KYBERNETIKA ČÍSLO 1, ROČNÍK 3/1967

O možnostech matematického modelování sémantiky přirozeného jazyka*

NIKOLAJ SAVICKÝ

1. Abychom mohli popsat jazykovou strukturu jako souhru výrazu a významu, musíme nejprve nezávisle na sobě popsat „výrazovou substanci" (v tomto případě distribuci) a „významovou substanci" (tj. vyjadřované pojmy).

2. Veškeré dosavadní popisy struktury byť i jediného slova, včetně popisů metodologicky i materiálově nejnáročnějších, jsou neúplné. Myslím, že jde o zásadní neúplnost formalizovaného popisu přirozeného jazyka, analogickou zásadní neúplnosti formalizace matematiky, a vyplývající ze stejných příčin — infinitního charakteru objektu.

* * *

Zpracování překladových pomůcek, k nimž patří nejen algoritmy strojového překladu, nýbrž např. i slovníky běžného typu, je jednou z nejdůležitějších aplikací lingvistiky. Předpokladem pro zlepšení těchto pomůcek je systematizace popisu lexikální sémantiky jazyka (viz L. Kopeckij [7], O. Leška [8], U. Weinreich [15]). Tento úkol je zařazen do plánu Ústavu jazyků a literatur ČSAV.

Existuje mnoho návrhů, jak provést tuto systematizaci. Pokusím se 1. navrhnout metodu, jak ověřit výsledky některých těchto postupů vzájemným porovnáním, 2. identifikovat obtíže, které při sémantickém popisu vznikají.

1. Za východisko můžeme vzít např. citovaný článek Weinreichův [15]. Autor v něm požaduje mj., aby se u slova přísně rozlišovaly významy kontrastní, které se mohou projevovat ve stejném kontextu, a významy komplementární, které jsou zcela závislé na kontextu.

Abychom toto rozlišování mohli systematicky a ověřitelným způsobem provádět, musíme umět nezávisle na sobě popsat typy kontextů a významové jednotky, aniž předem činíme nějaké předpoklady o tom, jaký je způsob přiřazení mezi významem a kontextem.

Metodou popisu kontextů je distribuční a transformační analýza. Její aplikaci na lexikální materiál rozpracovala nejpodrobněji skupina sovětských lingvistů rozvíjejících tzv. strukturální lexikologii (Ju. D. Apresjan [2], [3], [4], [5], M. P. Muravickaja [10], V. I. Perebejnosová [11], V. M. Rusanivskij [12]). Jejich práce jsou velmi cenné zejména tím, že své postupy aplikují na rozsáhlém materiálu. Ověřování, zda stanovené jednotky a třídy jsou sémanticky významné, je však u nich zcela intuitivní.

* Předneseno na *druhé konferenci o kybernetice,* která se konala v Praze ve dnech 16. — 19. listopadu 1965.

Zjišťování jednotek tzv. signifikativního významu [3], nezávislé na konkrétním jazyce, může probíhat v podstatě dvojím způsobem:

a) pomocí překladu (N. D. Andrejev a S. Ja. Fitialov [1], K. Čulík [6], P. Sgall [13]; tento způsob je universální (zkušenost ukazuje, že přeložit se dá téměř vše), ale ne příliš přesný, neboť výsledek podstatně záleží např. na tom, jakou množinu jazyků uvažujeme;

b) pomocí logické analýzy vyjadřovaných pojmů nebo tzv. rozkladu na sémantické komponenty, což je způsob velmi blízký (viz přehled v [3]). Musíme zde rovněž přihlížet k pojmovým soustavám nejrůznějších věd (botaniky, zoologie, anatomie, chemie atd.). Tento způsob je nejpřesnější z možných, ale je neúplný; např. tzv. intenzionální vztahy mezi výroky (vyjadřované slovesy sentiendi a dicendi apod.) působí zatím při logické analýze velké obtíže.

Zdá se, že z charakteristiky těchto dvou způsobů popisu významu vyplývá, že se mohou navzájem dobře doplňovat.

Vzájemný vztah distribučních a významových jednotek vyplyne ze způsobu vzájemného přiřazování; výsledek přiřazení, tj. lexikální jednotku, která je vymezena jak distribučně, tak i sémanticky, můžeme nazvat lexémem. Přesněji řečeno, *lexém* slova *a* je jednotka v rámci slova *a* (podmnožina množiny všech výskytů slova *a*), charakterizovaná jistým souborem relevantních rysů; za *relevantní rys lexému* považujeme distribuční nebo transformační vlastnost výskytů slova takovou, že této vlastnosti lze přiřadit sémantickou interpretaci („sémantický komponent"). Tím se poněkud odchylujeme od užívání termínu „lexém" u Apresjana [2], u něhož označuje jednotku čistě distribučně-transformační. Jinak řečeno, na rozdíl od Apresjana, považuji signifikativní význam za součást významu strukturního, a nikoliv za jev paralelní.

2. Zavedení pojmu lexému umožní přesněji ukázat, jakým způsobem se projevují obtíže sémantického popisu. Ukazuje se totiž, že každé zjemnění použitých kritérií a každé rozšíření zkoumaného materiálu vede k dalšímu štěpení stanovených jednotek (např. lexémů v rámci jednoho slova). Nedospíváme ke stabilním jednotkám.

Toto neomezené štěpení zná z vlastní praxe každý lexikograf. Málokdy se mu podaří dosáhnout pocitu, že popsal všechny významy zkoumaného slova. Vezměme však zvlášť nápadný příklad: M. P. Muravickaja [10] zpracovala podle Apresjanových postupů 400 výskytů ruského slovesa „нести" a rozdělila je na 32 třídy („значимости"). Avšak už poměrně nevelký slovník, kterého nebylo využito při zpracování (Russisch-deutsches Wörterbuch, red. Lochowiz a Leping, Leipzig 1953) uvádí mj. konstrukci „из-под пола несет", pro kterou Muravickaja nemá vůbec distribuční schéma, a „курица несет яйца", která nezapadá beze zbytku do žádné ze stanovených tříd. Rovněž práce V. I. Perebojnosové o anglickém slovesu „make" [11] je prokazatelně neúplná, čehož je si autorka vědoma.

Bylo by možno statisticky zkoumat, zda při rozšiřování zkoumaného korpusu se blížíme ke stanovení uzavřené množiny lexémů v rámci nějakého zadaného slova. Předpokládám, že výsledek by byl negativní, a to z těchto důvodů:

Na rozdíl od situace v popisu zvukové stránky jazyka, kde se podařilo překonat štěpení fonetických jednotek rozlišením relevantních rysů od nerelevantních, nemůžeme žádný rozdíl významu nebo distribuce považovat s klidným svědomím za nerelevantní; v nejbližším nově zkoumaném textu může být aktualizován, učiněn relevantním. Tuto možnost aktualizace lze ověřit takto: jakmile se pomocí překladu nebo logické analýzy přesvědčíme, že nějaké slovo má několik významů, které nejsou v dosavadních slovnících rozlišovány, můžeme, aniž porušíme gramatiku jazyka, utvořit takovou „transformaci", která tento rozdíl vystihne (např. „A je B" může mít různé významy; můžeme je rozlišit takto: „A je totéž co B", „A je jedním z B" apod.).

Jinak řečeno, vyslovuji hypotézu, že *inventář relevantních rysů lexémů, tj. distribučních nebo transformačních vlastností, kterým odpovídá nějaký významový rozdíl, je potenciálně nekonečný*, čili, což je totéž, že každý konečný seznam těchto relevantních rysů je neúplný a lze ho doplnit.

V tom je další rozdíl proti koncepci Apresjanově, který na několika místech explicitně formuluje opačný názor a tvrdí, že budeme-li zkoumat jen ty sémantické rozdíly, které jsou distribučně nebo transformačně vyjádřeny, pak dospějeme ke konečné množině.

Ve světle toho, co bylo uvedeno o vztahu významu a výrazu ve struktuře lexému, vidíme, že konečnost množiny relevantních rysů lexémů by mohla být prokázána trojím způsobem:

a) buď přímo udáním konečného a úplného seznamu relevantních rysů lexémů;

b) nebo udáním konečného a úplného seznamu výrazových, tj. distribučních vlastností, schopných diferencovat lexémy (o tom, že ke splnění tohoto požadavku máme daleko, viz Melčuk J. A. [9], str. 15—16);

c) nebo konečně udáním konečného a úplného seznamu elementárních jednotek signifikativního významu („sémantických komponent"). Vzhledem k tomu, že jazyk je schopen vyjádřit celou matematiku, zdá se, že splnění tohoto požadavku je znemožněno Gödelovou větou o neúplnosti formalizované aritmetiky (vždyť axiomy implicitně určují obsah nedefinovaných pojmů dané teorie, a tudíž každý nový nezávislý axiom tak či onak pozměňuje soubor nedefinovaných, elementárních pojmů). Srov. též [14], str. 42 ruského překladu.

Vzhledem k tomu, že jde o množiny prvků dále nerozložitelných, nejsou-li tyto množiny konečné, nemohou být ani rekursivně spočetné.

Zásadní neúplnost sémantického modelu neznamená, že bychom se měli vzdát jeho budování. Spíše naopak. Vzhledem k tomu, že množina uvažovaných prvků ve fonologickém modelu je konečná, je tento model z matematického hlediska málo zajímavý. Teprve na sémantické úrovni se blížíme k tomu, abychom postihli skutečnou matematickou strukturu jazyka, která nemůže být triviální.

Vztah mezi fonologií a strukturní sémantikou se zdá být z tohoto hlediska obdobný vztahu mezi Booleovou algebrou (algebrou konečné množiny) a aritmetikou přirozených čísel. V sémantice, stejně jako v aritmetice, místo jednoho definitivního formalismu musíme počítat s nekončící posloupností formalismů stále mocnějších.

Pro praktickou práci z toho vyplývá požadavek vědomě upustit od snahy po nedosažitelné úplnosti a nahradit ji snahou po správném výběru prvků, které je nutno zahrnout do popisu. Jde o to, že jednotlivé lexémy nejsou rovnocenné co do svého významu v textu. Frekvenční výzkumy nám umožní stanovit jejich relativní důležitost. V jistém smyslu dospíváme k rehabilitaci zásady „čím větší slovník, tím více významů".

(Došlo dne 19. listopadu 1965.)

LITERATURA

[1] Андреев Н. Д., Фитиалов С. Я.: Язык-посредник машинного перевода и принципы его построения. Тезисы совещания по математической лингвистике. Ленинград 1959, 53—60.

[2] Апресян Ю. Д.: О понятиях и методах структурной лексикологии. Проблемы структурной лингвистики. Изд. АН СССР, Москва 1962, 141—161.

[3] Апресян Ю. Д.: Современные методы изучения значений и некоторые проблемы структурной лингвистики. Проблемы структурной лингвистики. Изд. АН СССР, Москва 1963, 102—148.

[4] Апресян Ю. Д.: О сильном и слабом управлении. Вопросы языкознания *XIII* (1964), 3, 32—49.

[5] Апресян Ю. Д.: Опыт описания значений глаголов по их синтаксическим признакам. Вопросы языкознания *XIV* (1965), 5, 51—66.

[6] Čulík K.: Některé problémy teorie jazyků. Kybernetika a její využití. Nakl. ČSAV, Praha 1965, 276—290.

[7] Kopeckij L. V.: Aktuální otázky dvojjazyčného slovníku. O vědeckém poznání soudobých jazyků. Nakl. ČSAV, Praha 1958, 191—194.
[8] Leška O.: Systém v lexiku. Československá rusistika *VIII* (1963), 2, 64—67.
[9] Мельчук И. А.: Автоматический синтаксический анализ. Изд. Сибирского отделения АН СССР, Новосибирск 1964.
[10] Муравицкая М. П.: Некоторые вопросы полисемии. Изд. Киевского унив., Киев 1964.
[11] Перебейнос В. И.: К вопросу об использовании структурных методов в лексикологии. Проблемы структурной лингвистики. Изд. АН СССР, Москва 1962, 163—173.
[12] Русанівський В. М.: Спроба визначення семантичних груп дієслів на основі формальних критеріїв. Структурно-математична лінгвістика. Киев 1965, 56—64.
[13] Sgall P.: Perspektivy matematické a aplikované lingvistiky. Kybernetika a její využití. Nakl. ČSAV, Praha 1965, 263—275.
[14] Taube M.: Computers and Common Sense. 1961; rus. překl., Москва 1964.
[15] Weinreich U.: recense na Webster's Third New International Dictionary. International Journal of American Linguistics *XXX*, (1964) 4, rus. překl. Вопросы языкознания *XIV* (1965), 1, 128—132.

SUMMARY

On Possibilities of Matematical Modelling of Natural Language Semantics

NIKOLAJ SAVICKÝ

1. If we want to describe the structure of a language as the interplay of expression and content, we must first describe independently of one another „the expression substance" (in this case the distribution) and „content substance" (in this case the expressed notions).

2. All existing descriptions of the structure even of a single word, including descriptions most advanced methodologically, are incomplete. The author considers this incompleteness to bear an intrinsic and insuperable character. The incompleteness of the formalized description of natural language is probably analogous to the incompleteness of the formalized arithmetic, both proceeding from the same causes — the infinite character of the described object.

Nikolaj Savický, Ústav jazyků a literatur ČSAV, Thunovská 22, Praha - Malá Strana.