Posudek na disertační práci Mgr. Jiřího Miličky *Teorie komunikace jakožto explanatorní princip přirozené víceúrovňové segmentace textů*

Doc. Mgr. Václav Cvrček, Ph.D.
Ústav Českého národního korpusu, FF UK

Disertační práce Mgr. Jiřího Miličky *Teorie komunikace jakožto explanatorní princip přirozené víceúrovňové segmentace textů* je příkladem studie, na níž je především třeba ocenit myšlenkový postup a až v druhé řadě výsledky, ke kterým dospívá (jakkoliv jsou zajímavé). Jejím hlavním tématem je segmentace jazyka, její oprávněnost v popisu, přirozenost a zejména funkce. Přitom, jak už z názvu práce vyplývá, autor si neklade za cíl dospět k důkazům ontogenetické nebo fylogenetické povahy, které by přesvědčivě doložily proč a jak segmentace v jazycích vznikla a je využívána, ale volí cestu funkcionální, tedy hledání vhodného explanačního mechanizmu, který celou problematiku vhodně operacionalizuje. Tímto mechanizmem je pro něj teorie komunikace, která mu vzhledem k tématu poskytuje neotřelý úvahový rámec, v němž je možné dospět k přesvědčivým a inovativním závěrům.

Práce je rozdělena do sedmi kapitol (nepočítaje úvod a závěr), přičemž těžiště práce tkví především v kapitolách 5 až 7. První kapitola nazvaná *Teorie komunikace a jazyk* poskytuje základní definice termínů, s nimiž se v průběhu práce dále operuje (informace, kanál, komplexita, redundance apod.). Výklad je zde i v celém zbytku práce veden na české prostředí až s neobvyklou snahou, aby text byl srozumitelný i pro ty čtenáře, kteří se v problematice teorie komunikace neorientují. To je třeba pokládat za významný klad celé práce, který plně vyváží i zřídka se objevující formální nedostatky či drobné chyby (viz níže).

Druhá kapitola se věnuje Miličkovu specifickému pojetí lingvistiky v kontextu vědy. Elegantně zde vychází z aparátu teorie komunikace představeného v předchozí kapitole a aplikuje ho na vědu jako takovou, která je v jeho pojetí způsobem „jak sběr a kompresi dat provozovat kontrolovatelně a explicitně“ (s. 24). Bylo by možné jistě polemizovat s jeho rozdělením lingvistických směrů na ty, které abstrahují od uživatelů jazyka (kam řadí i korpusovou lingvistiku) a soustředí se výhradně na jazykový systém, ty, které inkorporují uživatele jazyka jako psychosociální jednotku do popisu (např. kognitivní lingvistika či sociolingvistika), a konečně ty, které abstrahují od uživatelů i od jazykového systému ve prospěch mechanických jazykových modelů (zejm. NLP). Diskutabilní je zde především to, zda je oprávněné zařazovat směry, které nezkoumají přímo uživatele jazyka jako psychosociální entity, ale pouze jejich jazykové produkty – texty (analogicky k behaviorálním metodám v psychologii), za směry abstrahující od uživatele en bloc (srov. kognitivně lingvistické práce těžící pouze textů jako produktů jazykové činnosti bez

korekce výzkumem jejich psycholingvistických, neurolingvistických či kognitivních mechanizmů nebo naopak korpusové výzkumy variability jazyka zohledňující nejrůznější sociolingvistické proměnné). Je přitom zřejmé, že rozdíl mezi první a druhou skupinou lingvistických směrů je pouze v tom, které empiricky měřitelné veličiny považuje za adekvátnější obraz jazykových (či obecně kognitivních) mechanizmů – zda texty nebo např. reakční časy, dotazníky či MRI.

Argumentační východisko celé práce je shrnuto v kapitole třetí. K ní se váže i moje nejzásadnější připomínka k celé disertaci zasluhující možná podrobnější diskusi v rámci obhajoby. Autor v úvodu kapitoly tvrdí, že jejím cílem je doložit, že jazyk nutně musí obsahovat distinktivní rysy a morfémy a že ostatní celky logickou nezbytností nejsou (s. 34). Navzdory tomu, že v dalším textu práce opakovaně tvrdí, že v kapitole 3 bylo prokázáno, že tato hypotéza platí, žádný zde uvedený argument nepovažuju za natolik průkazný, aby vylučoval možnost logické nezbytnosti jiné (nadmorfematické) segmentace. Autor zde z teorie informace dedukuje, že ke sdělování informace je třeba alespoň jednu distinkci (jeden distinktivní rys). Takový závěr nepřekvapí; je analogický k laickému pozorování, že se znalostí jednoho slova se nedomluvíme (nebo že k zakódování minimální informace – jednoho bitu – potřebujeme inventář alespoň dvou „symbolů", třeba 0 / 1, true / false nebo zvuk / ticho, na jejichž základě odlišujeme informaci od „pozadí"). Pomocí konkatenace nebo vzájemným ovlivňování (kompozicionalitou) pak můžou vznikat nějaké minimální jednotky; podstatné v této souvislosti podle mého názoru je, že za minimální tyto jednotky označujeme proto, že je podrobněji členit z hlediska informace neumíme, a tudíž se *jeví* jako nutné a samostatné segmenty. Jejich logickou nezbytnost to podle mého názoru sice *sensu stricto* nepotvrzuje, nicméně lze to považovat za akceptovatelné. Pokud jsem ale textu porozuměl správně, nikterak z toho, co bylo řečeno v kapitole 3, nevyplývá, že by segmentace na vyšších (nadmorfematických) rovinách nemohla analogicky být logickou nutností. Stejně jako jsou morfémy minimálními segmenty z hlediska další nedělitelnosti informace, mohou být třeba slova, věty nebo odstavce minimálními (a tedy logicky nutnými) segmenty vyšších rovin z hlediska nedělitelnosti např. myšlenky. Fakt, že existují v jazycích výpovědi (či jejich ekvivalenty) nebo vyšší ucelené (literární) útvary, takovou nezbytnost podporuje i empiricky (což ostatně konstatuje i Jiří Milička v úvodu kapitoly 4).

Segmentaci na vyšších rovinách se věnuje kapitola 4. V této souvislosti stojí za poznámku autorův postoj k problematice lingvistické aktivity ve vztahu k jazyku a mluvčím: „lingvisté nestojí nad jazykem, ale jsou přirozenou součástí jeho vývoje (podobně jako ekologové často spoluutvářejí biotopy, které následně zkoumají, a jako ekonomové zasahují do ekonomiky)." (s. 41) Lingvisté jsou zajisté ovlivněni svojí vlastní jazykovou praxí (sami jsou mluvčími) a tato zkušenost se nutně (ať už vědomě nebo podvědomě) promítá do jejich popisů jazyka. Neznamená to nicméně, že by

museli být nezbytnou součástí jazykového vývoje, což je občas vydáváno za legitimní zdůvodnění pro intervencionalismus při jazykové regulaci. To, zda a do jaké míry budeme jako lingvisté aktivní složkou v procesu jazykového vývoje, je totiž především záležitostí volby; fakt, že si např. česká jazykověda vybrala cestu zásahů namísto role nezúčastněného a „minimálně intervenujícího" pozorovatele, je sice politováníhodné, nicméně v této otázce druhotné, a neměli bychom to rezignovaně přijímat jako běžné, nutné nebo normální.

Kapitola se dále věnuje především segmentaci na úrovni slov a vět, kombinuje přístup diachronní a synchronní s poukazy k různým jazykům a psacím soustavám. Z popisu segmentace „podle mluvčích" (v opozici k popisu „podle lingvistů") vyplývá silná pozice slova, jako přirozené jednotky textu (ačkoli různé jazyky segmentují totéž sdělení různě). V této souvislosti se autor dopouští možná přílišného optimismu, který by mu odborníci na segmentaci a anotaci korpusů mohli závidět, když tvrdí, že „segmentace na slova je jeden z velmi mála způsobů přidávání redundance, které můžeme použít při přepisování již hotového textu, aniž bychom ho radikálně měnili" (s. 44). Tokenizace, tj. segmentace na slova – a to dokonce i v případě jazyka s relativně dlouhou a ustálenou tradicí písemné podoby, jako je čeština – je anotace jako každá jiná, a je tudíž netriviální interpretací textu. Z praxe ČNK můžu doložit, jaké problémy působí texty, které přicházejí vlivem konverzních procesů poškozeny na úrovni slovní segmentace, jak je kolikrát taková závada téměř neodstranitelná či dokonce obtížně automaticky identifikovatelná.

Je přitom trochu s podivem, že v této souvislosti nebyla zmíněna přemístitelnost slova (v rámci věty v jazycích, u kterých to umožňují slovosledné restrikce), která z pohledu laického uživatele jazyka přispívá výrazně ke konstituování přirozené segmentace na úrovni slova. Právě přemístitelnost (spolu s celistvostí, tj. s nemožností vložit segment dovnitř slova) hraje zásadní úlohu pro intuitivní vnímání slova jako přirozené jednotky (ve srovnání s morfémem, výše vymezeným jako minimální segment).

Argumentační oblouk celé práce vrcholí v kapitole 5, která pomocí pokusů se zašuměným kanálem dokládá, že segmentace na rovně slov a rovinách vyšších se v jazycích mohla vyvinout za účelem efektivního vkládání redundance. Autor zde postupně modeluje kanál tak, aby byl komputačně zvládnutelnou a přitom co nejvěrnější aproximací reálné mluvené komunikace. S jeho využitím pak detailně analyzuje různé způsoby vkládání redundance s ohledem na množství přenesené informace. Takto vytvořené modely pak testuje na redundanci v jazyce. Dospívá k vícerozměrnému modelu vkládání redundance, která se jeví být nejvhodnější z hlediska množství správně přenesené informace a tento model následně identifikuje s jednotlivými jazykovými rovinami v jazyce.

Principy vkládání redundance dále podrobněji popisuje kapitola 6. Autor prochází

jednotlivými úrovněmi, od distinktivních rysů a hlásek přes morfémy, slova a věty až k textům. Autor v sobě nezapře filologa v pasáži 6.9, kde se zastavuje ke „spekulativnímu literárněvědnému odpočinku" a komentuje redundanci v poezii. K polemice zde vybízí autorovo tvrzení, že „[k]aždé omezení, které poetická forma klade na autora, vnáší do textu redundanci" (s. 104). Ačkoliv to tak v mnoha případech může být, najdeme mnoho protipříkladů – refrén v Poeově Havranovi je přesným opakem redundantní informace, protože s každým opakováním dostává jiný význam a je tak vlastně kontextově pokaždé jinak disambiguován, stejně tak rým nemusí být pouhým „paritním bitem" držícím pohromadě formu básně, ale prvkem nesoucím svébytný a izolovaný význam atp.

V kapitole 7 se dosavadní výklad o teorii informace, redundanci a s ní související segmentaci spojuje s kvantitativně lingvistickým mainstreamem v podobě Menzerath-Altmannova zákona. Jiří Milička zde inovativně propojuje předchozí výsledky týkající se delimitační informace s Menzerathovým vztahem a porovnává ho s modelem, který pro tento vztah navrhl G. Altmann. Svůj model Milička ověřuje na českých a arabských datech, přičemž výsledky jsou na rovinách, kde to rozsah dat umožňuje, poměrně přesvědčivé. Je přitom poněkud paradoxní, že vyšší roviny (odstavce, kapitoly) budou z hlediska zásad kvantitativní lingvistiky pravděpodobně netestovatelné, protože materiál potřebného rozsahu překračuje rámec jednoho díla, což je požadavek pro mnoho badatelů v této oblasti nepřekročitelný.

Text obsahuje zcela zanedbatelné množství formálních chyb (překlepy, zdvojená slova) a je přehledně uspořádán. Z drobných věcných nedostatků bych ještě jmenoval tyto:

- Na s. 12 v pozn. 3 by nejspíš mělo být uvedeno, že po „11" následuje „10" 73krát, a nikoli 74krát.
- Příklad ztrátové komprese na s. 12 nespíš neměl popisovat situaci různě dlouhého vstupu, ale princip zanedbání části informace; zápis `(11(10){73}){3}` je ekvivalentní vstupu o délce 444 znaků, zatímco „ztrátově komprimovaný" zápis `(10){225}` generuje vstup o délce 450 znaků (nedochází tak pouze k zanedbání sekvence „11", ale i k přidání dalších 3 dvojbitů „10"). Z hlediska srozumitelnosti výkladu by bylo lepší ztrátovou kompresi vysvětlit na identicky dlouhých vstupech.
- V příkladu na s. 84 se systematicky objevuje chyba, kdy je samohláskám *u* a *ú* přiřazen rys „přední" a samohláskám *i* a *í* rys „zadní".
- V pozn. 16 na s. 86 se mluví o oblibě asociačních měr; pokud bych odhadoval žebříček oblíbenosti, na předních místech by asi nebylo MI-score, ale spíše logDice (používané ve velmi rozšířené aplikaci SketchEngine), log-likelihood (užívané hodně germanisty) či z-score (tradičně užívané hlavně v pracích britských lingvistů).

– Tvrzení, že v češtině „předložky musí stát bezprostředně před jmény, ke kterým se váží“ (s. 99) není zcela přesné, protože se v češtině nezřídka objevují případy dvou předložek za sebou (*přišel s pro mě nečekaným úlovkem*) nemluvě o situaci, kdy mezi předložku a jméno je vložen další rozvíjející člen (*o jarním kvítí*).

Jiří Milička napsal inovativní a inspirativní studii s neotřelým tématem a způsobem zpracování. V českém prostředí možná trochu neobvyklý anglosaský esejistický styl může kontrastovat s hloubkou myšlenek, záběrem témat i s exaktním zpracováním empirických či matematických částí. Je však třeba ocenit, že čtenářovo pochopení textu leží autorovi na srdci, čímž se vymyká běžné produkci kvalifikačních prací u nás. Jiří Milička touto prací znovu prokázal, že je originálním badatelem, který se obtížně vměstnává do standardních škatulek lingvistických oborů a směrů. Jeho v pravdě interdisciplinární přístup k tématům, jimž se věnuje, jazyková rozkročenost mezi evropskými a asijskými jazyky i schopnost nahlížet problematiku z netradičních úhlů slibuje do budoucna mnoho cenných výsledků (pokud se rozhodne ve vědecké kariéře pokračovat).

Předložená disertační práce Mgr. Jiřího Miličky *Teorie komunikace jakožto explanatorní princip přirozené víceúrovňové segmentace textů* bezpochyby splňuje podmínky na kvalifikační práce tohoto typu běžně kladené, a doporučuji ji proto k obhajobě.

V Praze, 11. února 2016