STRUKTURA TEXTU A SÉMANTIKA

PhDr. Luděk HŘEBÍČEK DrSc.
Orientální ústav AV ČR

Abstract
This essay could be comprehended as a brief introduction to quantitative linguistics. How are brain structures connected with semantics of language? The answer is to be found in the text.

Dvacáté století bylo v lingvistice dobou převratnou. Na začátku století se o to zasloužil strukturalismus. Když se moje generace v padesátých letech lingvisticky vzdělávala, byly myšlenky strukturalismu sice prohlašovány za buržoazní pavědu, ale my jsme chápali, že jde o nejnovější výdobytky teorie jazyka. Strukturalismus přinesl termín „lingvistika"; tím se teoretický obor odlišil od filologie a byl zdůrazněn rozdíl mezi vědeckým a předvědeckým pojetím jazyka. Měli bychom být vděčni strukturalismu za to, že každou chvíli z našich lingvistických úst a per splyne slovo „struktura" nebo „systém". Ve skutečnosti ovšem i strukturalismus byl jen filologií, nikoliv lingvistikou. Musíme být vděčni především profesoru **Vladimíru Skaličkovi** za podněty k tomu, aby se úvahy lingvistů poohlédly též po metodických postupech jiných věd.

Věda je obor lidské činnosti, který ve vztahu k pozorovaným jevům razí pojmy, specifické konvence, a formuluje hypotézy. Nikoliv však libovolné hypotézy, třeba nějaké teoreticky se tvářící výroky, ale odmítnutelné nebo testovatelné hypotézy, které do té doby, než jsou odmítnuty, mají platnost vědeckých zákonů. Tímto způsobem pojímají vědu moderní epistemologové (odkazuji především na **Karla Reimunda Poppera** a **Maria Bungeho**).

Ještě v padesátých letech, ale i později, jsme se učili, že jazyk představuje specifický systém, jehož vlastností je, že se skládá z relativně nezávislých podsystémů. Kdo tomu nevěří, ať nahlédne do libovolné gramatiky nebo učebnice konkrétního jazyka, kde najde nejdříve kapitolu o hláskovém systému, potom o morfologii, slovu a konečně o syntaxi. Z nepravdy je však tento nazor usvědčován *každým* písemným nebo mluvním projevem, v němž jednotky různých podsystémů vytvářejí pevný jednotný systém, jehož složky dokážou mluvčí i příjemci uvést do souvislosti. Jev, kterému budeme říkat **text**, byl pro teorii jazyka vědecky neuchopitelný. O textové lingvistice se začlo mluvit v šedesátých letech. Ale i v tomto oboru došlo jen k formulování pojmů a jejich vztahů spolu s některými konvencemi, zejména s ohledem na gramatické struktury konkrétních jazyků, nikoliv však k formulování zákonů (hypotéz).

Velmi jednoznačná stanoviska k těmto věcem formuloval **Gabriel Altmann**, německý lingvista slovenského původu a Skaličkův žák, například v knize Wimmer et al. (2003: 13-18).

V roce 1980 publikoval Gabriel Altmann článek v němž formuloval a odvodil Menzerathův zákon, o kterém nyní mluvíme jako o **Menzerathově-Altmannově zákonu**. Paul Menzerath v roce 1928 publikoval pozorování týkající se délky slov v počtu slabik a délky slabik ve slově. Zjistil, že čím delší je slovo v počtu slabik, tím kratší je v průměru

délka slabiky v počtu hlásek. Toto významné pozorování však zůstalo až do roku 1980 v podstatě nepovšimnuto.

Altmann zavedl dva pojmy: **jazykový konstrukt** a jeho **konstituent**. To byl zásadní krok při formulaci lingvistického Menzerathova-Altmannova zákona, který věcně souvisí s obecným **principem konstituence**. Ve skutečnosti zmíněný zákon je variantou tohoto principu a můžeme jej slovně formulovat takto:

Čím delší (větší, složitější atd.) je jazykový konstrukt, tím kratší (menší, jednodušší atd.) je v průměru jeho konstituent.

Altmann zároveň odvodil algebraickou podobu tohoto zákona:

$$y=Ax-b$$

kde *y* je průměrná velikost konstituentu,
x je průměrná velikost konstruktu,
A a *b* jsou parametry.

Když se nad touto formulací zamyslíme, objeví se některé lingvisticky zajímavé vlastnosti daného vztahu.

Především je zřejmé, že popisované veličiny charakterizují dvě různé jazykové jednotky: konstrukt jako určitou jazykovou jednotku a konstituent nebo konstituenty jako jednotky, z nichž se konstrukt skládá. Tím jsou k sobě vztaženy dvě sousední jazykové úrovně a zároveň je obecně definována **jazyková úroveň**. Za jazykovou úroveň a její jednotky považujeme to, co splňuje Menzerathův-Altmannův zákon.

Jazykové úrovně známe z pozorování konkrétních jazyků. Víme že existují tyto jednotky:

hláska – morf – (slabika) – slovo – syntaktická konstrukce – věta.

Máme tedy možnost kdykoliv ověřit platnost tohoto zákona. To se stalo v řadě prací, především v publikaci Altmann & Schwibbe et al. (1989). V aplikaci zákona vždy nejdříve zjistíme počet slabik ve slově *x* a k jednotlivým hodnotám *x* přiřadíme průměrnou délku slabiky *y*.

Působení zákonitostí toho typu, k němuž patří i Menzerathův-Altmannův zákon, nezjistíme ovšem pozorováním jednoho slova v textu. Systém, jehož součástí je slovo, funguje na probabilistických principech, musíme tedy pracovat s nějakým statistickým souborem. Působení zákonů se v jazykových systémech projevuje jako **tendence směřující k nějakému stabilnímu stavu**. Při zkoumání zákona nejdříve uspořádáme zjištěné hodnoty do tabulky, nakreslíme jejich graf a ptáme se, jak se pozorovaná křivka odlišuje od teoretické menzerathovské křivky. Existují prostředky, jak zjistit, zda je rozdíl mezi dvěma křivkami významný nebo nevýznamný.

Uvažujme však dále o spektru jazykových úrovní, přičemž nejvyšší jednotkou je věta. Skutečně v jazycích existují jen výše uvedené druhy jednotek? (Připomínám, že ve fonetice jsme schopni délku hlásek měřit docela snadno a zkoumat platnost zákona i pod úrovní hlásek – tolik k dolní příčce žebříčku úrovní.) Kolem roku 1988 jsem se zabýval otázkou existence **nadvětné jazykové úrovně** a nejdříve jsem se ptal, zda tuto úroveň tvoří texty jako její jednotky. Otázka tedy zněla, zda text je konstrukt, jehož konstituenty jsou věty (obecně říkejme **segmenty** textu). Záporná odpověď na tuto otázku je nasnadě. Ptal jsem se, jak je tedy tvořena souvislost vět, čím v textu věty spolu souvisí.

Při zkoumání pozorovaných textů jsem se zaměřil na jazykové projevy vzniklé v přirozených podmínkách jazykové komunikace, (1) které jsou souvislé (2) a jsou schopny poskytnout měřitelná a nezkreslená data.

Rozhodl jsem se uvažovat lexikální jednotku v textu ve dvou různých okolích:

(A) **v segmentu textu**, který je nejčastěji syntaktickým segmentem (větou, klauzí) nebo metrickým segmentem (např. veršem);

(B) vyšším okolím lexikální jednotky je **soubor všech segmentů, v nichž se daná lexikální jednotka vyskytuje**.

Nechci uvádět všechny podrobnosti tohoto postupu, je tu celé řada otázek a specifických konvencí formulovaných s ohledem na konkrétní jazyk. Podstatné však je konečné zjištění, že (B) se chová jako konstrukt a (A) jako jeho konstituent ve smyslu Menzerathova-Altmannova zákona.

To bylo ověřeno na textech různých stylů a v jazycích různých typů. Jestliže tato teorie je správná, můžeme na jejím základu vyslovit několik **závěrů**:

1. V textech existuje skrytá a dosud nepopsaná jazyková úroveň.
2. Každá lexikální jednotka vytváří v textu sémantický konstrukt, jehož konstituenty jsou segmenty v nichž se vyskytuje.
3. Každý segment textu je konstituentem těch sémantických konstruktů, tvořených lexikálními jednotkami, které se v něm vyskytují.
4. Text je systém, jehož dynamika je tvořena přeměnou slovního tvaru v lexikální jednotku a tedy v sémantický konstrukt.
5. Text má povahu lingvistické jednotky, kterými jsou např. věta, slovo či hláska.
6. Systém jazykových úrovní je množina tvořená na principu **soběpodobnosti**, má tedy podstatné vlastnosti **fraktálu**.

Tyto a ještě některé další důsledky stručně vylíčené teorie ukazují, že text nemůže být vyloučen z jazykového systému bez zkreslení tohoto systému. Naopak – co je jazyk, poznáme teprve, když zahrneme mezi jeho prvky i to, čím jsou k sobě vázány věty (spolu se svými strukturami) v jejich přirozené souvislosti, tedy v textu.

Na základě těchto poznatků jsem si dovolil formulovat axiom, podle něhož dynamiku textového systému vytváří vztah, který nazývám **sémantickou specifikací**. Jde o následující formulaci:

> Kolokací lexikálních jednotek v segmentech textu a v sémantických konstruktech jsou specifikovány významy lexikálních jednotek, které skutečně mají v textu.

Tato formulace není nikterak objevná, všichni víme, že lexikální význam se v textech proměňuje. Význam tohoto axiomu spočívá v jeho opření o teorii, v jejímž jádru je princip konstituence.

S ohledem na kolokaci jednotek a sémantickou specifikaci lze vytvořit **grafický obraz**, který se zakládá na dvou množinách:

(1) na množině uzlů grafu (*V*) čili lexikálních jednotek a

(2) na množině hran (*E*) čili jejich kolokacích.

V nějakém libovolném textu lze zjistit, že *V* má prvky vyznačující se jednak frekvencí, což je dimenze, kterou v grafu sémantické struktury textu můžeme, ale nemusíme vzít v úvahu; dále počtem hran, které s daným uzlem incidují a konečně vzdáleností, jakou má daný uzel od všech ostatních uzlů. Tuto vzdálenost měříme v počtu hran.

V této chvíli se ocitáme ve sféře **teorie grafů a sítí**, kde se můžeme opřít o poznatky rozvinuté v tomto matematickém oboru. Proto odkazuji na práci Watts (1999), ale mohl bych odkázat na mnoho dalších prací nádherného a v humanitních oborech s výhodou využitelného oboru.

Poznamenejme ještě, že sémantický obraz textu, jak jsem jej načrtl, by asi poskytl dosti nepřehlednou strukturu, lexikální jednotky a jejich vztahy jsou četné. Proto je nezbytné rozčlenit sémantickou strukturu na nějaké **jádro a periferii** textu, přičemž se můžeme opřít právě o frekvenční vlastnosti. Např. lexikální jednotky, které se vyskytují v textu jen jednou (tak zvaná hapax legomena, která ve vzorci Menzerathova-Altmannova zákona mají vynikající postavení, jejich četnost je evidentně rovna parametru *A*) lze z grafu vyčlenit a pracovat s ostatními jednotkami jako s jádrem. Graf sémantického jádra textu lze charakterizovat např. podle **centrality** (*betweenness centrality*). S výsledky analýzy grafu pak můžeme dále pracovat a využít jich pro objektivní a smysluplnou explanaci textu.

Dosud jsme o platnosti principu konstituence pro texty v přirozeném jazyce uvažovali tak, že jsme pro každé *x* vypočítali *y* bez rozlišení jednotlivých lexikálních jednotek. Nyní však vypočítejme **průměrnou velikost konstituentu** pro každou jednotlivou lexikální jednotku. Mějme tedy obecně nějakou lexikální jednotku *i* v daném textu; ta se vyskytuje v segmentech o délkách *s1, s2, ..., si ,..., sf* a nechť jejich suma je *Si*. Symbol *f* je v daném případě frekvencí lexikální jednotky *i*, tedy vlastně *fi*. Zaveďme veličinu *wi*, která je průměrnou délkou segmentů, v nichž se *i* v textu vyskytuje, čili

wi = Si / fi

O této veličině můžeme mluvit jako o **kontextuální váze lexikální jednotky** v textu. *GRAF* naznačuje, jak se tato veličina chová v textech různých jazyků a různých stylů.

Ve všech případech podobných analýz získaly prominentní postavení jednotky s kontextovou vahou max *wi*(*fi*), čili lexikální jednotky, které mají v grafech postavení „vrcholů vln“ této veličiny. Statistická analýza těchto maxim potvrzuje, že vyhovují Menzerathovu-Altmannovu zákonu. Z jednotlivých grafů je vidět, že tyto „vrcholy vln“ skutečně vytvářejí menzerathovskou křivku. Můžeme tvrdit, že veličina max *wi* (*fi*) zobrazuje **sémantický atraktor textu**, k tomu viz např. Williams (1997).

Když jsme dále hledali vyšší úrovně, tj. úroveň nadřazenou úrovni textu s konstrukty, k nimž texty tvoří konstituenty, pokus rozvíjet dál tuto teorii se dostal až do oblasti obecné sémantiky a teorie poznání. To však je problematika příliš rozsáhlá, aby pro ni jedna přednáška poskytovala dostatečný prostor.

Literatura:

Altmann, G. (1980): Prolegomena to Menzerath's law. *Glottometrika 2,* Brockmeyer, Bochum.
Altmann, G. & Schwibbe, M. H. et al. (1989): *Das Menzerathsche Gesetz in informations-verarbeitenden Systemen.* Olms, Hildesheim.
Watts, D. J. (1999): *Small worlds. The dynamics of networks between order and randomness.* Princeton U. P., Princeton (N. J.).
Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S. & Wimmerová, S. (2003): *Úvod do analýzy textov.* Veda, Bratislava.
Williams, G. P. (1997): *Chaos theory tamed*. Joseph Henry Press, Washington, D. C.