšiml, že Čech píše o procentuálních *rozdílech* a jejich vyhodnocení. Mimochodem, Štíchův ilustrativní příklad vykazuje 80% rozdíl mezi výskyty variant A a B (v případě 90% rozdílu by šlo o poměr variant 5 % vs. 95 %, v případě 10% rozdílu o poměr 45 % vs. 55 %). Kromě obrany toho, že „jde o informačně cenný fakt,

který bychom bez korpusu nezjistili", a s Čechovým původním textem nesouvisející poznámky o „zajímavosti" však tento příklad neobsahuje nic, co by se týkalo problematiky měření a interpretace velikosti rozdílů (srov. Čech, 2014, s. 178n). O tom, že sledování poměru frekvencí může být „informačně cenný fakt, který bychom bez korpusu nezjistili", nepochybujeme. Jak to však souvisí s původním Čechovým textem, nám ale není jasné.

Zcela mimo kontext původního Čechova textu jsou pak odstavce týkající se vztahu doloženosti (resp. nedoložitelnosti) a noremnosti (resp. nenoremnosti) (s. 78–79), kterých si prý Čech není vědom.

Dodejme, že se Štíchovým závěrem, že jsme „nepodali ani jeden rozumný a obhajitelný důvod pro to, proč by výzkum gramatiky současné psané češtiny neměl být založen na práci s velkým elektronickým korpusem" (s. 82), zcela souhlasíme. Problém je, že nic takového nebylo naším cílem.

3. REPREZENTATIVNOST

Pojmu reprezentativnost v souvislostech korpusové lingvistiky je věnován druhý kritizovaný text (Chromý, 2014). Chromý se snaží reprezentativnost jako jeden z nejzákladnějších a důležitých pojmů empirické vědy a statistiky aplikovat na korpusovou lingvistiku. Upozorňuje přitom, že je reprezentativnost v korpusové lingvistice chápána odlišně, což považuje za problematické, protože se v korpusové lingvistice se statistickou analýzou dat běžně pracuje. Reprezentativnost je standardně chápána jako míra, s jakou vzorek (v tomto případě korpus) odpovídá populaci (tj. reprezentuje ji). Chromý dochází k tomu, že takzvané reprezentativní korpusy češtiny jsou ze statistického hlediska nereprezentativní, a poukazuje na problémy, které to způsobuje z hlediska jejich využití.

Štíchova kritika se zde s kritizovaným textem opět míjí. Především je autorovi neprávem podsouváno, že „pochybuje o tom, zda lze vůbec nějaký korpus pokládat za reprezentativní" (Štícha, 2015, s. 81), což je v kontrastu s původním textem, ve kterém je reprezentativním korpusům věnován celý oddíl (Chromý, 2014, s. 191–192). Štícha se i v tomto případě vyhýbá věcné argumentaci zaměřené na tvrzení v původním článku a místo toho se věnuje obraně korpusu jako užitečného nástroje. Nepředkládá však jediný argument pro to, že by tzv. „reprezentativní" korpusy češtiny (například SYN2010 či SYN2015) byly skutečně reprezentativní ve statistickém slova smyslu. Místo toho prostě tvrdí, že reprezentativní jsou, protože jsou velké (s. 81), a vyzdvihuje to, že textů v korpusu je více, než běžný lingvista během svého života přečte (s. 79). Jenže velikost korpusu sama o sobě reprezentativnost nezaručuje (i když s ní určitým způsobem souvisí). Pokud bychom například vytvořili korpus na základě hodinových rozhovorů s celkově 100 000 lidmi z Prahy, stěží bychom mohli vůbec předpokládat, že to bude korpus reprezentativní pro celou ČR, ačkoliv to bude korpus obrovský (alespoň v kontextu mluvených korpusů). Argument, že korpus přesahuje lingvistovu introspekci, je ve Štíchově textu opět úhybným manévrem, protože Chromý v kritizovaném textu netvrdí opak, ani neargumentuje pro to, že by využití introspekce mělo být vhodnější než použití korpusu. O introspekci Chromý dokonce vůbec nemluví.

Zajímavé je, že Štícha spojuje reprezentativnost s kvalitou. Píše, že „pokud jde někomu o to, že z popisu gramatiky určitého jazyka dané doby unikne nějaký poznatek jen proto, že v tomto korpusu schází nějaký návodový text, psaný třeba navíc špatnou překladovou češtinou, že v něm chybí zápisy soudních rozhodnutí, kdejaká vnitřní směrnice a všemožné jiné odrůdy jazyka úředního stylu, patřícího k těm nejhorším jazykovým výplodům lidského ducha, pak rád souhlasím s názorem, že korpus nezastupuje celý jazyk“ (s. 79). Jinde si klade otázku: „Jde nám snad o to vědět, jak píše ten nejhorší stylista, jaké gramatické zvláštnosti obsahuje kdejaký návodový text či kdejaká interní směrnice?“ (s. 78) Na tuto otázku si pak odpovídá mimo jiné tvrzením, že „se snad dokážeme spokojit s poznatky o tom, jak píše většina vzdělaných lidí“. Štícha vlastně představuje velmi úzkou představu korpusu jako vzorku textů vyprodukovaných vzdělanými lidmi, a jeho ideálem je tak korpus vymezený negativně, tedy takový, v němž některé texty začleněny nejsou. Stejně jako v případě představy obecného cíle korpusové lingvistiky je i zde Štíchovo pojímání věci úzké a nezohledňuje řadu možností, které nám korpus jako vzorek textů může nabízet.

4. ZÁVĚR ANEB PROČ SI NEROZUMÍME

Domníváme se, že Štíchova kritika našich dvou textů je založena na nepochopení toho, o co nám šlo. Naší snahou bylo kriticky přistoupit k některým aspektům korpusové lingvistiky, které jsou běžně chápány jako neproblémové. Nikoliv však s cílem vyvolání nějaké bitvy, nebo snad kvůli tomu, abychom od korpusové lingvistiky někoho odváděli, nýbrž s cílem tyto aspekty problematizovat, učinit je předmětem zamyšlení a diskuse. Taková problematizace základních prvků určité disciplíny (v tomto případě korpusové lingvistiky) je podle našeho názoru důležitá, a pokud je oprávněná, může být disciplíně jedině ku prospěchu.

Štícha ve svém textu deklaruje úzké chápání korpusové lingvistiky jako disciplíny, jejímž cílem je primárně zachytit morfologickou a syntaktickou normu vzdělaných lidí. Domníváme se, že toto úzké vnímání problematiky může být jednou z příčin jeho neporozumění našim argumentům a z toho plynoucí potřeby postoje spočívajícího v obraně korpusů a korpusové lingvistiky jako takových.

LITERATURA

CVRČEK, V. – KOVÁŘÍKOVÁ, D. (2011): Možnosti a meze korpusové lingvistiky. *Naše řeč*, 94, 113–133.

ČECH, R. (2014): Jen popis s čísly? Perspektivy korpusové lingvistiky. *Naše řeč*, 97, 171–184.

CHROMÝ, J. (2014): Korpus a reprezentativnost. *Naše řeč*, 97, 185–193.

ŠTÍCHA, F. (2015): Perspektivy korpusové lingvistiky: deskripce, nebo explanace. *Korpus – gramatika – axiologie*, 6, 75–82.

TOGNINI-BONELLI, E. (2001): *Corpus Linguistics at Work*. Amsterdam – Philadelphia: John Benjamins.

GRANTOVÁ PODPORA

Tento text vznikl za podpory projektu Univerzity Karlovy Progres č. 4, Jazyk v proměnách času, místa, kultury.