Proč (někdy) nemíchat texty aneb Text jako možná výchozí jednotka lingvistické analýzy[1]

Radek ČECH | Katedra českého jazyka FF OU
Pavel KOSEK | Ústav českého jazyka FF MU
Ján MAČUTEK | Katedra aplikovanej matematiky a štatistiky, Fakulta matematiky, fyziky a informatiky Univerzity Komenského v Bratislavě a Matematický ústav Slovenskej akadémie vied
Olga NAVRÁTILOVÁ | Ústav českého jazyka FF MU

Why not to mix texts (sometimes): the text as a possible default unit of linguistic analysis

The paper deals with two important questions in linguistic research: 1) What do we actually model when we model language usage? and 2) What is an appropriate sample or 'text unit' for the analysis of language usage? In the beginning, we critically discuss several approaches to the analysis of language behaviour. Then, we introduce the most important characteristics of both Zipf's linguistic theory and synergetic linguistics. We focus in particular on the aspects of these theories which are connected to the above-mentioned questions. Specifically, we emphasize that one of the fundamental features of these theories is the assumption that there are linguistic laws which govern human language behaviour and which can be best detected by observing the language behaviour of an individual (in a particular context). As a consequence, if the goal of the research is to examine laws of this kind, the individual text is used as a basic unit for the analysis. The mixing of texts can, in some cases, lead to the "concealing" of the laws, as is presented in an example. We also offer another example which shows how characteristics of the same law (in this case, the Menzerath-Altmann law) differ in different texts. Finally, we emphasize that using individual texts in linguistic research is but one possible approach to analysis, i.e. we do not attempt to make it a linguistic research dogma.

Key words: analysis, language behaviour, model, sample, text

Klíčová slova: analýza, jazykové chování, model, text vzorek

1 Úvod

V současné lingvistice můžeme nalézt několik zásadně odlišných odpovědí na to, co má být předmětem lingvistického výzkumu. Například strukturalistická lingvistika si klade za cíl modelovat abstraktní jazykový systém ve smyslu langue, generativní lingvistika se pokouší vysvětlit fungování tzv. jazykové kompetence, tj. mentální

[1] Příspěvek vznikl v rámci řešení grantového projektu *Vývoj českých pronominálních enklitik* (GA17-02545S), 2017–2019.

schopnosti člověka vytvářet správné jazykové struktury a rozumět jim, lingvistické směry zaměřené na analýzu toho, jak se jazyk používá, zase vytvářejí modely jazykového chování. Každý z přístupů má bezpochyby své výhody a limity, které byly a jsou velmi intenzivně diskutovány. V tomto článku se pokusíme představit jeden z aspektů výzkumu založeného na analýze jazykového chování. Chceme ukázat, že určitá teoretická východiska, která mimo jiné definují předmět výzkumu, vedou k tomu, že někdy volíme text za výchozí jednotku jazykového bádání. Navíc budeme na konkrétních příkladech ilustrovat, jaké důsledky pro interpretaci výsledků výzkumu může mít míchání textů. Rádi bychom hned v úvodu zdůraznili, že jsme si dobře vědomi toho, že volba textu jako výchozí jednotky není bez problémů. Ale „problematické" jsou i všechny ostatní přístupy a v principu zřejmě neexistuje ani ideální teorie, ani ideální metoda.

V první části článku se zaměříme na kritické zhodnocení toho, jak je možné interpretovat modely řečového chování, a stručně shrneme základní principy Zipfovy jazykové teorie a synergetické lingvistiky, přičemž se zaměříme na ty aspekty obou přístupů, které vedou k volbě textu jako základní jednotky. V další části poukážeme na problematiku míchání textů a možných důsledků tohoto postupu.

2 Interpretace modelů jazykového chování

Vyjděme z jednoduchého předpokladu, že texty (ať mluvené, či psané) jsou projevem pozorovatelného jazykového chování a představují empirický základ pro vytváření modelů tohoto chování.[2] Co tyto modely reprezentují? Na první pohled je odpověď triviální – reprezentují výše zmíněné jazykové chování. Máme tedy text či sumu textů, například v podobě jazykového korpusu, a model pak představuje určité zobecnění vlastností daného textu či korpusu. V nejjednodušším případě se jedná o klasifikačně založené popisy vybraných jazykových jednotek. Klasifikace ovšem většinou bývá jen prvním krokem. Následuje induktivní vyvození pravidelností či pravidel pozorovaného chování. Na základě těchto pravidelností je pak možné formulovat predikce, tj. očekávání, že stejné pravidelnosti bude možné pozorovat i v dalších textech či korpusech. Všimněme si, že tento krok stojí na předpokladu, že pozorovatelné jazykové chování mluvčích je projevem nějakých obecnějších principů a mechanismů. Jde-li nám o pochopení těchto principů (tj. nechceme-li zůstat u pouhé klasifikace a snažíme-li se *vysvětlit* vlastnosti jazykového chování), model již ve skutečnosti nereprezentuje jazykové chování per se, nýbrž reprezentuje vlastnosti něčeho, co de facto stojí za konkrétními řečovými projevy. Z této perspektivy se již otázka „co tyto modely reprezentují?" jeví v poněkud jiném světle a zdá se, že ade-

[2] Ostatními přístupy, zejména těmi, které jsou založeny na introspekci, či „laboratorními" přístupy analyzujícími chování mluvčích při reakci na podněty nejrůznějšího typu, se zde zabývat nebudeme.

kvátní odpověď na ni nabízí nějaká forma dualismu, tak jak ji známe třeba v podobě langue-parolové dichotomie jazykovědného strukturalismu – předmětem modelování je pak například abstraktní jazykový systém ve smyslu langue.[3] Nechme nyní stranou teoretické a metodologické problémy související s dichotomickým přístupem (srov. Hopper, 1987; Kořenský, 1987; Komárek, 1999; Newmeyer, 2003; Laury – Ono, 2005; Čech, 2005; 2007; 2017; Beneš, 2015) a sledujme, jaké důsledky má toto teoretické východisko při volbě jazykového materiálu. Je-li langue chápán jako abstraktní systém, který nějakým způsobem sdílejí mluvčí daného jazyka, a je-li jeho analýza založena na zkoumání jazykového chování, pak je ideálem mít k dispozici všechny jazykové projevy a z nich odvodit vlastnosti langue. Samozřejmě, mít takový soubor jazykových projevů je ideál prakticky nedosažitelný. V každém případě pak lze ale tento soubor chápat jako „populaci" a pokusit se vytvořit adekvátní „vzorek" této populace. Zde ovšem narážíme na další často diskutovaný problém faktické nereprezentativnosti vzorků v lingvistice, za všechny srov. Chromý (2014) a literaturu zde uvedenou. Zde je třeba zdůraznit, že velikost vzorku – ve smyslu čím větší korpus vytvoříme, tím více se přiblížíme onomu ideálnímu stavu – problém s nereprezentativností neřeší. Bez ohledu na všechny problémy související s dichotomickým přístupem je ale asi jasné, že adekvátní jednotkou tohoto typu analýzy bude *soubor* textů (tj. jazykový korpus), nikoliv text jediný.

Zcela specifický přístup v rámci langue-parolového přístupu představuje pojetí Cvrčkovo (2014). Ten chápe langue nikoliv jako abstraktní systém, ale jako sumu všech jazykových promluv, srov. jeho charakteristiku kontextu jako obecného fenoménu vztahujícího se „ke všem jazykovým jevům ve všech jejich realizacích, což vede ke kvantitativnímu zkoumání langue jako celku" (Cvrček, 2014, s. 11–12). Jedná se o velmi specifické pojetí langue, které se zásadně liší od tradičního přístupu k tomuto pojmu v lingvistice. Co se však týká vymezení adekvátní jednotky jazykové analýzy, i zde bude výchozím materiálem soubor textů.

Nyní se pokusme na problematiku modelu jazykového chování a jeho interpretace podívat z jiné perspektivy. Začneme nikoliv od samotných projevů tohoto chování, tj. jazykových promluv (ať psaných, či mluvených), z nichž jsou *induktivně* odvozována pravidla a pravidelnosti jazykového systému, ale vezměme jako výchozí bod obecný princip lidského chování, z něhož budeme *deduktivně*[4] vyvozovat vlastnosti jazykového systému. To je přístup, který byl v lingvistice představen Zipfem (1935; 1949). Oním obecným principem, který má zásadní vliv na lidské chování, včetně jeho verbální složky, je podle Zipfa tzv. princip nejmenšího úsilí.

[3] Langue-parolová dichotomie je běžně přijímána jako východisko lingvistického výzkumu u korpusových lingvistů, za všechny srov. Čermák (2017).

[4] Rádi bychom zdůraznili, že nám nejde o to preferovat dedukci před indukcí, jak to například činí Popper (1997), ale jde nám zde o nastínění určité badatelské perspektivy.

Ten v podstatě znamená následující: máme-li nějaký cíl, hledáme co nejúčinnější způsob, jak ho dosáhnout, přičemž bereme v potaz relativně široký kontext našeho konání. Jinými slovy, zdroje jsou omezené, proto je třeba s ohledem na požadovaný cíl volit takovou strategii, která je pro nás výhodná (a to i z dlouhodobého hlediska), kvůli tomu, abychom omezené zdroje co nejefektivněji využili. Například, existuje-li velká pravděpodobnost, že budeme nuceni v příštích letech psát na klávesnici každý den několik hodin (a že rychlost psaní bude mít navíc třeba i vliv na výši naší mzdy), tak je pro nás pravděpodobně výhodné vynaložit relativně velké úsilí na osvojení si psaní všemi deseti prsty, protože ve výsledku (tj. s ohledem na vynaložené úsilí při psaní v následujících letech a s ohledem na „zisk" plynoucí z této činnosti) to bude znamenat menší úsilí, než kdybychom na klávesnici psali dvěma prsty. Zde je třeba zdůraznit, že se jedná o stochastický princip a že při volbě dané strategie se jedná o odhad, který vlivem nejrůznějších faktorů nemusí vyjít. V případě psaní na klávesnici se může třeba ukázat, že díky rozvoji technologií psát na klávesnici vůbec nebude potřeba nebo přijdeme o ruku nebo najdeme úplně jiné zaměstnání atp., a že náš odhad byl tudíž špatný. Podle Zipfa má tento princip rozhodující vliv jak na naše chování, tak na vlastnosti systémů, které lze vnímat jako výsledek našeho chování – třeba na podobu lidského jazyka.

Na základě tohoto principu je možné vyvozovat vlastnosti zkoumaného systému, ideálně ve formě empiricky testovatelných hypotéz. Za všechny uveďme vztah mezi frekvencí slova a jeho délkou (čím je slovo frekventovanější, tím je v průměru kratší), mezi frekvencí a polysémií (čím je slovo frekventovanější, tím má více významů), mezi slovesnou valencí – respektive tzv. plnou valencí (viz Čech et al., 2010) – a komplexitou syntaktického stromu (čím více slov je bezprostředně závislých na slovese, tím kratší jsou v průměru syntaktické fráze, které jej rozvíjejí) atd. Systém v tomto pojetí je výsledkem jednotlivých interakcí mezi mluvčími, přičemž tyto interakce na jedné straně udržují stabilitu systému (aby bylo možné se vůbec domluvit), na straně druhé jej neustále mění (například vlivem kontextu) – proto Zipf, mimo jiné, ve 30. letech 20. století označoval svůj přístup termínem dynamická filologie.

Princip nejmenšího úsilí sice představuje výchozí strategii řečového chování lidí, ale při jeho realizaci na něj působí okolnosti nejrůznějšího typu a ty mají zásadní vliv na to, jak se projeví. Řečeno obecněji, princip je sice stále stejný, ale jeho parametry se díky kontextu mění. Například, mluvíme-li k malým dětem o tom, co budeme dělat o víkendu, na jedné straně, nebo diskutujeme-li na třídní schůzce s dospělými lidmi o problematice kvality jídla ve školní jídelně, na straně druhé, naše řečové chování sice vychází ze stejných principů, ale kontext způsobuje určité změny. Dále, jinak se v těchto kontextech zřejmě bude projevovat desetiletý mluvčí v porovnání s mluvčím padesátiletým. Testujeme-li pak platnost jednotlivých hypotéz (viz výše), zdá se být teoreticky adekvátnější použít takový jazykový materiál, který je co nejhomogennější, tj. promluvu daného mluvčího v daném kontextu. V určitých

případech totiž může dojít k tomu, že rozdílnost kontextu (či lépe řečeno tzv. hraničních podmínek) může způsobit, že pokud texty smícháme, testovanou hypotézu zamítneme právě z důvodu silného vlivu různých kontextů – parametry jsou pak tak rozdílné, že nedovolí, aby byl „vidět" mechanismus, na jehož základě byla hypotéza formulována (konkrétní příklady viz v části 3).

Samozřejmě že požadavek na homogenitu vzorku je určitý ideál, zejména v případě psaných textů – málokdy máme k dispozici takový psaný text, který byl vytvořen „jedním tahem" bez toho, aby byl následně korigován, případně redigován někým jiným.[5] Navíc, míchání textů (tj. de facto působení jiných kontextů) někdy vede k tomu, že předpokládaný mechanismus se projeví ještě silněji než v případě analýzy jednotlivých textů (či promluv). Ze Zipfova přístupu a důsledků z něj plynoucích podle našeho názoru nevyplývá jednoznačný závěr, na jehož základě bychom měli „dogmaticky" akceptovat buď analýzy založené pouze na množině textů, nebo pouze na textech (promluvách) jednotlivých. Zipfův přístup nás ale vede k určité obezřetnosti, zejména v případě práce se souborem textů. Je při tom důležité, že tato „obezřetnost" má poměrně silné teoretické zakotvení. Navíc se může stát inspirací k poněkud odlišnému pohledu na možnosti analýzy jazykového chování obecně.

Přestože se Zipf (1935; 1949) ve svých knihách zabývá různými rovinami a vlastnostmi jazykového systému, jeho přístup má více méně heuristický charakter. Jako pokus o systematické rozpracování Zipfova pojetí lze chápat koncept synergetické lingvistiky (Köhler, 1986; 2005), který ze Zipfa vychází a který – inspirován myšlenkami samoorganizace a samoregulace systémů (Eigen, 1971; Prigogine – Stengersová, 2001), a zejména synergetickou teorií (Haken, 1983) – představuje pokus o vytvoření obecné lingvistické teorie.[6] Jedním z teoretických východisek synergetické lingvistiky je předpoklad, že podoba jazykového systému je výsledkem samoregulačních mechanismů vycházejících z interakce mezi tzv. požadavky mluvčího a posluchače. V tomto pojetí je jazykový systém v principu systémem dynamickým, který se neustále mění a vyvíjí, ale právě díky samoregulačním a samoorganizačním mechanismům je v určitém rovnovážném stavu. Tento stav je de facto „produkt" adaptace vůči okolí. Jinak řečeno, mluvčí (který má ovšem jindy roli posluchače) využívá aktuální stav systému k dosažení svého komunikačního cíle, přičemž má možnost jej do jisté míry ovlivňovat a měnit, zejména pokud tato změna vede k naplnění jeho komunikačních cílů.

---

[5] Jedním z mála textů, o kterém máme doloženo, že byl napsán bez následných úprav, je Škvoreckého povídka Bassaxofon, srov.: „V polosnu tří extatických dnů jsem napsal Bassaxofon. Když byl hotový, nemusel jsem změnit jedinou řádku" (Škvorecký, J., Příběh neúspěšného saxofonisty. Praha: 2002).

[6] Velmi srozumitelně hlavní principy synergetiky a synergetické lingvistiky představila v českém prostředí Uhlířová (1995).

Pokud jde o vymezení jednotky, která se používá při jazykové analýze, převažuje v synergetice pohled, který důrazně preferuje práci s jednotlivými texty. Důvody tohoto přístupu velmi přehledně shrnuje Uhlířová (1995, s. 10):

> Text jako předmět zkoumání synergetické lingvistiky je uvažován, jak vyplývá z výše řečeného, z hlediska procesuálního: Předpokládá se, že vzniká – jakožto nová kvalita – za určitých výchozích podmínek a určitými strategiemi, které mohou působit buď souhlasně, anebo proti sobě. Výchozí podmínky zůstávají, alespoň v bezpříznakových případech, v průběhu procesu generování textu rozumně stabilní (alespoň to o nich lze předpokládat), takže text lze považovat za synergetický systém. Má se za to, že vnitřní stabilitu textu, resp. její míru, v krajním případě pak nedostatek stability, lze odhalit za předpokladu, že budeme zkoumat text jako celek. To je kardinální zásada synergetické lingvistiky; synergetická lingvistika nepracuje s náhodnými, systematickými ani jinými výběry z textů, protože má za to, že při jakémkoli výběru existuje riziko, že se mohou některé vlastnosti textu „ztratit". Jen text jako celek, jako individuální opus s celistvou strukturou, může představovat organizovaný, rovnovážný systém. Tento systém, resp. jeho rovnovážnost, je hledána v jazykových prostředcích, jimiž je vytvářen.

Shrnuto, pokud se rozhodneme modelovat řečové chování mluvčích, měli bychom explicitně stanovit, v jakém teoretickém rámci pracujeme. V současné lingvistice zaměřené na analýzu toho, jak se jazyk používá, totiž neexistuje všeobecně přijímaný konsensus ohledně toho, co je vlastně modelováno.

3 Praktické příklady

V této části budeme na dvou příkladech ilustrovat, jak rozdílné mohou být výsledky, které jsou založeny jednak na analýze skupiny textů, jednak na analýze textů jednotlivých. V jedné z našich analýz (Čech et al., 2019) jsme sledovali vztah mezi tzv. plnou valencí predikátu – tj. počtem výrazů, které jsou bezprostředně závislé na predikátu (viz Čech et al., 2010) – a pozicí příklonky *sě* ve staré češtině.[7] Konkrétně jsme předpokládali, že čím je větší plná valence predikátu, tím je menší pravděpodobnost, že se příklonka *sě* realizuje v tzv. postiniciální pozici klauze (tj. za prvním syntaktickým elementem).[8] Tento předpoklad jsme ověřovali na jazykovém materiálu skládajícím se z osmi biblických knih Bible olomoucké (Genesis, Job, Izaiáš, Sírachovec, Matouš, Lukáš, Skutky, Zjevení). Výsledky založené na analýze korpusu složeného z těchto osmi knih (tabulka 1 a obrázek 1) vedly k zamítnutí predikce – namísto předpokládané negativní korelace jsme zjistili korelaci pozitivní, byť statisticky nevýznamnou (na základě Kendallova korelačního koeficientu $t = 0{,}79$, p-hodnota = 0,49).

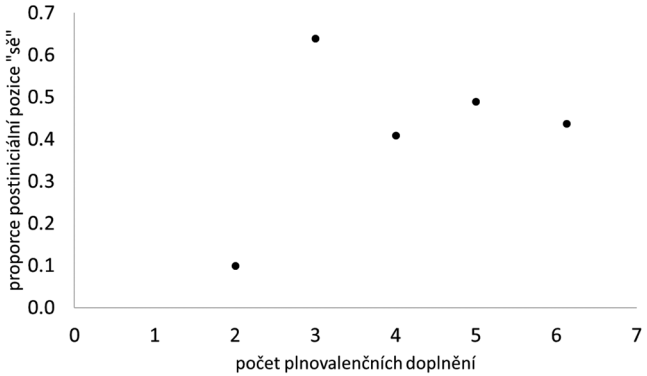
---

[7] Až do začátku 20. stol. se čeština vyznačovala konkurencí mezi postiniciální pozicí enklitika (*V zahradě **se** starý strom rázem **skácel***) a kontaktní pozicí enklitika v bezprostřední blízkosti svého syntakticky/morfologicky nadřazeného výrazu (*Starý strom **skácel se** v zahradě rázem* / *Starý strom v zahradě **se skácel** rázem* / *Starý strom v zahradě rázem **se skácel*** / *Starý strom v zahradě rázem **skácel se***).

[8] K problematice vymezení pozic příklonek ve staré češtině viz Kosek et al. (2018a; 2018b).

PV	2P	non2P	proporce 2P
2	2	19	0,1
3	133	75	0,64
4	81	117	0,41
5	47	49	0,49
6,13	13	18	0,44

Tabulka 1: Počty plnovalenčních doplnění predikátu (PV), výskytů enklitika *sě* v postiniciální pozici (2P), nepostiniciální pozici (non2P) a proporce výskytů enklitika *sě* v postiniciální pozici ve vzorku sestaveném z osmi biblických knih Bible olomoucké (Genesis, Job, Izaiáš, Sírachovec, Matouš, Lukáš, Skutky, Zjevení).



Obrázek 1: Proporce postiniciálního *sě* vzhledem k počtu plnovalenčních doplnění predikátu v korpusu složeném z biblických knih Genesis, Job, Izaiáš, Sírachovec, Matouš, Lukáš, Skutky, Zjevení.

Při ověřování předpokládaného vztahu na jednotlivých textech jsme ovšem zjistili, že u čtyř textů (Matouš, Lukáš, Skutky, Genesis) jsou výsledky ve shodě s predikcí, u dvou výsledky v protikladu a u dvou nebylo možné pro omezené množství dokladů výsledky vyhodnotit. Pro ilustraci jsou v tabulkách 2 a 3 a na obrázcích 2 a 3 prezentovány výsledky, které nejlépe odpovídají predikci, v tabulce 4 a na obrázku 4 výsledek, který je s ní v rozporu.[9]

[9] Je třeba zdůraznit, že tyto výsledky mají pouze ilustrativní hodnotu a představují první vhled do dané problematiky. Vzhledem k malému rozsahu dat (všechna data je nutno anotovat ručně) a porušení předpokladu nezávislosti výběru nebyl předpokládaný vztah testován standardními statistickými nástroji, resp. pokud bychom je v tomto případě použili, došli bychom k statisticky nesignifikantním rozdílům mezi naměřenými hodnotami (na hladině významnosti $\alpha = 0{,}05$). Jinými slovy, rozdíly mezi hodnotami je možno přičíst náhodě. Jedná se však o pilotní analýzu, jejímž cílem bylo ověřit určitý badatelský směr.

PV	2P	non2P	proporce 2P
2,7	15	10	0,6
4	18	18	0,5
5	13	15	0,46
6,18	4	7	0,36

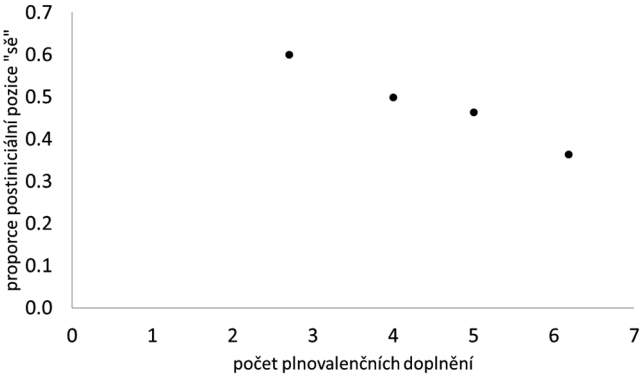
Tabulka 2: Počty plnovalenčních doplnění predikátu (PV), výskytů enklitika *sě* v postiniciální pozici (2P), nepostiniciální pozici (non2P) a proporce výskytů enklitika *sě* v postiniciální pozici v knize Skutky Bible olomoucké.

PV	2P	non2P	proporce 2P
2,86	20	16	0,66
4	18	18	0,5
5,16	8	11	0,42

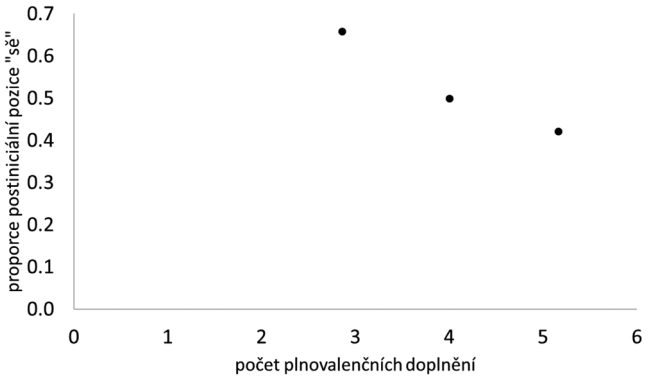
Tabulka 3: Počty plnovalenčních doplnění predikátu (PV), výskytů enklitika *sě* v postiniciální pozici (2P), nepostiniciální pozici (non2P) a proporce výskytů enklitika *sě* v postiniciální pozici v Lukášově evangeliu Bible olomoucké.

PV	2P	non2P	proporce 2P
2,92	23	15	0,61
4	11	21	0,34
5,25	11	9	0,55

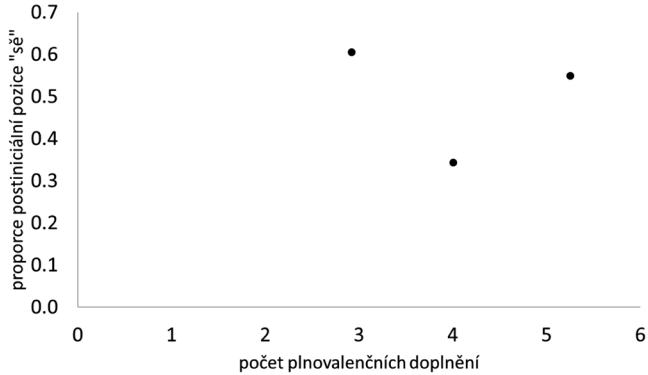
Tabulka 4: Počty plnovalenčních doplnění predikátu (PV), výskytů enklitika *sě* v postiniciální pozici (2P), nepostiniciální pozici (non2P) a proporce výskytů enklitika *sě* v postiniciální pozici v knize Job evangeliu Bible olomoucké.



Obrázek 2: Proporce postiniciálního *sě* vzhledem k počtu plnovalenčních doplnění predikátu ve Skutcích.



Obrázek 3: Proporce postiniciálního *sě* vzhledem k počtu plnovalenčních doplnění predikátu v Lukášově evangeliu.



Obrázek 4: Proporce postiniciálního *sě* vzhledem k počtu plnovalenčních doplnění predikátu ve knize Job.

Bližší pohled na analyzovaný vzorek odhaluje, že platnost dané predikce je zřejmě mimo jiné ovlivněna typem textu. Texty, jejichž analýza není v rozporu s daným předpokladem, mají narativní charakter, kdežto ostatní texty nesou výrazné rysy básnického textu, v případě knihy Job pak jde evidentně o velmi různorodý text, v němž se míchají dialogy s částmi meditativními.

Jiným příkladem je testování platnosti Menzerathova-Altmannova zákona (Altmann, 1980; Cramer 2005), podle něhož by mezi délkami jednotek sousedních jazykových rovin měl být systematický vztah. Tento vztah vyjadřuje funkce (1)

$$y(x) = ax^b e^{-cx},$$

kde x reprezentuje délku tzv. konstruktu (například slova), y průměrnou délku jednotek, ze kterých se konstrukt skládá, jde o tzv. konstituenty (ve vztahu ke slovu půjde

o slabiky), *e* je Eulerovo číslo, *a*, *b*, *c* jsou parametry. Platnost zákona jsme testovali na korpusu sestaveném z pěti různorodých textů: prezidentského projevu K. Gottwalda z roku 1949, prezidentského projevu V. Havla z roku 1990, odborného textu V. Cvrčka Za ještě tvrdší kodifikační diktát? (Naše řeč, 2006, s. 26–29), básně Zlatý kolovrat K. J. Erbena a textu žákyně 6. třídy Já a první žárovka publikovaný ve školním časopise Červotoč (<http://www.sedmikraska.cz/dilna/cervotoc.php>). Konstruktem je v této analýze slovo, jeho délka je měřena v počtu slabik (osa *x* na obrázcích 5, 6 a 7), konstituentem délka slabiky měřená v počtu hlásek (průměrné hodnoty jsou na ose *y* na obrázcích 5, 6 a 7).

Výsledky analýzy jsou prezentovány v tabulce 5. Testujeme-li korpus jako celek, docházíme k vynikající shodě dat s modelem (vyjádřeno koeficientem determinace R^2). V případě jednotlivých textů je sice shoda dat s modelem o něco menší, ale stále dosahuje velmi vysokých hodnot. Sledujeme-li však hodnoty parametrů u jednotlivých textů, zjišťujeme velké rozdíly (s výjimkou parametru *a*, který zhruba odpovídá největší průměrné délce slabik – ta je u všech textů velmi podobná). Viditelně zde hraje velkou roli vliv žánru. Gottwaldův a Havlův projev mají hodnoty parametrů *b* a *c* velmi blízké. Naprosto odlišně se vzhledem k dalším textům chová Erbenův Zlatý kolovrat, jenž má opačné hodnoty znamének těchto parametrů. Specifické chování můžeme vidět dále jak u Cvrčka, tak u žákovského textu.

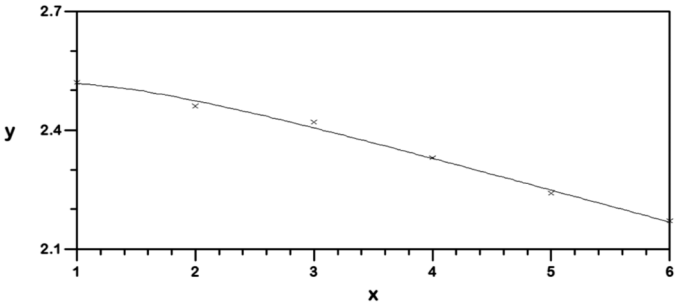
	vše dohromady	**Gottwald**	**Havel**	**Cvrček**	**Erben**	**žákovský text**
a	2,63	2,62	2,59	2,55	2,64	2,57
b	0,04	0,12	0,13	0,25	–0,15	0,05
c	–0,04	–0,07	–0,07	–0,1	0,01	–0,04
R^2	0,99	0,95	0,91	0,93	0,94	0,95

Tabulka 5: Hodnoty parametrů *a*, *b*, *c* a koeficientu determinace R^2 u pěti analyzovaných textů.

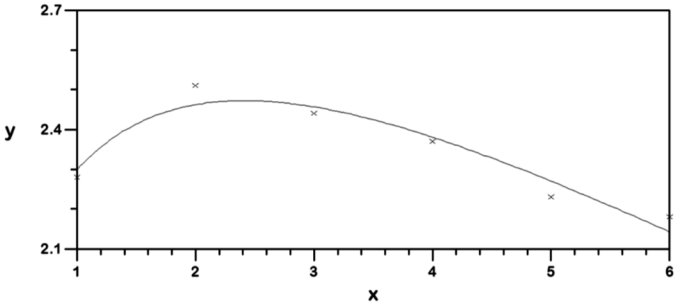
Dané rozdíly jsou ještě zřejmější, sledujeme-li grafické vyjádření vztahu mezi daty a modelem. Pro ilustraci uvádíme výsledky u celého korpusu (obrázek 5) a dále u textu Cvrčkova (obrázek 6), Erbenova (obrázek 7) a textu žákovského (obrázek 8). Pozorujeme-li například průběh funkce u Cvrčka, vidíme, že v průměru nejdelší slabiky jsou u slov o délce dvou slabik. U Erbena jsou pak nejdelší slabiky u slov nejkratších, v případě žákovského textu není rozdíl v průměrné délce slabik u jednoslabičných a dvojslabičných slov.

Sledování jednotlivých textů ukazuje, že daný mechanismus se v jednotlivých textech projevuje rozdílným způsobem, což je ve shodě s pohledem jak Zipfova přístupu, tak synergetické lingvistiky – mluvčí přizpůsobuje své chování danému kontextu. V tomto ohledu je zajímavé se vrátit k výsledku měření založeném na korpusu jako celku, u nějž jsme dosáhli nejlepší shody dat modelem (viz tabulka 5). Na první pohled by se možná mohlo zdát, že tento výsledek je nejlepším důkazem toho, že je smysluplnější pracovat s většími vzorky než s jednotlivými texty. Je však

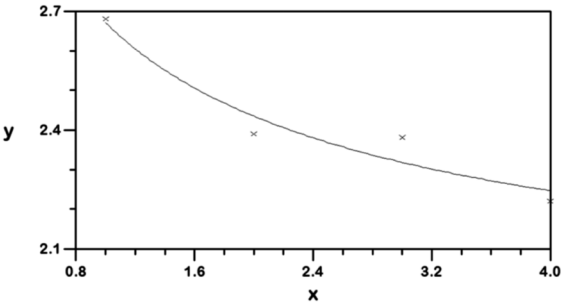
třeba si ale uvědomit, že přestože model sedí lépe na celý korpus, tak „zakrývá“ rozdíly, které vyjadřují odlišné působení daného mechanismu (viz tabulka 5 a obrázky 5–7). Jinými slovy, „zakrývá“ vliv tzv. hraničních podmínek, které omezují jeho platnost. V našem případě by se dalo uvažovat o vlivu žánru či autorství.



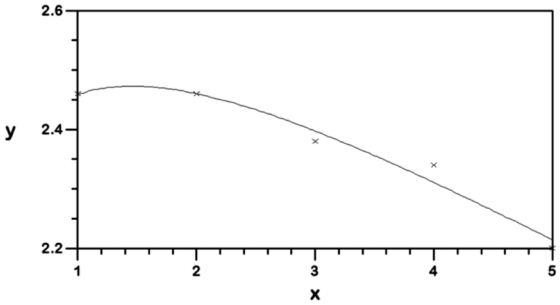
Obrázek 5: Vztah délky slova (měřené v počtu slabik; osa *x*) a průměrné délky slabiky (měřené v počtu hlásek; osa *y*) ve všech pěti textech prezentovaných v tabulce 5.



Obrázek 6: Vztah délky slova (měřené v počtu slabik; osa *x*) a průměrné délky slabiky (měřené v počtu hlásek; osa *y*) v textu V. Cvrčka Za ještě tvrdší kodifikační diktát?, viz tabulka 5.



Obrázek 7: Vztah délky slova (měřené v počtu slabik; osa *x*) a průměrné délky slabiky (měřené v počtu hlásek; osa *y*) v básni Zlatý kolovrat K. J. Erbena, viz tabulka 5.



Obrázek 8: Vztah délky slova (měřené v počtu slabik; osa *x*) a průměrné délky slabiky (měřené v počtu hlásek; osa *y*) v textu žákyně 6. třídy Já a první žárovka, viz tabulka 5.

4 Závěr

Jedním z problémů analýzy jazykového chování je otázka toho, co je vlastně v tomto typu analýz modelováno. V tomto článku jsme se pokusili ukázat, že odpověď není zdaleka tak jednoznačná, jak by se na první pohled mohlo zdát. Navíc je tato otázka spojena s volbou jazykového materiálu a interpretací výsledků, konkrétně jde o to, zda pracovat se soubory textů, či texty jednotlivými. Oba přístupy mají bezpochyby své výhody a limity. Zde jsme se zaměřili především na určité problémy, které jsou spojeny s analýzou jazykových korpusů skládajících se ze souboru textů. Rádi bychom na závěr zdůraznili, že tento článek vnímáme jako pilotní studii, jejíž ambicí je pouze nastínit základní kontury dané problematiky.

LITERATURA

Altmann, Gabriel (1980): Prolegomena to Menzerath's law. In: Rüdiger Grothjan (ed.), *Glottometrika 2*. Bochum: Brockmeyer, s. 1–10.

Beneš, Martin (2015): Máme se vzdát dichotomického pohledu na „jazyk"? *Studie z aplikované lingvistiky / Studies in Applied Linguistics*, 6(2), s. 181–191.

Cramer, Irene M. (2005): Das Menzerathsche Gesetz. In: Reinhard Köhler – Gabriel Altmann – Rajmund G. Piotrowski (eds.), *Quantitative Linguistik / Quantitative Linguistics: Ein internationales Handbuch / An International Handbook*. Berlin – New York, NY: De Gruyter, s. 659–688.

Cvrček, Václav (2014): *Kvantitativní analýza kontextu*. Praha: Nakladatelství Lidové noviny.

Čech, Radek (2005): Komunikace versus systém, nebo komunikace versus model? *Slovo a slovesnost*, 66(3), s. 176–179.

Čech, Radek (2007): Language system – linguistics as an empirical science. *Săpostavitelno ezikoznanie*, 32, s. 42–49.

Čech, Radek (2017): Jazykověda bez langue: odpověď Martinu Benešovi. *Studie z aplikované lingvistiky / Studies in Applied Linguistics*, 8(1), s. 103–110.

Čech, Radek – Pajas, Petr – Mačutek, Ján (2010): Full valency: verb valency without distinguishing complements and adjuncts. *Journal of Quantitative Linguistics*, 17(4), s. 291–302.

Čech, Radek – Kosek, Pavel – Navrátilová, Olga – Mačutek, Ján (2019): Full valency and the position of enclitics in the Old Czech. *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*. Paris: Association for Computational Linguistics, s. 83–88.
Čermák, Franišek (2017): *Korpus a korpusová lingvistika*. Praha: Karolinum.
Eigen, Manfred (1971): Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10), s 465–523.
Haken, Herrman (1983): *Synergetics: An Introduction*. Berlin – New York, NY – Heidelberg – Tokyo: Springer.
Hopper, Paul (1987): Emergent grammar. In: Jon Aske – Natasha Beery – Laura Michaelis – Hana Filip (eds.), *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Grammar and Cognition*. Berkeley, CA: Berkeley Linguistics Society, s. 139–157.
Chromý, Jan (2014): Korpus a reprezentativnost. *Naše řeč*, 97(4–5), s 185–193.
Köhler, Reinhard (1986): *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.
Köhler, Reinhard (2005): Synergetic linguistics. In: Reinhard Köhler – Gabriel Altmann – Rajmund G. Piotrowski (eds.), *Quantitative Linguistik / Quantitative Linguistics: Ein internationales Handbuch / An International Handbook*. Berlin – New York, NY: De Gruyter, s. 760–774.
Komárek, Miroslav (1999): Komunikace versus systém? *Slovo a slovesnost*, 60(3), s. 187–193.
Kořenský, Jan (1987): K procesuálnímu modelování řečové činnosti. *Slovo a slovesnost*, 48(3), s. 177–189.
Kosek, Pavel – Navrátilová, Olga – Čech, Radek – Mačutek, Ján (2018a): Word order of reflexive sě in finite verb phrases in the first edition of the old Czech Bible translation (Part I). *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, 135(3), s. 177–188.
Kosek, Pavel – Navrátilová, Olga – Čech, Radek – Mačutek, Ján (2018b): Word order of reflexive sě in finite verb phrases in the first edition of the old Czech Bible translation (Part II). *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, 135(3), 189–200.
Laury, Ritva – Ono, Tsuyoshi (2005): Data is data and model is model: you don't discard the data that doesn't fit your model! *Language*, 81(1), s. 218–225.
Newmeyer, Frederick J. (2003): Grammar is grammar and usage is usage. *Language*, 79(4), s. 682–707.
Popper, Karl R. (1997): *Logika vědeckého bádání*. Praha: OIKOYMENH.
Prigogine, Ilya – Stengersová, Isabelle (2001): *Řád z chaosu: Nový dialog člověka s přírodou*. Praha: Mladá fronta.
Uhlířová, Ludmila (1995): O jednom modelu rozložení délky slov. *Slovo a slovesnost*, 56(1), s. 8–14.
Zipf, George K. (1949): *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Cambridge, MA: MIT Press.
Zipf, George K. (1949): *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley.

Radek Čech
Katedra českého jazyka FF OU
Reální 5, 701 03 Ostrava
cechradek@gmail.com