SBORNÍK PRACÍ FILOZOFICKÉ FAKULTY BRNĚNSKÉ UNIVERZITY
STUDIA MINORA FACULTATIS PHILOSOPHICAE UNIVERSITATIS BRUNENSIS
A 36, 1988

KAREL PALA

SYNTAKTICKÁ ANALÝZA A PÁDOVÉ RÁMCE

1. ÚVOD

V tomto příspěvku je předmětem naší pozornosti automatická syntaktická analýza českých vět a problematika pádových rámců. V první části příspěvku si povšimneme syntaktického analyzátoru ANAL1 (Pala, 1985), v druhé části se budeme věnovat jeho rozšířené a sémanticky orientované verzi — ANAL2, v němž je uplatněna myšlenka pádových rámců a pádové gramatiky.

2. CHARAKTERISTIKA ANALYZÁTORU ANAL1

Syntaktický analyzátor ANAL1 je implementován v systému TPT/WANDER (Benešovský, Šmídek, 1984) na počítači PDP 11/34 v ÚVT UJEP v Brně. Je interaktivní, není příliš velký, zabírá asi 100 KB paměti (vnitřní i disky) a umožňuje grafické zobrazení výsledků syntaktické analýzy českých vět — syntaktických grafů-stromů — na terminálu VT 100.

Je schopen pracovat s vymezenou podmnožinou českých vět tvořenou jednoduchými českými větami obsahujícími libovolně složité substantivní skupiny v přímých i předložkových pádech, adverbiální a adjektivní skupiny v přímých i předložkových pádech, adverbiální a adjektivní skupiny a skupiny slovesné — maximálně třísložkové. Pokud jde o souvětí, ANAL1 si poradí se souvětími obsahujícími spojky *že*, *aby* a se vztažnými zájmeny *který*, *kdo*, *co*, připravena je analýza souvětí s dalšími spojovacími výrazy *jak*, *až*, *když*, *protože*, *jestliže*, *pak*, *a*, *ale*, *nebo*. Testy byly prováděny na souboru čítajícím 150 vět s úspěšností kolem 90 %. Získávané stromové struktury vět jsou víceznačné, lze získávat všechny možné analýzy, i když jsme se zatím spokojili se získáním jedné analýzy.

Souhrnně je ANAL1 implementován jako soubor sémantických (TPT—) sítí: 1) VOCAB.WAN, 2) CZECH.WAN, 3) ANAL.WAN a 4) WORK.WAN. Stručně řečeno, síť VOCAB.WAN obsahuje slovník asi 600 českých slovních tvarů, síť CZECH.WAN obsahuje hierarchii lingvistických pojmů potřebných

pro syntaktickou analýzu a vztahů mezi nimi — obě tyto sítě jsou datové. Síť ANAL.WAN představuje vlastní program analýzy, vlastní analyzátor s procedurální nekontextovou gramatikou češtiny tvořenou asi 70 pravidly a síť WORK.WAN je pracovní — do ní se v průběhu analýzy vstupní věty ukládají dílčí výsledky, tj. podstromy jednotlivých větných složek a v případě úspěchu i výsledná stromová struktura celé vstupní české věty.

Povšimneme si nyní poněkud podrobněji jednotlivých uvedených sítí. První z nich je VOCAB.WAN — slovník českých slovních tvarů. V rámci ANAL1 se nevyplatí provádět morfologickou analýzu, rozsah použitého slovníku je příliš malý, i když se počítá s jeho interaktivním doplňováním. U každého slovního tvaru je ve slovníku uveden jeho slovní druh a všechny jeho gramatické kategorie, které může mít. Počítá se i uvedením sémantických rysů různého druhu, např. ZIV (životnost), NEZ (neživotnost) a dalších, i když v této fázi práce se jich nevyužívalo.

Druhou datovou sítí je CZECH.WAN, která je tzv. typovou sítí a obsahuje hierarchii pojmů potřebných pro danou aplikační oblast — tedy v našem případě — pojmů lingvistických. Jinak řečeno, síť CZECH.WAN obsahuje formalizovanou taxonomii slovních druhů, syntaktických složek a gramatických kategorií potřebných pro popis českých syntaktických struktur. Tato taxonomie je úplná a je v průběhu analýzy využívána vlastním analyzátorem (tj. sítí ANAL.WAN) ke kontrole správného užití lingvistických pojmů. U slovních druhů je použita klasická klasifikace (viz např. Havránek, Jedlička, 1960) doplněná ovšem o dosti podrobné subklasifikace (zejména u sloves). Syntaktické (větné) složky jsou vzhledem k použití složkové syntaktické koncepce vymezeny čistě formálně. Jejich větněčlenské funkce, pokud jsou získávány, jsou vyhodnocovány na základě pádů nebo, jak dále ukážeme, pádových rámců. Pracujeme tedy se složkami substantivními (přímými a předložkovými, zájmennými), slovesnými a adverbiálními a adjektivními. Pokud jde o gramatické kategorie, pracujeme se všemi obvyklými jmennými a slovesnými kategoriemi, s obvyklým rozlišením oznamovacích, rozkazovacích a tázacích vět a s potřebnými sémantickými rysy, které se v síti CZECH.WAN mohou vyskytovat v podobě atributů.

Síť ANAL.WAN, tj. vlastní analyzátor a program analýzy, je napsán v jazyce WANDER, jenž je programovacím jazykem vyšší úrovně pro manipulaci se sítěmi a grafovými strukturami. Základními částmi analyzátoru jsou 4 velké skupiny nekontextových pravidel:

a) soubor pravidel pro rozpoznávání substantivních skupin,
b) soubor pravidel pro rozpoznávání slovesných skupin,
c) soubor pravidel rozpoznávajících adverbiální a adjektivní skupiny,
d) soubor pravidel pro skládání jednotlivých větných složek do celé věty.

I když rámec použitých pravidel je nekontextový, díky jejich formulaci v podobě procedur získáváme možnosti pracovat s kontextovými závislostmi, a to až v rozsahu celé věty. Tato vlastnost pravidel-procedur se výrazně uplatňuje zejména u slovesných skupin, které mají často nespojitý charakter, např. ve spojeních typu: „*On by včera večer napsal ten program...*". Dále mají pravidla-procedury tu vlastnost, že samy dovedou budovat podstrom složky, kterou právě analyzují.

K síti ANAL.WAN patří i dvě relativně samostatné procedury SHODA a PÁDY, jejichž úkolem je testovat gramatickou shodu v rodě, čísle a pádě

uvnitř substantivních skupin a dále mezi subjektovou substantivní skupinou a predikátovou slovesnou skupinou (v čísle, osobě a jmenném rodě). Tím tyto procedury jednak kontrolují syntaktickou správnost složek uvnitř analyzované věty a jednak představují jistou techniku zpětného prohledávání (backtrackingu), když omezují falešné kroky při možném chybném spojení složek.

Dodejme ještě, že celková strategie analýzy použitá v ANAL1 je shora dolů, v určitých bodech analýzy se však uplatňuje strategie zdola nahoru, která celkově činí analýzu efektivnější.

K síti WORK.WAN jen poznamenáme, že s ní vlastně komunikuje uživatel — lingvista, když si nechává na obrazovce terminálu zobrazovat výsledky analýzy — grafy stromy vstupních českých vět. Vlastní grafické zobrazení zajišťuje program PEPNET, jehož autorem je J. Gerbrich.

3. ANAL2 A PÁDOVÉ RÁMCE

Zkušenosti s analyzátorem ANAL1 a úvahy kolem jeho praktické použitelnosti spolu s rozbory úspěšnosti jednotlivých typů analyzátorů publikovanými v literatuře (viz např. Frederking, 1985) vedou k pokusům o smíšené analyzátory. Analyzátory tohoto typu na počátku analýzy vstupní věty využívají syntaktických způsobů rozpoznávání slovních druhů a složek, které nesou nebo by mohly nést význam, v klíčových bodech se však obracejí k sémanticky orientovaným pravidlům, která určují, jak dokončit analýzu věty, nebo které analýzy jsou v případě víceznačnosti — sémanticky interpretovatelné. Zdá se tedy, že sémanticky orientované analyzátory si mohou snáze poradit s některými typickými jevy v přirozených jazycích, např. s víceznačností lexikální i konstrukční, s elipsami a anaforickou referencí, s porovnáními a koordinací.

Z lingvistického hlediska je jasné, že základní relační a také formálně nejsnáze identifikovatelnou složkou české věty je určitý tvar slovesný. Navíc každé sloveso se v závislosti na svém významu pojí s určitými participanty, aktanty, rolemi či hloubkovými sémantickými pády (viz např. Fillmore, 1968). Slovesa se podle participantů, s nimž se pojí, seskupují do významových tříd, takže kombinace sloves s typickými participanty lze popsat pomocí formálních pravidel, jimž se podle Fillmora říká pádové rámce.

Soubor takových formálních pravidel tvoří pak pádovou gramatiku, která představuje sémantický popis určitých typů vět (podle toho, které třídy sloves jsou do ní pojaty) a jíž může být s výhodou použito jako jádra syntaktického analyzátoru.

V této souvislosti není od věci poznamenat, že v oblasti umělé inteligence se nezávisle objevila myšlenka rámců — jejím autorem je M. Minsky (1975). Podobnost rámců a pádových rámců byla velmi brzy rozpoznána a došla značného rozšíření jak v pracích orientovaných na porozumění přirozenému jazyku, tak v pracích týkajících se rozpoznávání obrazců, scény a analýzy vidění (Winston, 1977).

Pádové rámce popisují význam slovesa uvedením jeho typických participantů, tj. chápeme-li sloveso jako relační prvek (predikát), pak sémantická povaha participantů představuje zároveň idiosynkratickou charakteristiku jednotlivých participantů — argumentů. Dále je známo, že pro jednotlivé

participanty existují i typické způsoby jejich syntaktické, formální realizace v povrchové struktuře věty. Údaje tohoto druhu lze přičlenit k pádovým rámcům a s výhodou jich využívat v syntaktickém analyzátoru obsahujícím pádovou gramatiku. Např. pro participanty typu agens je typická a téměř závazná formální realizace v podobě nominativní substantivní skupiny, podobně narazíme-li v české větě na substantivní skupinu v přímém akuzativu, jde s největší pravděpodobností o formální realizaci participanta typu objektivu. Stejně tak participant typu adresát má nejtypičtější realizaci v podobě dativní nepředložkové substantivní skupiny a tak bychom mohli pokračovat dále.

Budujeme-li nějaký konkrétní fragment pádové gramatiky (např. pro češtinu), musíme se vyrovnat s několika problémy:

a) je potřeba rozhodnout, jak bude vypadat soubor vhodných participantů, s nimiž budeme pracovat. V literatuře lze najít dvě protikladná řešení této otázky, např. Fillmore (1968) pracuje s malým počtem hloubkových pádů — kolem 10, jiní (Sgall a kol., 1986) pracují až s 30. Popis s větším počtem participantů je nepochybně přesnější, i když může být méně obecný, a ne tak elegantní, nicméně se zdá, že pro konkrétní aplikaci bude nejvhodnější jistý kompromis počítající zhruba asi s 20 participanty.

b) je potřeba rozlišit (i) obligatorní (vnitřní) participanty, (ii) fakultativní participanty, (iii) volné modifikátory chápané často jako implicitní. Pro naše účely pokládáme za vhodné následující členění: (i) obligatorní:
agens,
objektiv,
adresát,
patiens.

V případě potřeby zavedeme jejich vhodné varianty, např. u sloves označujících zpracování informace bude užitečné pracovat s nositelem informace, příjemcem informace apod. (ii) fakultativní:
benefaktiv/or
instrument,
původ,
sociativ,
kauzativ.
(iii) volné modifikátory:
způsob (vlastní, zřetel, míra),
lokativ,
temporativ,
komparace.
I zde počítáme s doplněním dalších modifikátorů, jestliže si to daná aplikační oblast vyžádá.

Uvědomujeme si, že výběr inventáře participantů je do jisté míry arbitrární a že pro podobné třídy sloves lze mít i dosti různé inventáře participantů. Je to dáno jednak použitím různých teoretických východisek a jednak tematickým zaměřením dané problémové oblasti a pohledem na ni.

(iv) Obtížným problémem je rozhodování o aritě jednotlivých sloves, tj. rozhodování o počtu argumentů-participantů, s nimiž se jednotlivá slovesa mohou pojit. Někteří autoři soudí (Sgall, Materna, 1983), že u jednoho slovesa lze rozlišit predikáty s různou aritou, např.:

(1) Pavel píše dopis perem.

(2) Pavel píše Evě.

Věty (1) a (2) by také měly být popsány různými pádovými rámci. Pro praktickou aplikaci v syntaktické analýze se však jeví užitečnějším pracovat u slovesa *psát* s jedním pádovým rámcem odpovídajícím větě:

(3) Pavel píše dopis Evě perem.

Některé participanty lze pak chápat jako implicitní (defaults) a v konkrétních větách se přirozeně nemusí vyskytovat, v odpovídajícím pádovém rámci jsou však uvedeny a mohou pro ně dokonce být zavedena sémantická inferenční pravidla podobná sémantickým postulátům.

Dále v hrubých rysech naznačíme strukturu analyzátoru ANAL2, jehož ústředním prvkem — jádrem je právě soubor pádových rámců — pádová gramatika.

Počáteční kroky jsou u ANAL2 stejné jako u ANAL1 — přečte se vstupní věta a provede se lexikální analýza všech slov a ke každému slovu se přiřadí ze slovníku jeho slovní druh a příslušné gramatické kategorie. Existuje zde i možnost interaktivního doplňování slov do slovníku s případným vygenerováním všech potřebných tvarů.

Ve vstupní větě se najde slovesný tvar a vytvoří se celá slovesná skupina. Aktivuje se procedura „hlídač rámců", která si najde pádový rámec analyzovaného slovesa a začne řídit průběh analýzy celé věty. Hlídač rámců se zejména snaží zjistit, která substantivní skupina odpovídá kterému participantu z pádového rámce. Analýza jednotlivých substantivních, adjektivních a adverbiálních skupin přitom probíhá s použitím pravidel-procedur definovaných v ANAL1.

Jestliže „hlídač rámců" najde pro daný pádový rámec formální realizace všech participantů a ve větě již nezbylo žádné neanalyzované slovo, analýza úspěšně končí a vytvoří se vhodná hloubková reprezentace vstupní věty. Pokud ve větě zbyly nějaké složky, je potřeba rozhodnout, zda jsou to podsložky již analyzovaných participantů nebo zda jde o volné modifikátory, případně i o implicitní participanty. Právě rozhodování tohoto typu jsou velmi obtížná, jsou spojena s konstrukční víceznačností a vyžadují chytře organizované zpětné prohledávání, abychom získávali ne všechny možné analýzy vstupní věty, nýbrž jen ty, které připouštějí rozumné sémantické interpretace.

Zvlášť bude „hlídač rámců" pracovat se slovesem *být*, neboť to se pojí s více pádovými rámci a je tudíž víceznačné. Samostatně je také potřeba pracovat s případy, kdy jednotlivým participantům odpovídají vedlejší věty. Zde se jeví jako výhodné použít heuristická pravidla spojená přímo s některými konkrétními českými slovy, např. s jednotlivými spojkami a odkazovacími výrazy. Podobného druhu jsou i pravidla (použitá již v ANAL1) vztahující se k hranicím mezi jednotlivými větnými složkami — tyto hranice poskytují velmi užitečné informace pro rozhodování o dalším průběhu analýzy.

Slovník pro ANAL2 bude odlišný od slovníku ANAL1 v tom, že bude mít dvě části: první obsahuje všechna slova kromě sloves a podle potřeby bude doplněn o vhodné sémantické rysy. Jeho samostatnou podčástí bude slovník idiomů, který je nutný pro analýzu idiomatických spojení dodávajících vstupům větší přirozenost.

Druhá část slovníku: slovník sloves + pádové rámce — představuje tedy vlastní pádovou gramatiku.

Závěrem lze říci, že ANAL2 představuje pokus rozšířit ANAL1 ve směru

přirozenějších vstupů a větší robustnosti. Použití pádových rámců poskytuje také dobré východisko k definitivní sémantické analýze v termínech konstrukcí, jak jsou definovány v transparentní intenzionální logice (Tichý, 1976).

LITERATURA

BENEŠOVSKÝ, M.—ŠMÍDEK, M.: Testování programů. In: Sborník SOFSEM '84, VVS Bratislava, 1984, s. 34—36.
FILLMORE, Ch.: The case for case, In: Universals in Linguistic Theory, ed. by E. Bach and R. Harms, Holt, Rinehart and Winston, 1968.
FREDERKING, R. E.: Syntax and semantics in natural language parsers, výzk. zpráva Dept. of Computer Science, Carnegie-Mellon University, May 1985.
HAVRÁNEK, B.—JEDLIČKA, A.: Česká mluvnice. Praha, SPN, 1960.
MATERNA, P.—SGALL, P.: Optional participants in a semantic interpretation (arity of predicates and case frames of verbs). Prague Bulletin of Mathematical Linguistics, *39*, 1983, s. 27—37.
MINSKY, M.: A framework for representing knowledge. In: Psychology of Computer Vision, ed. by P. Winston, Mc-Graw Hill 1975.
Pala, K.: Syntaktický analyzátor pro češtinu. SPFFBU, *A 33*, 1985, v tisku.
SGALL, P.—HAJIČOVÁ, E.—PANEVOVÁ, J.: The meaning of the sentence and its semantic and pragmatic aspects. Praha, Academia 1986, s. 198—199.
TICHÝ, P.: Intensional logic. Manuscript. University of Otago, 1976, 730 s.
WINSTON, P.: Artificial intelligence. Addison-Wesley, Reading-London, 1975.

SYNTACTIC ANALYSIS AND CASE FRAMES

In this paper we briefly describe the interactive natural language parser for Czech-ANAL1 and its semantically based extension — ANAL2.

The first parser ANAL1 has been implemented on the PDP 11/34 as a semantic net consisting of four nets:

1. VOCAB.WAN—a dictionary of Czech word forms (about 600 items),
2. Type net CZECH.WAN containing the hierarchy of the linguistic notions that appear to be necessary for the description and parsing of the basic Czech syntactic structures,
3. ANAL.WAN—the parser itself and its program written as a set of the context-free rule based procedures in the higher programming language WANDER,
4. Temporary net WORK.WAN in which the intermediate results of the analysis are stored and from which they can be drawn and displayed on the screen of the terminal.

ANAL2 is a semantically based extension of the ANAL1 and it is built on the idea of the case frames of verbs and case grammar. The suggested case frame grammar is, of course, a fragment designed for a specialized problem domain, particularly, for programming. It includes about 150 most frequent verbs and their case frames. The set of deep cases includes about 20 cases and we distinguish the obligatory ones, the optional ones and so called free modifiers. The core of the parser ANAL2 is a "frame keeper"—the procedure that is able to control the whole course of the parsing and that tries to solve the ambiguous points in the analysis.