RADEK ČECH

Jen popis s čísly? Perspektivy korpusové lingvistiky

Just a description with numbers? Perspectives of corpus linguistics

The aim of the article is both to point to the descriptive character of a majority of corpus linguistic analyses and to argue that this character (which is manifested by the classification, sorting or labelling of language data) represents a limit in corpus linguistic research. Further, an experimental approach (in the sense of empirical testing of a hypothesis) is proposed as a possible way of overcoming this limit. Finally, some methodological aspects of current state of corpus linguistics, namely the notion of representativeness and the interpretation of quantification, are critically discussed.

Key words: corpus linguistics, description, hypothesis, quantification

Klíčová slova: korpusová lingvistika, popis, hypotéza, kvantifikace

„The lack of hypotheses is almost pathological."
(Altmann, 2012, s. 11)

„… it may appear surprising that statistical methods are not that widespread in linguistics. This is all the more surprising because such methods are very widespread in disciplines with similarly complex topics such as psychology, sociology, economics. To some degree, this situation is probably due to how linguistics has evolved over the past decades…"
(Gries, 2009, s. 4)

1 Úvod

Vznik elektronických jazykových korpusů a zejména jejich obecná dostupnost v posledních zhruba patnácti letech bezpochyby přinesly do lingvistiky výrazný impuls. Bez levných osobních počítačů a internetu by ale korpusová lingvistika (dále KL) byla pravděpodobně jen záležitostí hrstky lingvistů, kteří by měli to štěstí mít k dispozici technická zařízení pro většinu zcela nedostupná.[1] V důsledku toho by i počet jazykových korpusů byl vzhledem k dnešku asi zanedbatelný (stejně tak i jejich kvalita). S jistou nadsázkou snad můžeme říct, že za rozvoj KL z velké části vděčíme jednak počítačovým inženýrům, jednak komplexu nejrůznějších faktorů (obchodních, politických, sociálních atp.), který vedl k tomu, že dnes pro většinu lidí není problém mít osobní počítač připojený k internetu nejen

[1] Srov. vzpomínku J. Svartvika na dobu počátků KL: „At the time, processing large amounts of linguistic data required enormous computer resources. Henry Kučera mentioned that, when they produced the concordance of the corpus, they had to use the complete mainframe capacity of the Brown University Computer Unit – in all other university departments, computer-based research activities had to be suspended for a full day!" (Svartvik, 2007, s. 20).

v práci, ale i doma. Na první pohled je tento fakt možná jen dalším anekdotickým příkladem toho, že na charakter vědy a její soudobý stav mají vliv nejrůznější okolnosti, které s daným vědním oborem zpravidla nemají mnoho společného, a že historie vědy je spíše než směřováním k stále „dokonalejšímu" či „pravdivějšímu" poznání sledem událostí, jež jsou výsledkem vlivu nejroztodivnějších příčin (srov. Feyrabend, 2001). Domnívám se však, že pokud se chceme pokusit porozumět současnému stavu KL, diskusím a sporům, které vedou její protagonisté, a vlastně celé její současné podobě a perspektivám jejího vývoje, je potřeba vzít tento fakt vážně. O co jde? Pomalý nástup KL od konce 50. let 20. století, který lze vnímat jako autentickou reakci na metodologické a teoretické problémy předchozích přístupů (srov. Svartvik, 2007), byl od poloviny 90. let vystřídán jejím obrovským rozmachem. Ten však nebyl způsoben masovou kritickou metodologickou a teoretickou reflexí soudobé lingvistiky. Jak jsem uvedl výše, jednalo se spíše o určitou shodu okolností. Ani pohled na diskuse týkající se metodologie a teorie lingvistiky v odborných časopisech na přelomu 80. a 90. let totiž nenasvědčuje, že by se většina lingvistů začala „bouřit" proti soudobému stavu, a že by se tak vznik jazykových korpusů dal interpretovat jako reakce na stav oboru. Spíše se zdá, že pro většinu lingvistů se jazykový korpus stal fenoménem, který se dost nečekaně objevil a s jehož existencí bylo třeba se nějakým způsobem vyrovnat. A současný stav KL je z velké části právě projevem tohoto „vyrovnávání se". Samozřejmě, přítomnost nového nástroje se může stát (a zpravidla stává) inspirací pro nové myšlenky a rozvoj oboru, což je v případě korpusové lingvistiky reprezentováno zejména tzv. korpusově řízeným (corpus-driven) přístupem – ten vychází z předpokladu, „že v reprezentativním a rozsáhlém vzorku textů je vše, co lingvista k analýze potřebuje" (Cvrček, 2013, s. 219; pro shrnutí viz McEnery – Hardie, 2012). Z určitého nadhledu je však celý rozmanitý proud KL svázán s deskriptivně strukturálním paradigmatem lingvistiky více, než by se snad na první pohled mohlo zdát, a to i v případě „radikálního" corpus-driven přístupu. Osobně se domnívám, že právě tato svázanost je jednou z příčin toho, že možnosti, které nabízejí jazykové korpusy, jsou využívány jen v omezené míře. Jinými slovy (abych se vyhnul případným nedorozuměním): to, že lingvisté používají jazykové korpusy k *jakémukoliv* účelu, je samozřejmě zcela v pořádku; na druhou stranu je jistě užitečné vzít na vědomí, že jazykové korpusy (a s nimi související možnosti rychlého zpracování doposud nevídaného kvanta jazykových dat) jsou nástroje, které umožňují mnohem více, než jen popisovat a třídit jazykové jevy (čehož si možná většina lingvistů pořád není dost dobře vědoma – srov. slova S. Griese v mottu tohoto článku).

V následujících řádcích a) se zaměřím na základní cíl naprosté většiny korpusověLingvistických analýz, tj. *popis* jazyka spočívající v *klasifikaci* jazykových jednotek, a pokusím se ukázat, jak sledování tohoto cíle (podle mého názoru zbytečně) omezuje možnosti poznání toho, jak funguje přirozený jazyk; b) dále

se budu věnovat teoretickým a metodologickým problémům, které používání jazykových korpusů pro lingvistickou analýzu provázejí a se kterými se současná KL podle mého názoru stále nedokázala adekvátně vyrovnat; c) budu ilustrovat, jak pojetí hypotézy, tak jak se s ním pracuje v experimentálních vědách, může znamenat překonání deskriptivního charakteru lingvistické analýzy.

2 Od popisu k… popisu s čísly

Představitelé hlavního proudu české KL se otevřeně hlásí k tradici strukturalismu: na jedné straně sice KL hodnotí jako samostatnou jazykovědnou disciplínu (na úrovni např. kognitivní lingvistiky, generativní lingvistiky nebo psycholingvistiky), na straně druhé tvrdí, že KL „[s]vými metodologickými východisky navazuje na strukturalismus, přináší ovšem do jazykovědného výzkumu specifické nástroje a metody“ (Cvrček – Kováříková, 2011, s. 113). Bez ohledu na to, zda budeme KL považovat za samostatnou disciplínu (záleží na kritériích), praxe, nahlížena z perspektivy korpusověllingvistických publikací, ukazuje, že hlavní rozdíl mezi předkorpusovou lingvistikou a KL opravdu spočívá v metodologických aspektech výzkumu (v prvé řadě je to důraz na empirii, odklon od používání introspekce a uvědomění si významu kvantifikace) a v používání nových nástrojů, tj. elektronických korpusů. V podobném duchu, tj. jako specifický metodologický postup, je hodnocena KL i ve světovém kontextu, za všechny srov. „What is corpus linguistics? It is certainly quite distinct from most other topics you might study in linguistics, as it is not directly about the study of any particular aspect of language. Rather, *it is an area which focuses upon a set of procedures, or methods, for studying language* (although, as we will see, at least one major school of corpus linguists does not agree with the characterisation of corpus linguistics as a methodology).“ (McEnery – Hardie, 2012, s. 1, kurziva RČ).

Metodologie jistě patří k nejdůležitějším aspektům výzkumu. Ovšem volba každé metody je závislá na badatelském cíli, jenž rozhoduje o povaze každého vědeckého výzkumného projektu. Jaké jsou tedy cíle KL? Sledujme je např. prostřednictvím některých citací z článku Cvrčka a Kováříkové (2011), ve kterém jsou vymezeny základní charakteristiky KL (kurziva RČ):

- „korpus si nachází svoji cestu i ke gramatickému *popisu*“ (s. 113);
- „Korpusová lingvistika se tak postupně etablovala jako samostatná jazykovědná disciplína […], která se zaobírá všemi běžnými rovinami jazykového *popisu*.“ (s. 113);
- „se tak setkáváme s nepochopením toho, co korpus a korpusová lingvistika přináší nového a jaké možnosti se tím lingvistickému *popisu* otevírají.“ (s. 114);
- „Frekvence je pro zkoumání jazyka velmi podstatná zejména tím, že podává informaci o centru a periférii jazykových jevů, podle čehož by měl být strukturován i *popis* jazyka“ (s. 116);

- „Korpus je navíc nedocenitelným nástrojem *popisu*, protože je vzorkem skutečného a realizovaného úzu“ (s. 117);
- „Její součástí jsou způsoby vytěžování korpusů, tedy způsoby zkvalitňování a doplňování *popisů* jazyka, objevování nových vztahů a vynalézání nových konceptů sloužících pro *popis* jazykové reality.“ (s. 122);
- „Na základě zkoumání materiálu v korpusu se ale kromě zpřesňování dosavadních vědomostí vytvářejí i nové koncepty *popisu*.“ (s. 126).

Nahlédneme-li do korpusověLingvistických publikací, je na první pohled evidentní, že popisem se v naprosté většině případů myslí *klasifikace* jazykových jednotek, doplněná o informaci týkající se kvantifikace (zpravidla jde o procentuální vyjádření či různá „skóre“) a následnou interpretaci. Pro ilustraci cílů předkorpusové lingvistiky srovnejme citaci z *Mluvnice češtiny 3* (Daneš et al., 1987, kurziva RČ):

- „Úkolem syntaxe […] je podat systematický *popis* a výklad výrazové, významové a komunikativní struktury jazykových jednotek“ (s. 7).

Z hlediska badatelského cíle je tedy KL spíše jen technologicky odlišnou variantou deskriptivně zaměřené strukturální lingvistiky. Není přitom samozřejmě pochyb o tom, že kvantifikace vnáší do popisu/klasifikace nové obohacující prvky, které nám umožňují lépe pochopit fungování jazyka. Ale i kvantifikovaný popis/klasifikace není zpravidla ničím jiným než tříděním prvků daného systému podle *námi zvolených* kritérií. Klasifikace je jistě prvním nezbytným krokem, který nám umožňuje zavést do „chaosu“ prvků určitý řád, odhaluje podobnosti/rozdílnosti mezi prvky systému atp. Explanační síla popisu/klasifikace je však velmi malá. Navíc bývá velmi obtížné (respektive nemožné) rozhodnout, která ze dvou či více klasifikací je lepší – např. jak rozhodnout, zda je lepší popis syntaxe bez větného členu „doplněk“, či s ním? Nebo zda klasifikovat řadové číslovky jako adjektiva, nebo jako číslovky? Je lepší klasifikace příslovečných určení do dvanácti, nebo do dvaceti skupin? Nemožnost rozhodnout mezi jednotlivými způsoby popisu je důsledkem toho, že žádný popis ve smyslu klasifikace nemá pravdivostní hodnotu; jinými slovy, nemáme nástroje, pomocí nichž bychom mohli rozhodnout, zda námi zvolený popis/klasifikace odpovídá nějakým „reálným“ vlastnostem systému, či ne.[2]

Popis jako badatelský cíl je ovšem dobře pochopitelný, jestliže přistoupíme na základní axiom strukturalistického paradigmatu – jazykovou dichotomii ve smyslu langue–parole. Pokud předpokládáme, že za ohromnou variabilitou

[2] V tomto smyslu je velmi cenné sledovat vývoj analytické filozofie, který lze velmi stručně charakterizovat jako cestu od snahy nalézt kritéria pravdivosti (za všechna srov. Wittgenstein, 1993a; Russel, 1975) až k radikálnímu odmítnutí této snahy (opět jen namátkově Wittgenstein, 1993b; Quine, 1991; Rorty, 2012). Přehled tohoto vývoje podává Peregrin (2005).

konkrétních jazykových projevů existuje „daný“ systém jednotek a vztahů mezi nimi, pak se tento cíl zdá být racionálním východiskem. Na druhou stranu „[j]azykový systém [míněno ve smyslu langue, RČ] není bezprostředně zkoumatelný žádnou ze známých lingvistických metod“ (Cvrček – Kováříková, 2011, s. 124), z čehož plynou všechny problémy popisu/klasifikace, jak je popisuji výše (tj. zejména absence kritérií, na jejichž základě by se dalo rozhodnout, která z klasifikací je lepší, ve smyslu lépe odpovídající langue).[3]

Jiným způsobem k obhajobě popisu jako badatelského cíle přistupují představitelé corpus-driven přístupu, kteří předpokládají, že přímé a teorií nezatížené pozorování dat nás dovede k správnému poznání jazyka, protože teorie je „v datech“ samotných (tj. jazykovém korpusu).

Pohled na praxi KL tak odhaluje, že KL je pevně spojena s deskriptivně strukturalistickým přístupem, zejména cílem jazykových analýz. Díky technickým vymoženostem, které představují jazykové korpusy a snadná možnost přístupu k nim, KL de facto „jen“ rozšiřuje rozsah a možnosti popisů, přičemž obvykle přidává informaci o frekvenci (případně kvantifikuje poměry mezi zkoumanými jevy) – není tedy vlastně ničím jiným než popisem s čísly.

Jak jsem uvedl výše, popis ve smyslu klasifikace je prvním nezbytným výzkumným krokem. Neměl by však být, ať už kvantifikovaný, nebo ne, krokem finálním. A to ani v případě, že je doplněn o „sebezajímavější“ interpretaci nejrůznějšího druhu, což je snad nejlépe patrné z pohledu na historii a současný stav jiných vědních oborů. Vždyť přece cílem fyziky není třídit předměty, např. na teplé a studené, ale pokusit se zachytit a vysvětlit mechanismy, které řídí nebo ovlivňují vlastnosti předmětů, např. vztah teploty a tlaku. Podobně i cílem biologie už dávno není klasifikovat jednotlivé organismy, ale např. vysvětlit mechanismy související s dědičností atp. V tomto smyslu se zdá rozumné pokusit se překonávat popisně klasifikační charakter lingvistiky a zaměřit se na analýzu mechanismů, jimiž se řídí naše jazykové chování a jež mají rozhodující vliv na charakter přirozeného jazyka (viz část 5).

3 Nástrahy kvantifikace

V souvislosti s tím, co je napsáno v části 1 tohoto článku, si dovolím tvrdit, že relativně razantní nástup kvantifikace v lingvistice byl způsoben spíše technickými okolnostmi (tj. vznikem jazykových korpusů a nástrojů umožňujících rela-

[3] Zajímavý je postřeh R. Rortyho, který se týká povahy pojmů majících charakter platónského idealismu, mezi něž lze podle mě řadit i pojem langue (srov. Čech, 2005a, 2005b): „Problém platónských pojmů nespočívá v tom, že jsou „nesprávné“, nýbrž v tom, že o nich nelze mnoho říci – zejména je není možné „naturalizovat“ či jinak spojit se zbytkem zkoumání, kultury či života“ (Rorty, 2012, s. 292). S ohledem na lingvistiku bychom se pak mohli zeptat: Ztrácíme něco, pokud se vzdáme pojmu langue? Jaký je jeho reálný vliv na povahu jazykových analýz a jejich interpretaci?

tivně jednoduchý přistup k nim) než teoreticky motivovaným úsilím lingvistů kvantifikovat; výjimku v tomto ohledu tvoří kvantitativní analýzy M. Těšitelové a jejích spolupracovnic a spolupracovníků (Těšitelová, 1980; Těšitelová et al., 1985; Novák, 1964). Tento fakt je možná příčinou toho, že se v KL objevují problémy související právě s kvantifikací, které korpusoví lingvisté podle mého názoru adekvátně neřeší, přestože je na ně často poukazováno.

Jedním z nejvýraznějších je opakovaně diskutovaný problém reprezentativnosti jazykových korpusů.[4] Přestože jde o problém zásadní a *principiální* (srov. Králík, 2013), který má významný vliv na povahu interpretace jazykových analýz (i těch popisně klasifikačních), nejsou námitky proti pojetí reprezentativnosti náležitě vyvráceny ze strany zastánců reprezentativnosti. Například Cvrček a Kováříková (2011) zcela otevřeně poukazují na význam pojmu reprezentativnosti pro KL, srov.: „Korpus je sice rozsáhlým vzorkem, ale stále je jen částí celkové populace, kterou představují všechny texty a promluvy daného jazyka. Je jasné, že průzkum provedený na nereprezentativním vzorku může být významně zkreslený (vzhledem k celkové populaci), stejně jako by byl zkreslený průzkum stranických preferencí provedený pouze na jedné sociální skupině nebo v jednom místě.“ (s. 131). Aniž by však zmínili teoretické námitky proti konceptu reprezentativnosti, konstatují, „1) že všechny výtky vůči reprezentativnosti jsou výtkami proti konkrétnímu sestavení daného korpusu, nikoli proti principu korpusové lingvistiky, a dále pak, 2) že je plně v moci badatele se s tímto nedostatkem vyrovnat sestavením vlastního materiálu na základě nabídky poskytovatele korpusu.“ (s. 132). Nechme stranou bod 2), který je aplikovatelný jen pro případ speciální jazykové analýzy (např. specificky vymezeného žánru, případně analýzy jazyka konkrétního autora atp.) a zaměřme se na reprezentativnost *obecných* jazykových korpusů.

Předně, jakýkoliv obecný korpus je vždy souborem různých textů, jejichž jazyk je ovlivněn autorstvím, žánrem atd. Z toho plyne, že zkoumáme-li nějaký jazykový jev, pravděpodobnost jeho výskytu je pro každý text, autora, dílo atd. jiná, což má následující důsledek: „zvětšujeme-li počet pozorování *pomocí* a *uvnitř* korpusu, obecně se nepohybujeme v homogenním prostoru. Extrémní růst počtu sledovaných jevů proto nemusí nutně znamenat extrémní zpřesnění poznatků. Jsou-li z hlediska axiomů pro každý text (autorské dílo) i pro každý okamžik limitní hodnoty – pravděpodobnosti – odlišné, vzniká otázka, jak vůbec obecně chápat hodnoty zjištěné z korpusu“ (Králík, 2013, s. 50–51). Jinými slovy: znamenají hodnoty či rozdíly mezi pozorovanými hodnotami opravdu to, co si o nich intuitivně myslíme, když pracujeme s rozsáhlými obecnými jazykovými korpusy? To je principiální problém.

[4] Pro podrobný přehled této problematiky viz Chromý (2014).

Otázku reprezentativnosti dále komplikuje fakt, že v jazyce nelze stanovit tzv. základní soubor (populaci), vzhledem k němuž by bylo možné určit míru reprezentativnosti daného vzorku (tj. v našem případě korpusu). V tomto ohledu jsou jednotlivé korpusy, i ty tzv. reprezentativní, spíše jen vzorky (samples) jazyka, sestavené podle konvenčně přijatých kritérií (srov. Šulc, 2001), nikoliv vzorkem reprezentujícím povahu jazyka jako takového (tu prostě nedokážeme vzhledem k jeho charakteru určit).

Jak je vidět, výše uvedené námitky jsou principiální a stoupenci reprezentativnosti by na ně měli adekvátně reagovat. Pro ilustraci uveďme reakci jedné z nejvýznamnějších postav KL, G. Leeche (2007), který na jedné straně uznává, že pojem reprezentativnosti je problematický a nedosažitelný, ale na straně druhé se ho nehodlá vzdát s tím, že je jakýmsi ideálem, ke kterému přese všechnu problematičnost směřujeme: podle něj je reprezentativnost otázkou míry, nikoliv otázkou dichotomickou ve smyslu reprezentativní vs. nereprezentativní. Jeho postoj však celý problém nijak neřeší – jak můžeme totiž mluvit o míře reprezentativnosti, když ji nemáme k čemu vztáhnout?

Kuriózní na celém problému ne/reprezentativnosti je fakt, že pokud přejdeme za hranice popisně klasifikačního přístupu, ztrácí tento pojem v jazykovém výzkumu svůj význam (viz část 5).

Další otázkou, která si při korpusové analýze zaslouží velkou pozornost, je otázka měření a jeho interpretace. Označíme-li nějaký jev číslem, např. spočítáme-li frekvenci určitého jazykového jevu, získáváme informaci, která nám sama o sobě nic neříká. Smysl měření – v KL se jedná zpravidla o měření četnosti výskytu – spočívá v tom, že můžeme různé velikosti (např. frekvence) pozorovaných jevů porovnávat. Na tom samozřejmě není nic složitého ani záludného, pokud se nedopustíme triviálních početních chyb. Zjištěné absolutní hodnoty se v praxi většinou normalizují do formy přehledného procentuálního vyjádření. Jenže pouhé konstatování, že pozorované jevy se vyskytují s různou frekvencí (byť normalizovanou), neříká o moc více než prosté označení jevu číslem. Kvantifikace začíná mít smysl v případě, kdy se snažíme předpokládané (a posléze zjištěné či nezjištěné) rozdíly interpretovat, tj. dáváme do souvislosti změřený rozdíl s vlastností jinou, např. formou jazyka, žánrem, sémantickými vlastnostmi, délkou, komplexitou, morfologickou produktivitou atd. Pokud nám nejde o nic více než o ilustrativní popis, je typickým výsledkem např. gramatika doplněná o informaci různého poměru popisovaných jevů vzhledem k tzv. registrům (Biber et al., 1999). Při pokusech o hlubší interpretaci se však každý badatel musí vyrovnávat s otázkou, jak interpretovat *různou velikost* zjištěných rozdílů. Intuitivně víme, že 90% rozdíl ve výskytu s velkou pravděpodobností znamená rozdíl významný, tj. že je způsoben nějakým předpokládaným mechanismem, třeba vlivem žánru. Ale jak je to s jinými poměry, které nejsou na první pohled evidentní – 5%, 10%, 15%, 30%…? A dále, dobře víme, že procenta mohou vyjadřovat absolutní hodnoty

lišící se v řádech a variabilitě, takže statistická interpretace stejného procentuálního rozdílu může být vzhledem k povaze změřených hodnot naprosto rozdílná (srov. Čech, 2012 a následnou reakci v Cvrček – Kodýtek, 2013). Jinými slovy, pokud nechceme používat kvantifikaci jen jako ilustrativní doplněk popisu, dostáváme se na pole statistiky, která pro neškoleného badatele skýtá řadu nástrah. A vzhledem k tomu, že naprostá většina lingvistů nebyla a není systematicky školena v tomto oboru, dochází k interpretacím, které jsou z hlediska statistiky problematické (eufemisticky řečeno) – stačí nahlédnout do korpusových analýz, v nichž převažují procentuálně vyjádřené rozdíly většinou doplněné komentáři typu, že pozorovaný rozdíl je „malý", „větší", „téměř stejný", že „se velmi liší" apod.

Jak je vidět, používání kvantifikace s sebou přináší řadu netriviálních metodologických důsledků. Osobně se domnívám, že adekvátní vyrovnání se s nimi je v současné době jednou z největších výzev KL.

4 Není hypotéza jako hypotéza

Význam kvantifikace obecně spočívá v tom, že umožňuje vyjádřit velikost rozdílu vlastnosti, jež je pozorována u více jevů. Například můžeme porovnávat délku slov, jejich frekvenci, velikost inventáře fonémů v jednotlivých jazycích atp. Jak jsem uvedl v části 3, analýza pozorovaných rozdílů dostává smysl až v případě, že jsou tyto rozdíly vztáhnuty k nějaké jiné vlastnosti. Konkrétně, například zjištění, že slovo „hypotéza" se ve všech tvarech v stomilionovém korpusu (SYN2010) vyskytuje 1634× a slovo „test" (opět ve všech tvarech) 8242×, samo o sobě nic neříká. Pokud však frekvenci vztáhneme k pozorování délky slov, situace se mění: analýza většího počtu slov velmi rychle odhalí vztah mezi těmito dvěma vlastnostmi, což mimochodem platí i pro slova „hypotéza" a „test". Pouhá korelace (byť velmi silná) však pro pochopení fungování pozorovaných vlastností ještě nemusí vůbec nic znamenat.[5] Analýza vztahu mezi dvěma a více vlastnostmi získává na významu až prostřednictvím *teoretického* zdůvodnění, tj. ve chvíli, kdy začneme o tomto vztahu uvažovat ve formě hypotézy (srov. Zipf, 1935, 1949; Köhler, 2005).

Stejně jako naprostá většina slov je i slovo „hypotéza" slovem polysémním, stačí nahlédnout třeba do *Slovníku spisovného jazyka českého*. V diskurzu experimentálních věd má však tento výraz poměrně jednoznačný význam – stabilita tohoto významu je dána formálními pravidly, na jejichž základě je možné rozlišit, který výrok lze za hypotézu považovat a který nikoliv. Tato pravidla jednak určují formu tvrzení, jednak charakterizují povahu praktických důsledků, které

[5] Srov. korelaci mezi měsíčním průměrným počtem lidí, kteří navštěvují hroby svých předků, a měsíčním průměrným počtem přezutých zimních pneumatik u automobilů – tato korelace je jen výsledkem toho, že zákon nařizuje přezout pneumatiky v termínu, kdy je Svátek zemřelých.

z daného tvrzení musejí vyplývat, abychom mohli dané tvrzení označit za hypotézu. V tomto smyslu se za hypotézu považuje tvrzení (Gries, 2009, s. 11), které

a) se týká více než jednoho jevu či případu;
b) má alespoň implicitně strukturu podmínkového souvětí, tj. „*jestliže…, pak…*“, případně „*čím…, tím…*“ (např. čím je slovo frekventovanější, tím je kratší);
c) je falzifikovatelné (tj. vyvratitelné) prostřednictvím experimentu, který dovoluje rozhodnout, zda predikce formulovaná prostřednictvím hypotézy je vyvrácena, či ne.

Nahlédneme-li do praxe experimentálních věd, zjistíme, že se tento standard dodržuje, tj. pokud se mluví o hypotézách, pak v tomto smyslu – např. je minimální pravděpodobnost, že se v časopise z oboru experimentálních věd bude v odborné studii používat výraz „hypotéza“ ve smyslu domněnky či tvrzení, které nelze falzifikovat.

Naproti tomu v lingvistice se s tímto pojmem zachází mnohem volněji, srov.

- „V hláskosloví ani v jiných rovinách vodňanského herbáře nejsou prokazatelné další nářeční jevy z oblasti, kde nedošlo ke vzniku vibranty *ř* (východomoravské území), dáváme proto přednost hypotéze, že se jedná o nedbalý zápis“ (Černá, 2005, s. 76);
- „V nich se překlad Františka Vrby jeví jako silně zatížený mužským genderovým úhlem pohledu a estetikou vnímání; potvrzuje se tak původní hypotéza, že se spíše „staví na stranu“ mužského hrdiny, resp. autorského tvůrce a erotické líčení prezentuje spíše z jeho perspektivy“ (Širokovská, 2004, s. 23);
- „Proč není samo slovo *plémě* ve staročeských textech doloženo v očekávaném významu, o tom lze vznášet různé hypotézy.“ (Šimandl, 2007, s. 238);
- „Hypotéza 2.1: Co-text je věrným zrcadlem (situačního) kontextu v tom smyslu, že všechny pro danou komunikační situaci relevantní kontextové vlastnosti jsou co-textem explicitně reflektovány, a mají tedy nějaký textový korelát. […] Hypotéza 2.2: (Textový) kontext věrně reflektuje všechny vlastnosti jazykových jevů relevantní pro jejich užití.“ (Cvrček, 2013, s. 24)[6];
- „Vycházeje z toho, že teorie valence i přes zjevná slabá místa představuje dobrý konstrukt lingvistické teorie, pokusím se nyní představit hypotézu *modifikované valenční teorie* (MVT) a formulovat základní principy této teorie.“ (Karlík, 2001, s. 171n.).

Výrazem „hypotéza“ jsou označovány domněnky, předpoklady, klasifikace, empiricky netestovatelné výroky atd. V žádném případě není mým úmyslem používání výrazu „hypotéza“ v těchto významech označit jako špatné, nevhodné

[6] V tomto případě autor explicitně zmiňuje, že se jedná o empiricky netestovatelné hypotézy a chápe je jako premisy.

či nevědecké (vše záleží na volbě kritérií „vědeckosti“), ani nepředpokládám, že by výše uvedení autoři neznali význam tohoto slova ve smyslu testovatelného tvrzení – spíše je vidím jako projev určitého specifika lingvistického diskurzu, což se dá chápat jako důsledek toho, že se lingvistika neetablovala jako experimentální věda. Z hlediska lingvistické praxe (zejména té části, která k lingvistice přistupuje jako k experimentální vědě) je však nezbytně nutné, aby bylo vždy jasné, v jakém smyslu je tento výraz použit.

5 Za hranice popisu

Ilustrujme „cestu“ od popisu/klasifikace k teoreticky odůvodněným hypotézám na příkladu analýzy slovesné valence. Slovesná valence se tradičně definuje jako vlastnost slovesa „vázat na sebe určitý počet syntaktických pozic, determinovaný počtem sémantických aktantů, v jistých morfologických formách obsaditelných primárně jistými výrazy“ (Grepl – Karlík, 1998, s. 45). Dále se vychází z předpokladu, že „valenční vlastnosti sloves (i dalších slovních druhů) jsou velmi rozmanité. Nelze je tedy odvodit obecnými pravidly, je třeba je popsat pro jednotlivé položky“ (Lopatková et al., 2008, s. 8). Výsledkem komplexního popisu valence jsou pak valenční slovníky (Svozilová et al., 1997, 2005; Lopatková et al., 2008; Hlaváčková, 2008) zachycující valenční charakteristiky jednotlivých lexikálních jednotek prostřednictvím tzv. valenčních rámců, které vyjadřují počet, formu a někdy i význam valenčních doplnění. Kvantifikovanou podobu valenčního slovníku reprezentuje *PDT-Vallex* (Urešová, 2011), v němž je u každého valenčního rámce přidána informace o frekvenci daného rámce v *Pražském závislostním korpusu*.

Pokud se podíváme na fungování valence z perspektivy experimentálního přístupu, je třeba nejprve najít teoretické zdůvodnění vztahu valence a jiné vlastnosti jazyka, přičemž musí být splněna podmínka empirické testovatelnosti. Zdá se například rozumné předpokládat, že výraz, který se vyskytuje s vysokou frekvencí, se objevuje v mnoha různých kontextech, tudíž roste pravděpodobnost, že některé z těchto kontextů se pro daný výraz stanou typickými, v některých případech dokonce i obligatorními. Jedním z důsledků tohoto procesu by měl být nárůst počtu valenčních rámců výrazu, které nejsou ničím jiným než obligatorními kontexty. Na základě této teoretické úvahy lze vyslovit empiricky testovatelnou hypotézu: *čím je sloveso frekventovanější, tím větší počet valenčních rámců má*. Protože víme, že frekventovaná slova jsou v průměru kratší než slova méně frekventovaná, odvodíme další hypotézu: *čím je sloveso kratší, tím má více valenčních rámců.* A dále, různé kontexty mají bezpochyby vliv na sémantické změny daného výrazu, což nás dovede k další hypotéze: *čím má sloveso větší počet valenčních rámců, tím je větší jeho polysémie*. S nárůstem různých kontextů a polysémie, tj. i valenčních rámců, roste zase pravděpodobnost, že se v růz-

ných kontextech vyskytnou různá slova vyjadřující podobné významy, což nás vede k hypotéze: *čím má sloveso větší počet valenčních rámců, tím větší počet synonym má*. A takto bychom mohli pokračovat dále, srov. Strauss et al. (2008), Čech – Altmann (2011), Čech – Mačutek (2010), Čech et al. (2010), Gao et al. (2014). Tímto relativně jednoduchým způsobem se dostáváme za hranice popisu, navíc se nám daří vysvětlit fungovaní nejen jednoho jevu (v našem případě valence), ale modelovat také vztahy mezi jazykovými jevy, o kterých se zpravidla uvažuje izolovaně. Dále se ukazuje, že experimentální přístup k valenci, jak je zde v krátkosti představen, narušuje výše zmíněnou představu, že „valenční vlastnosti sloves (i další slovních druhů) jsou velmi rozmanité. Nelze je tedy odvodit obecnými pravidly" (Lopatková et al., 2008, s. 8).[7] A co je snad nejdůležitější, všechny zmíněné hypotézy se dají vztáhnout k obecným mechanismům ovlivňujícím naše jazykové chování a vysvětlit na teoretické rovině, např. prostřednictvím tzv. synergetické lingvistiky (Köhler, 2005).

Na příkladu experimentálně založené analýzy slovesné valence lze nejen ilustrovat, jak překonat její deskriptivní charakter směrem k explanaci, ale také to, že z hlediska analýz tohoto druhu nepředstavuje pojem reprezentativnosti jazykových korpusů (srov. část 3) žádný problém. Pokud totiž testujeme hypotézy, které reprezentují určité mechanismy (často odvozené z nějakých obecných principů, jako je např. Zipfův *princip nejmenšího úsilí* – srov. Zipf (1949) nebo jeho rozpracování v Köhler (2005)), předpokládáme, že ony mechanismy řídí jazykové chování *jednotlivých* uživatelů jazyka (aniž by si toho byli tito uživatelé vědomi – např. mechanismus řídící vztah frekvence a délky slova). Z toho plyne, že vezmeme-li jakýkoliv *jednotlivý* text (ať mluvený, či psaný), měly by se předpokládané mechanismy v textu projevit.[8] Takový mechanismus, pokud je patřičně ověřen prostřednictvím testů hypotéz, má status zákona. Jestliže ale analyzujeme vzorek jazyka sestavený z různých textů, zpravidla dospíváme k horším výsledkům (míru kvality výsledků lze určit na základě míry shody matematického modelu s daty), než když analyzujeme jednotlivé texty. To je snadno pochopitelné, když si uvědomíme, že se předpokládaný mechanismus projevuje u každého jednotlivce trochu jinak a že na něj má vliv např. volba tématu, vliv žánru atp. A když jsou pak všechny tyto vlivy „smíchány" v jednom vzorku, do jisté míry „pokřivují" pohled na fungování daného mechanismu – např. Grzybek a Stadlober (2007) odhalili, že v případě vytvoření korpusu, ve kterém jsou smíchány texty různých žánrů, přestává platit tzv. Arensův zákon (zákon vyjadřující vztah

[7] To samozřejmě neznamená, že by bylo možné prostřednictvím kvantitativních analýz vytvořit valenční slovník; existují však valenční vlastnosti sloves, které odvodit lze.

[8] Pokud se neprojeví, hledáme nejprve příčiny tohoto jevu a zpravidla docházíme k odhalení tzv. hraničních podmínek dané hypotézy – např. dadaistické texty či projevy lidí s Wernickovou afázií ze zřejmých důvodů mnohé mechanismy „nerespektují" – nebo hypotézu odmítáme a hledáme lepší vysvětlení.

mezi délkou slova a délkou věty). To samozřejmě neznamená, že neexistují jazykové vlastnosti, které je možné testovat na mixu různých textů. Celou problematiku[9] přehledně shrnují Wimmer et al. (2003, s. 21): „Ak chceme overovať nejaký zákon, tak nesmieme miešať texty, lebo v každom texte sú iné tzv. počiatočné podmienky. Dokonca je niekdy potrebné analyzovať oddelene aj jednotlivé kapitoly románu alebo vety symofónie. Dá sa napríklad ukázať, že zákon pre rozdelenie dĺžky slova je ľahko overiteľný napr. na jednotlivých Goetheho listoch. Čím viac listov však zlúčime do spoločného výberu, tým menej vhodným sa stáva daný model, lebo v každom liste má zákon iné parametre. V takových prípadoch sa odporúča prinajmenšom miešanie, kombiovanie modelov (porov. Altmann 1992), ale ešte vhodnejšia je separácia homogenních častí. Tak napríklad tzv. frekvenčný slovník jedného celého jazyka (žánru a pod.) je vhodný len na veľmi obmedzené teoretické účely, lebo je konštruovaný zo zmesi textov".

Shrnuto, v experimentálním přístupu založeném na statistickém testování hypotéz neznamená rozsáhlost dat, na nichž je hypotéza testována, žádnou „automatickou" výhodu (ve smyslu čím více dat, tím lépe), srov. také Rietvield et al. (2004).

6 Závěr

Korpusová lingvistika se podle mého názoru nachází, i vzhledem k historickým okolnostem souvisejícím s jejím vznikem a vývojem, na hranici dvou paradigmatických přístupů: svou orientací na popis (ve smyslu klasifikace doplněné interpretací) je pevně svázána se strukturalismem, na druhou stranu používání kvantifikace a důsledky z ní plynoucí před ní otevírají problémy, které nejsou, zdá se, v rámci strukturalistického paradigmatu řešitelné a které ji dostávají do úzkého kontaktu s paradigmatem jiným – paradigmatem, v němž kvantifikace hraje ústřední roli z důvodů teoretických a v němž má centrální postavení experiment založený na testování hypotéz. Korpusová lingvistika tak stojí na hranici mezi „popisem s čísly" a perspektivami, které jí nabízí přijetí experimentální metodologie. Jistě bude zajímavé sledovat, jakým směrem se v příštích letech bude její vývoj ubírat.

LITERATURA

ALTMANN, G. (1988): *Wiederholungen in Texten*. Bochum: Brockmeyer.

ALTMANN, G. (1992): Das Problem der Datenhomogenität. *Glottometrika*, 13, s. 287–298.

ALTMANN, G. (2012): Certain Differences between Qualitative and Quantitative Linguistics. *Czech and Slovak Linguistic Review*, 2, s. 6–15.

BIBER, D. – JOHANSSON, S. – LEECH, G. – CONRAD, S. – FINEGAN, E. (1999): *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education ESL.

CVRČEK, V. (2013): *Kvantitativní analýza kontextu*. Praha: Nakladatelství Lidové noviny.

[9] V obecnější rovině jde o problematiku tzv. homogenity a heterogenity jazykových dat, srov. Altmann (1988, 1992), Strauss et al. (2007), Grzybek (2013).

CVRČEK, V. – KODÝTEK, V. (2013): Ke klasifikaci morfologických variant. *Slovo a slovesnost*, 74, s. 139–145.
CVRČEK, V. – KOVÁŘÍKOVÁ, D. (2011): Možnosti a meze korpusové lingvistiky. *Naše řeč*, 94, s. 113–133.
ČECH, R. (2005a): Komunikace versus systém, nebo komunikace versus model? *Slovo a slovesnost*, 66, s. 176–179.
ČECH, R. (2005b): Limity (nejen jazykovědného) strukturalismu. *Slovo a slovesnost*, 66, s. 19–31.
ČECH, R. (2012): Několik teoreticko-metodologických poznámek k Mluvnici současné češtiny. *Slovo a slovesnost*, 73, s. 208–216.
ČECH, R. – ALTMANN, G. (2011): *Problems in Quantitative Linguistics 3*. Lüdenscheid: RAM-Verlag.
ČECH, R. – MAČUTEK, J. (2010): On the quantitative analysis of verb valency in Czech. In: P. Grzybek – E. Kelih – J. Mačutek (eds.), *Text and Language. Structures – Functions – Interrelations. Quantitative Perspectives*. Wien: Praesens Verlag, s. 21–29.
ČECH, R. – PAJAS, P. – MAČUTEK, J. (2010): Full Valency. Verb Valency without Distinguishing Complements and Adjuncts. *Journal of Quantitative Linguistics*, 17, s. 291–302.
ČERNÁ, A. M. (2005): Herbář z kodexu vodňanského a jeho ortografické zvláštnosti. *Naše řeč*, 88, s. 71–81.
Český národní korpus – SYN2010 (2010) [online]. Praha: Ústav Českého národního korpusu FF UK. <http://www.korpus.cz>.
DANEŠ, F. – GREPL, M. – HLAVSA, Z. (1987): *Mluvnice češtiny 3. Skladba*. Praha: Academia.
FEYERABEND, P. K. (2001): *Rozprava proti metodě*. Praha: Aurora.
GAO, S. – ZHANG, H. – LIU, H. (2014): Synergetic Properties of Chinese Verb Valency. *Journal of Quantitative Linguistics*, 21, s. 1–21.
GREPL, M. – KARLÍK, P. (1998): *Skladba češtiny*. Olomouc: Votobia.
GRIES, S. (2009): *Statistics for Linguistics with R: A Practical Introduction*. Berlin: Mouton de Gruyter.
GRZYBEK, P. (2013): Homogeneity and heterogeneity within language(s) and text(s): Theory and practice of word length modeling. In: R. Köhler – G. Altmann (eds.), *Issues in Quantitative Linguistics 3*. Lüdenscheid: RAM-Verlag, s. 66–99.
GRZYBEK, P. – STADLOBER, E. (2007): Do we have problems with Arens' law? A new look at the sentence-word relation In: P. Grzybek – R. Köhler (eds.), *Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday*. Berlin – New York: Mouton de Gruyter, s. 205–217.
HLAVÁČKOVÁ, D. (2008): *Databáze slovesných valenčních rámců VerbaLex* [online]. Disertační práce. Brno: Masarykova univerzita. Cit. 2014-04-24. <http://theses.cz/id/f4zej0?info=1;isshlret=VerbaLex;;zpet=/vyhledavani/?search=verbalex>.
CHROMÝ, J. (2014): Korpus a reprezentativnost. *Naše řeč*, 97, s. 185–193.
KARLÍK, P. (2001): Hypotéza modifikované valenční teorie. *Slovo a slovesnost*, 61, s. 170–189.
KÖHLER, R. (2005): Synergetic linguistics. In: R. Köhler – G. Altmann – R. G. Piotrowski (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*. Berlin – New York: de Gruyter, s. 760–774.
KRÁLÍK, J. (2013): Srovnávání nesrovnatelného. *Korpus – gramatika – axiologie*, 4, s. 48–52.
LEECH, G. (2007): New resources, or just better old ones? In: M. Hundt – N. Nesselhauf – C. Biewer (eds.), *Corpus Linguistics and the Web*. Amsterdam: Rodopi, s. 134–149.

LOPATKOVÁ, M. – ŽABOKRTSKÝ, Z. – KETTNEROVÁ, V. (2008): *Valenční slovník českých sloves*. Praha: Karolinum.
McENERY, T. – HARDIE, A. (2012): *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
NOVÁK, P. (1964): Two types of formulae in quantitative linguistics. *The Prague Bulletin of Mathematical Linguistics*, 2, s. 11–14.
PEREGRIN, J. (2005): *Kapitoly z analytické filozofie*. Praha: Filosofia.
QUINE, W. V. O. (1991): *Hledání pravdy*. Praha: Herrmann & synové.
RIETVIELD, T. – VAN HOUT, R. – ERNESTUS, M. (2004): Pitfalls in Corpus Research. *Computers and the Humanities*, 38, s. 343–362.
RORTY, R. (2012): *Filosofie a zrcadlo přírody*. Praha: Academia.
RUSSELL, B. (1975): *Zkoumání o smyslu pravdivosti*. Praha: Academia.
STRAUSS, U. – FAN, F. – ALTMANN, G. (2008): *Problems in Quantitative Linguistics 1*. Lüdenscheid: RAM-Verlag.
STRAUSS, U. – GRZYBEK, P. – ALTMANN, G. (2006): Word Length and Word Frequency. In: P. Grzybek (ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Dordrecht, NL: Springer, s. 277–295.
SVARTVIK, J. (2007): Corpus linguistics 25+ years on. In: R. Facchinetti (ed.), *Corpus linguistics 25 years on*. Amsterdam – New York: Rodopi, s. 11–26.
SVOZILOVÁ, N. – PROUZOVÁ, H. – JIRSOVÁ, A. (1997): *Slovesa pro praxi*. Praha: Academia.
SVOZILOVÁ, N. – PROUZOVÁ, H. – JIRSOVÁ, A. (2005): *Slovník slovesných, substantivních a adjektivních vazeb a spojení*. Praha: Academia.
ŠIMANDL, J. (2007): Proměny pohledů na slovo plemeno (a trochu i na reprezentativní korpusy). *Naše řeč*, 90, s. 237–246.
ŠIROKOVSKÁ, S. (2004): Dvojí překlad knihy D. H. Lawrence Milenec Lady Chatterleyové z perspektivy genderu. *Naše řeč*, 87, s. 14–24.
ŠULC, M. (2001): Tematická reprezentativnost korpusů. *Slovo a slovesnost*, 62, s. 53–61.
TĚŠITELOVÁ, M. (1980): *Využití statistických metod v gramatice*. Praha: Academia.
TĚŠITELOVÁ, M. et al. (1985): *Kvantitativní charakteristiky současné češtiny*. Praha: Academia.
UREŠOVÁ, Z. (2011): *Valenční slovník Pražského závislostního korpusu* (*PDT-Vallex*). Praha: Ústav formální a aplikované lingvistiky MFF UK.
WIMMER, G. – ALTMANN, G. – HŘEBÍČEK, L. – ONDREJOVIČ, S. – WIMMEROVÁ, S. (2003): *Úvod do analýzy textov*. Bratislava: Veda.
WITTGENSTEIN, L. (1993a): *Tractatus logico-philosophicus*. Praha: Oikúmené.
WITTGENSTEIN, L. (1993b): *Filosofická zkoumání*. Praha: Filosofický ústav AV ČR.
ZIPF, G. K. (1935): *The psycho-biology of language. An Introduction to Dynamic Philology*. Boston: Houghton-Mifflin – Cambridge: MIT Press (2nd edition, 1968).
ZIPF, G. K. (1949): *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology*. Cambridge: Addison-Wesley.

Katedra českého jazyka FF OU
Reální 5, 701 03 Ostrava
radek.cech@osu.cz