Sémantické slapy v textových strukturách

LUDĚK HŘEBÍČEK

Semantic cataracts in text structures

ABSTRACT: Assume a semantic space demarcated by a set of lexical units occurring in a text. Their transition into word forms is a dynamic semantic process connected with the collocation of the units in text segments. The aim is to find semantic constructs and constituents and to apply Menzerath-Altmann's law, which defines the relationship of language constructs and constituents to the semantic text level. Any text is expected to provide a characteristic arrangement of lexical units in their contextual bonds. This can be explained as a feature of semantic dynamism and possibly also of human thinking. We can metaphorically call it a *semantic cataract*.

Key words: language construct and constituent, Menzerath-Altmann's law, text segment, semantic construct, semantic (contextual) weight of lexical units

Klíčová slova: jazykový konstrukt a konstituent, Menzerathův-Altmannův zákon, textový segment, sémantický konstrukt, sémantická (kontextuální) váha lexikálních jednotek

Úvod

V moderní lingvistice je věnována zvýšená pozornost jazykovým strukturám, které přesahují úroveň věty (Ziegler – Altmann, 2002; Wimmer et al., 2003). Jde o úroveň, na níž formalizace popisu způsobem uplatněným klasickou lingvistikou na nižších jazykových úrovních naráží na obtíže. Hledá se tak dovršení celkové stavby jazykových systémů a podsystémů, mezi nimiž se jinak obtížně nachází princip či zákonitost, která je spojuje v jeden celek.

Zkušenost ukazuje, že typů analýzy textu může být nespočet i na vědecké úrovni poznání, navíc k nim patří analýzy prováděné intuitivně uživateli přirozených jazyků při produkci i recepci textů. Zjednodušeně řečeno, jde o nalezení systémových vazeb větných struktur v textu a o popis tohoto systému. Ve skutečnosti jde také o pochopení textu jakožto jazykové jednotky navzdory jeho enormní variabilitě.

Takový cíl přináší do lingvistiky některá zvláštní hlediska:

1. Postup poznání jazyka od věty k textu, k útvaru složenému z vět, jako by u textu končil, protože nad textem už v mezích lingvistiky žádný bezprostředně pozorovatelný jazykový útvar nenacházíme.
2. Ve vrcholných strukturách textu se jazykový kód transformuje do sémantického útvaru, v němž se těsná vazba na jazykové jednotky (zřetelná u slov v lexikálních významech, u vět stále ještě viditelná v jejich syntaktických strukturách) proměňuje ve složitě (možná i chaoticky) organizované sémantické útvary. Tím se textová lingvistika stává sémantickým oborem, který hledá formální utváření velmi členitých systémů. Ani ty se však nezdají být zachytitelné jinak než prostřednictvím jazyka v textech. Zdůrazněme, že navzdory přechodu od jazykového kódu k významovým strukturám stále ještě jde o lingvistickou sémantiku. Lingvistika je nenahraditelná, obecná sémantika i logika jsou obory, které se nezabývají jednotlivostmi lidského myšlení v jeho vaz-

bě na sémantický systém představovaný určitou individuální mentalitou. Naopak obory jako psychologie či psychiatrie operují s jazykem povýtce intuitivně.
3. Přechod do sémantického abstraktna neznamená ztrátu ukotvení získaných forem v jazyce. Jak jazykový kód, tak i uspořádání významových struktur nejsou vlastnostmi nějakého abstraktního prostoru; jsou důsledkem aplikace mozkových funkcí v podmínkách sociální komunikace. V textech nacházíme otisk mozkových funkcí, obraz jejich dynamiky při generování významů a jejich transformaci do jazykových struktur. Tato okolnost zvyšuje důležitost lingvistického výzkumu v daném oboru.

Základy teorie

O které jazykové zákonitosti se hledání textových struktur opírá?

Lze přijmout za prokázaný funkční vztah mezi frekvencí lexikální jednotky v textu a velikostí textového segmentu, v němž se tato jednotka vyskytuje. Je evidentní, že taková funkce nesporně musí být funkcí pravděpodobnostní, která nedává deterministicky přesné výsledky predikované na základě určitého vztahu, nýbrž výsledky odpovídající celkové tendenci vztahem vyjádřené. Historie popisu zmíněného funkčního vztahu byla již několikrát v literatuře popsána. Shrnout ji lze takto:

Jazykový konstrukt je obecně definován jako jednotka určité jazykové úrovně a jeho konstituent jako jednotka úrovně bezprostředně nižší. Jejich vztah popsal Gabriel Altmann a odvodil pro něj vzorec

(1) $y = Ax^{-b}$,

kde x = velikost konstruktu, y = velikost konstituentů vyjádřená střední hodnotou, přičemž A a b jsou parametry tohoto vztahu (Altmann, 1980; Altmann – Schwibbe et al., 1989).

Jeho platnost byla prokázána na gramatických úrovních jazyka, tedy pro jednotky od fonémů až po věty. Altmann pro tento vztah navrhl název „Menzerathův zákon“, později byl zcela oprávněně nazván „Menzerathovým-Altmannovým zákonem“. Podle tohoto vzorce s rostoucí velikostí konstruktu klesá velikost konstituentu.

Platnost tohoto zákona byla rozšířena i na textovou úroveň (Hřebíček, 1989, 1992, 1997, 2002, 2005). Na této v podstatě sémantické úrovni jsou jazykové konstituenty tvořeny segmenty textu. Segmentem je věta nebo nějaký jinak definovaný útvar o přibližné velikosti věty, např. verš. Podle této teorie sémantický konstrukt je tvořen souborem segmentů, v nichž se vyskytuje daná lexikální jednotka. Velikost konstruktu je rovna počtu segmentů, jež jsou jeho konstituenty. U konstituentů je velikost obvykle vyjádřena průměrným počtem slov v konstituentech daného konstruktu.

Působení tohoto zákona bylo ověřováno způsobem, který je dokumentován v tabulce 1.

Sémantický konstrukt je v této tabulce představen svou velikostí x. Pokud jde o konstituenty, byly sečteny velikosti všech segmentů, jež tvoří konstituenty konstruktů o velikosti x; z tohoto součtu byl vypočítán průměr y. Z pozorovaných hodnot y je zřejmé, že očekávaný pokles neplatí absolutně pro každou hodnotu zvlášť, ale projevuje se jako celková klesající tendence proměnné y spolu s rostoucím x.

Tabulka 1: Menzerathův-Altmannův zákon na úrovni sémantických konstruktů.

x	*z*	*y*
1	304	17,13
2	66	15,91
3	30	17,33
4	13	14,88

x	*z*	*y*
5	6	17,00
6	7	14,90
7	8	15,88
8	2	12,81

x – sémantický konstrukt, *z* – počet lexikálních jednotek, *y* – průměrná velikost konstituentu.
Text: Ivan Kraus, Klášterní ulice. In: Ivan Kraus, Číslo do nebe. Praha: Marsyas, 1993.

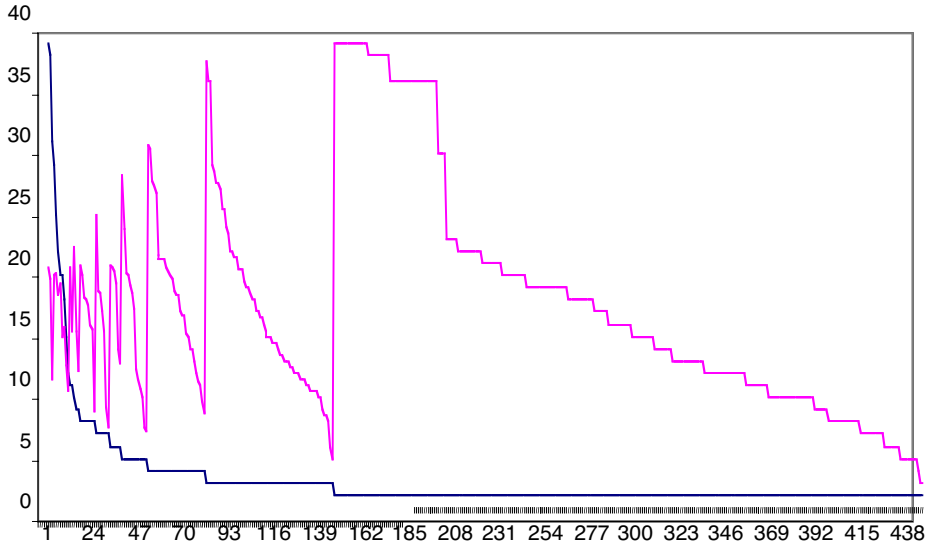
Sémantika textu je v navrženém způsobu analýzy zredukována na zobrazení vztahu dvou strukturních složek: lexikálních jednotek a textových segmentů. Ve skutečnosti ovšem v jazykových kódech k významům odkazují i další jednotky. Podle dosavadních zkušeností s různými jazyky by každý analyzovaný text měl být na počátku analýzy pečlivě interpretován (přepsán) tak, aby formální lexikální jednotky s gramatickou funkcí nebyly postaveny naroveň plnohodnotným sémantickým jednotkám. Různé typy morfémů, zejména ty s referenční funkcí by měly být nahrazeny příslušnými lexikálními jednotkami. Mělo by platit pravidlo, že určitý význam by měl být v jazykovém kódu textu po jeho interpretaci reprezentován jednou určitou lexikální jednotkou. To znamená, že všechny významové posuny slov v podobě synonymie, homonymie apod. by měly být rozpoznány a převedeny na příslušné lexikální jednotky tak, aby platilo zmíněné pravidlo. V citované literatuře je interpretace textu podrobně popsána a doložena na příkladech.

Při analýze zjišťujeme přítomnost určité lexikální jednotky v daném segmentu; jestliže v témže segmentu se vyskytuje určitá lexikální jednotka víckrát než jednou, počítáme její výskyt pouze jednou. Při této evidenci lexikálních jednotek tedy existuje rozdíl mezi frekvencí *i*-tého slova f_i a velikostí sémantického konstruktu x_i, odpovídajícího dané lexikální jednotce. Prakticky ovšem je tento rozdíl tak malý, že je statisticky nevýznamný. Proto o velikosti sémantického konstruktu můžeme mluvit též jako o frekvenci lexikální jednotky v textu a zavést tak identitu $x_i \equiv f_i$.

Slovo v kontextu

Sémantická struktura textu je pojem blízký pojmu „kontext“. Je to slovo s širokou sémantickou platností. I v lingvistice lze jeho význam považovat za více či méně intuitivní. Menzerathův-Altmannův zákon aplikovaný na sémantickou úroveň a chápaný jako základní kámen sémantické struktury textu vlastně zavádí dvě úrovně kontextu:
(a) úroveň segmentální a
(b) úroveň textovou.

Skutečně, lexikální jednotka např. ve větě je určována sémantickými vztahy uvnitř segmentu, ale spolu s ním působí na celý text v širším vztahu k dalším jednotkám, tedy přes meze dané segmentem. Evidentně je to důsledek rozlišení konstruktu a konstituentu na sémantické úrovni textu. Existují tedy úrovně kontextu dané kolokací slova



Obrázek 1: Sémantická váha w_i (křivka kaskádovitého nebo slapovitého tvaru) v závislosti na frekvenci f_i (křivka tvaru L); vodorovná osa zobrazuje jednotlivé lexikální jednotky i.

s jinými slovy jednak v segmentu a jednak prostřednictvím příslušného konstruktu v celém textu. Umístěním lexikální jednotky v segmentu a v sémantickém konstruktu dochází k proměně lexikální jednotky v slovní tvar, což představuje dynamiku, která vlastně definuje text jako lingvistickou jednotku. Bylo navrženo označit tuto změnu za *sémantickou specifikaci* lexikální jednotky.

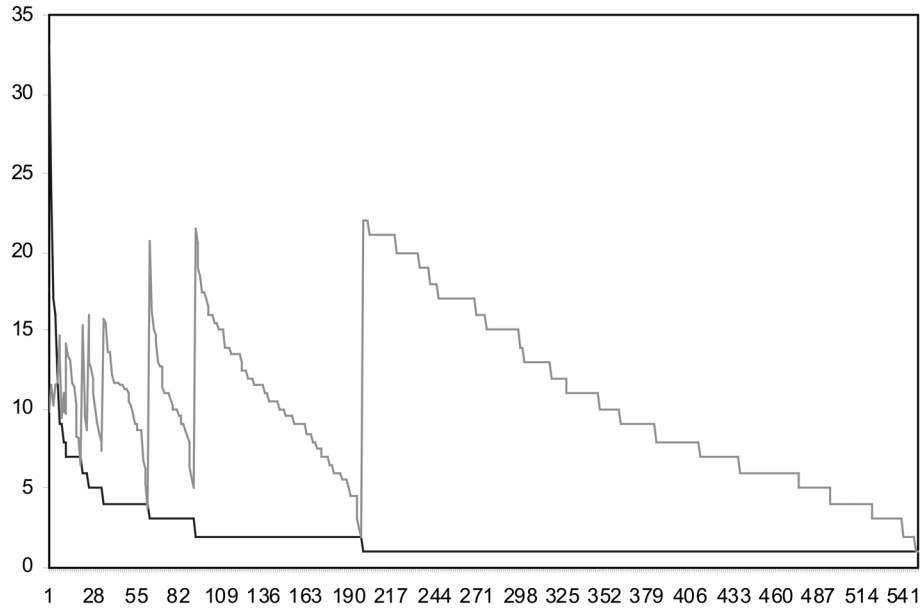
V tabulce 1 byl text a jeho struktura na sémantické úrovni charakterizován vztahem celkové délky příslušných segmentů k frekvenci lexikální jednotky, jak předepisuje Menzerathův-Altmannův zákon. Při takovém způsobu evidence působení tohoto zákona byly do jisté míry zanedbány lexikální jednotky. Jiný úhel pohledu, při němž jsou respektovány jednotlivé lexikální jednotky, dává jiný a dosti zajímavý obraz struktury textu. K tomuto účelu zaveďme pojem *sémantické* (nebo též *kontextuální*) *váhy* lexikální jednotky. Definujme ji vztahem

$$(2) \quad w_i = \frac{S_i}{f_i},$$

kde i je lexikální jednotka, f_i její frekvence v textu a S_i součet délek s_i všech segmentů, v nichž se v daném textu lexikální jednotka i vyskytuje.

Je zřejmé, že w_i je ve skutečnosti průměrnou délkou segmentu, která v textu přísluší lexikální jednotce i. Pomocí této veličiny lze sémantickou strukturu téhož textu, který je popsán v tabulce 1, zobrazit způsobem užitým v obrázku 1.

Složitá křivka slapovitého tvaru v tomto obrázku ve skutečnosti spojuje body odpovídající funkci $w_i(f_i)$, čili sémantické váze zobrazené jako funkce frekvence slova i. Zdůrazněme, že tento tvar křivky jsme dostali pro všechny texty souboru textů z růz-



Obrázek 2: Sémantická struktura textu jako na obrázku 1. Turecký text, kapitola z románu Demir Özlü, Bir Yaz Mevsimi Romansı (Romance jedné letní sezóny), Istanbul: Ada, 1990, s. 63–68.

ných jazyků. Do tohoto souboru byly zcela náhodně zařazeny texty různých stylů. Jako doklad uvádíme obrázek 2, který zobrazuje výsledky analýzy tureckého textu.

Obecně má křivka $w_i(f_i)$ tři charakteristické vlastnosti, které odpovídají následujícím veličinám:

1) max $w_i(f_i)$, tj. vrcholu „vlny“, někdy představovaném více než jedním i;
2) $<w_i(f_i)>$, což je průměr hodnot odpovídajících jedné vlně pro určitou frekvenci;
3) min $w_i(f_i)$, tj. nejnižší hodnotě dané vlny, rovněž někdy odpovídající více než jednomu i.

Zkráceně je píšeme max w_i, $<w_i>$, min w_i. Je zřejmé, že střední hodnota $<w_i>$ je totožná s proměnnou y Menzerathova-Altmannova zákona, čili na sémantické úrovni textu platí identita $y \equiv <w_i>$. Pokud jde o min w_i, u řady pozorovaných textů nelze pro tuto proměnnou vymezit nějakou charakteristickou vlastnost.

Naopak max w_i se ukazuje být dosti důležitou veličinou. Grafické zobrazení navozuje představu, že vrcholy jednotlivých slapových vln zobrazených touto proměnnou v textech různých jazyků tvoří menzerathovskou křivku (jak tuto křivku obvykle nazývá Gabriel Altmann).

Je-li tomu skutečně tak, pak se tím zvyšuje důležitost této sémantické struktury. Hodnoty tří charakteristických proměnných (pozorovaných v textu, k němuž se vztahuje obrázek 2) jsou uvedeny v tabulce 2. Zvolili jsme tento text proto, že je na něm zřetelně vidět, jak obtížně prokazatelný je zde pokles středních hodnot $<w_i>$ při rostoucí frekvenci. V tabulce 2 je vidět pokles od hodnoty $<w_i>$ = 10,36 k hodnotě 10,21. V tabulce

jsou vynechány údaje odpovídající jen jediné lexikální jednotce, u nichž celková tendence může být zkreslena. (Ve skutečnosti frekvenci slova sen „ty“ $f_i = 33$, což je nejfrekventovanější jednotka v tomto textu, odpovídá pokles na hodnotu $\langle w_i \rangle = 9{,}97$.)

Tabulka 2: Hodnoty sémantické váhy slov pozorované v tureckém textu (jde o týž text jako na obrázku 2).

f_i	max w_i	Očekávané max w_i	$\langle w_i \rangle$	min w_i
1	22,00	22,00	10,36	1,00
2	21,50	19,25	10,45	2,00
3	20,67	17,80	10,83	5,00
4	15,75	16,84	10,51	3,75
5	16,00	16,13	10,74	7,40
6	15,33	15,57	10,96	8,67
7	14,14	15,11	10,76	6,43
8	11,00	14,73	10,21	9,75
9	14,67	14,40	11,48	9,44

Dodejme ještě, že pokud jde o min w_i, kromě slabé korelace s frekvencí nelze u této veličiny pozorovat nějakou další charakteristickou vlastnost.

Jinak je tomu u proměnné max w_i, u níž je pokles vždy velmi zřetelný. Tento průběh se opakuje u všech zkoumaných textů bez rozdílu jazyka, což vede k zamyšlení nad příčinou takového průběhu dané funkce. Očekávané hodnoty max w_i jsou uvedeny ve třetím sloupci tabulky 2 a jsou vypočítány pomocí rovnice

$$\max w_i = 22 f_i^{-0{,}1929},$$

což je Menzerathův-Altmannův zákon formulovaný pro y s parametry $A = 22$, $b = -0{,}1929$ při identitě $x \equiv f_i$. Jestliže pro testování rozdílu mezi distribucí pozorovaných a očekávaných hodnot použijeme neparametrický Wilcoxonův test, dostaneme výsledek, podle něhož lze přijmout hypotézu o nesignifikantnosti rozdílu mezi oběma distribucemi, pozorovanou a očekávanou. To znamená, že v sémantické struktuře textu jsou pro její utváření významné maximální hodnoty konstituentů, čili maximální hodnoty velikosti textových segmentů pro jednotlivé lexikální jednotky. Obecně řečeno, veličina $w_i(f_i)$ roste jak při vstupu nové lexikální jednotky do textu, tak i při zvýšení frekvence lexikální jednotky, která již v textu existuje. Jestliže se zvětší max $w_i(f_i)$, znamená to, že nějaký rostoucí segment je vyplňován buď novou lexikální jednotkou, nebo se zvýšila frekvence jednotky, která se již v textu vyskytuje. Zdá se, že tuto vlastnost nenajdeme na jiných jazykových úrovních ve vztahu konstruktů a konstituentů než právě na úrovni sémantické.

Výhledy

Určité estetické uspokojení z vlnící se sémantické struktury je narušováno fluktuacemi pozorovaných hodnot kolem hodnot ideálních, tj. hodnot očekávaných (vypočítaných). Místy bychom raději viděli hodnoty jiné nebo někdy i žádné. V některých případech totiž nemá smysl uvažovat pozorované hodnoty, když průměrná velikost

konstruktu, odpovídající určité frekvenci, je dána třeba jen jedinou hodnotou, a proto je náchylná k odchylce. Takovou situaci matematická statistika umožňuje korektně řešit pomocí některého z testů extrémních odchylek.

Tento přístup uplatňujeme na té úrovni poznání sémantické struktury textu, na níž se nacházíme dnes, když stále ještě cítíme potřebu přesvědčovat se o tom, že zákonitý vztah mezi jazykovým konstruktem a jeho konstituenty na sémantické úrovni v textech skutečně existuje. Proto je nezbytné být při statistických experimentech naprosto korektní. Přesto se lze domnívat, že i na této úrovni poznání dostal pojem významového konstruktu lexikálních jednotek poněkud určitější podobu.

Při dalším výzkumu ovšem bude nezbytné soustředit se na jednotlivé pozorované odchylky od ideálních křivek a hledat příčiny, proč mají určitý tvar. V některých případech, zejména u uměleckých textů, půjde o záměr produktora textu zesílit jeho estetické působení. Výmluvným příkladem je poezie segmentovaná do stejně dlouhých nebo přibližně stejně dlouhých veršů, kdy jsou významové efekty segmentace dosahovány a zesilovány jinými prostředky. Jindy umělecká próza zařazuje do sekvence segmentů velmi rozsáhlé výčty (jak je známe např. z děl Karla Čapka nebo jak jsme je pozorovali v našem souboru u některých tureckých textů). Při zvukové reprodukci textů může být použito segmentace, která neodpovídá syntaktickému členění. Jinak ovšem větná syntax se zdá být tím nejpřirozenějším základem segmentace textů.

Odchylky od ideální struktury mohou být rovněž způsobeny patologickými příčinami, např. vadami mozkových funkcí. Fakt, že stejnou nebo podobnou slapovou strukturu nacházíme v textech místně, jazykově nebo stylisticky vzájemně velmi vzdálených vede k představě, že táž struktura a fungování mozku se prosazuje do textů bez ohledu na velké rozdíly v jazykových kódech. Pak lze očekávat, že vady a odchylky v mozkových strukturách a funkcích se budou prosazovat do jazykových projevů jako charakteristické odchylky od nějaké standardní sémantické struktury, které mají např. medicínský význam. Je možné, že tato okolnost postaví lingvistiku před nové úkoly.

LITERATURA

ALTMANN, G. (1980): Prolegomena to Menzerath's Law. *Glottometrika*, 2, s. 1–10.

ALTMANN, G. – SCHWIBBE, M. H., et al. (1989): *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim – Zürich – New York: Olms.

HŘEBÍČEK, L. (1989): The Menzerath-Altmann law on the semantic level. *Glottometrika*, 11, s. 47–56.

HŘEBÍČEK, L. (1992): *Text in Communication: Supra-Sentence Structures.* Bochum: Brockmeyer.

HŘEBÍČEK, L. (1997): *Lectures on Text Theory.* Prague: Oriental Institute.

HŘEBÍČEK, L. (2002): *Vyprávění o lingvistických experimentech s textem.* Praha: Academia.

HŘEBÍČEK, L. (2005): Text laws. In: R. Köhler – G. Altmann – R. G. Piotrowski (eds.), *Quantitative Linguistik: Ein internationales Handbuch / Quantitative Linguistics: An International Handbook.* Berlin – New York: Walter de Gruyter, s. 348–361.

WIMMER, G. – ALTMANN, G. – HŘEBÍČEK, L. – ONDREJOVIČ, S. – WIMMEROVÁ, S. (2003): *Úvod do analýzy textov.* Bratislava: Veda.

ZIEGLER, A. – ALTMANN, G. (2002): *Denotative Textanalyse: Ein textlinguistisches Arbeitsbuch.* Wien: Edition Praesens.

SUMMARY

Semantic cataracts in text structures

The search for text structures is an important subject for text linguistics. Recently, this problem appeared to be solvable with the help of Menzerath-Altmann's law. The formal structure of this law is expressed by formula (1). Its basic notions are *language construct* and its size (x), *language constituent* expressed by a mean-size value (y) and their relationship (given by parameters A and b). This is a general law valid for all language levels, including the semantic text level. In this way, any text (in a natural language with an uninterrupted sequence of sentences) actually becomes something like a language unit.

The standard manner of detection of this law at the semantic level is documented in Table 1, where z is the number of lexical units, each of which is a base for semantic constructs of a text. Their sizes equal the number of text segments, each containing a given lexical unit; the (usually approximate) identity of the semantic construct's size with frequency, i.e. $x \equiv f$, is thus defined. These segments are constituents of a construct and their average size is expressed in number of words.

Another manner of treating this structure considers each lexical unit i occurring in a text. The *semantic* or *contextual weight* of a unit is defined in formula (2), where S_i is the sum of segment sizes of a given i, and f_i is word frequency in the text. If the values of w_i are treated in relation to f_i as its function, i.e. as $w_i(f_i)$, and the Zipfian arrangement of word frequencies is applied to f_i and w_i, an L-formed curve for f_i together with a "cataract curve" for $w_i(f_i)$ is obtained, see Figure 1 concerning the same Czech text as in Table 1.

This typical arrangement of lexical units was observed in different texts and different languages, as is documented by the Turkish text in Figure 2. Both texts, Czech and Turkish, are items from literature (a short story and a chapter of a novel), but similar structures can be found in different kinds of texts.

Each "wave" of the semantic cataract represents a distribution of values that can be characterized by three variables: max $w_i(f_i)$, a mean value $\langle w_i(f_i)\rangle$ and min $w_i(f_i)$ – or simply max w_i, $\langle w_i\rangle$, min w_i. Each wave, however, seems to hang on the first of these variables. Its relevance for the whole structure is proven by the fact that the points of max w_i (or, simply, the peaks of the waves) are situated on a Menzerathian curve. This curve corresponds to the transformed formula (1), where y is substituted by max w_i. It is evident that this new curve better captures the basic idea of Menzerath-Altmann's law: the greater the construct, the smaller the constituent.

Any text is a complex system, therefore the descriptive variables and curves experience the interference of random fluctuations. However, it is rational to expect that in the future some typical deviations from the ideal forms of the characteristic curves will be found and combined with pathological digressions proper to certain brain functions.

Orientální ústav AV ČR, v. v. i.
Pod vodárenskou věží 4, 182 08 Praha 8
<hrebicek@orient.cas.cz>