БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ ACADÉMIE BULGARE DES SCIENCES
ИЗВЕСТИЯ НА МАТЕМАТИЧЕСКИЯ ИНСТИТУТ
BULLETIN DE L'INSTITUT DE MATHÉMATIQUES
Том (Tome) XIII



# КЪМ ВЪПРОСА ЗА „НУЛЕВИТЕ ЕЛЕМЕНТИ" ПРИ МАШИННИЯ ПРЕВОД И АВТОМАТИЧНАТА ОБРАБОТКА НА ЕСТЕСТВЕНИТЕ ЕЗИЦИ

Александър Людсканов

Общоизвестно е, че при превода (обикновен (ОП), т. е. извършван от човек преводач, или машинен (МП), т. е. извършван самостоятелно от автоматична сметачна машина (АСМ)) от един естествен език $L_i^N$ (входния) на друг естествен език $L_j^N$ (изходния) в редица случаи в изходното съобщение се появяват елементи, чиито праобрази не са явно зададени на графемичното или фонетичното равнище на входното съобщение. Ето защо, когато сравняваме на това равнище (графемично или фонетично; по-нататък ще видим, че същото се наблюдава и при сравняване на кои да е две еднакви равнища) въпросните две съобщения — входното (оригинала) с изходното (превеждащото), — ние интуитивно долавяме, че първото е „некомплектно" по отношение на второто, че в него има „празнини", отговарящи на „новопоявилите" се елементи във второто. Тези празнини ще наричаме *нулеви елементи* $\phi$ и ще се постараем на първо място да заменим тази интуитивна представа с едно логически обосновано понятие.

Въпреки че проблемата за нулевите елементи в даден естествен език (тук под език се разбира крайно множество от текстове, формулирани на този език) изпъква най-ясно при превод от този език на някакъв друг език, тя досега, доколкото ни е известно, не е била предмет на обобщени изследвания нито в трудовете, посветени на общата теория на превода (вж. например [1]—[6]), нито в онези работи, които си поставят специалната цел да установят, класифицират и евентуално да моделират и формализират процедурите по установяването на образи (преводни съответствия) (вж. например [7]). По същия начин в теоретическите работи, посветени на МП и изобщо на машинната обработка на информация, зададена във форма на естествените езици, тази проблема не е поставяна и разглеждана в обща форма и досега е била емпирично и частично решавана само в някои конкретни преводни алгоритми [8], [72]. Но проблемата за нулевите елементи не е само специална „преводаческа" проблема — в съответен аспект тя се поставя и при сравнително-типологическите лингвистически изследвания, и при изследванията, посветени на езика еталон [15], и при обсъждането на проблемата за така наречената

междуезикова интерференция и особено при разработването на конкретни модели, описващи различни езикови процедури.

Както ще проличи от по-нататъшното изложение, проблемата за нулевите елементи се свързва в последна сметка с начините на представянето на смисъла от средствата, принадлежащи на различни равнища на съответния език, и следователно конвергира с някои от основните проблеми на съвременната лингвистика проблемите на лингвистическата семантика (слаба[1]) и с проблемите, свързани с обособяването на отделните равнища на езика, които напоследък привличат вниманието на все повече изследователи (срв. например докладите, изнесени на Международната конференция по семиотика, състояла се през септември 1966 г. в Полша, и на Симпозиума по равнищата на езика, състоял се през април 1967 г. в Москва[2]). Всичко това наред с подчертаната на последните международни срещи по МП и математическа лингвистика (МЛ) императивна необходимост от задълбочаване на лингвистическите изследвания, по-широки обобщения и логически анализ на приеманите по интуиция понятия [16], [68], [69] налага да се разгледат „празнините", които условно нарекохме нулеви елементи, да се обобщят те под едно суперординирано общо понятие и да се анализират някои от неговите основни свойства (§ 1), а, от друга страна, първо, да се изложат и обобщят познатите процедури за третиране на нулевите елементи и, второ, да се види какви по-мощни възможности, допускащи формализиране[3] и следователно алгоритмизиране, разкриват в тази област най-новите насоки, по които вървят днес изследователите в областта на структурното езикознание и математическото моделиране на естествените езици, и да се предложи съответна процедура (§ 2). Това ще позволи да се установят общите проблеми, които се поставят при обработването на нулевите елементи при МП и изобщо при автоматичното третиране на естествените езици (§ 3), а така също да се набележи решението на някои от тях в светлината на предложената от автора концепция за оптимална селективна стратегия при МП (§ 4).

---

[1] Някои езиковеди обособяват с л а б а семантика, която се занимава с проблемите на еднозначността и тъждествеността на значението на езиковите изрази, и с и л н а семантика, която се занимава предимно с проблемите на истинността и съотношението на езиковия израз и неговия денотат [11].

[2] Интересно е да се напомни, че докато преди Втората световна в йна и непосредствено след нея „главна" тема на лингвистиката беше фонологията и донякъде морфологията, към края на четиридесетте години успоредно с първите стъпки по пътя на създаването на теорията на формалните граматики основният интерес на лингвистите се пренесе в областта на синтактичните структури, а днес прави още една крачка напред и все повече се премества в областта на семантиката (вж. например [14]) и на моделирането на механизма на свързването на звука и значението. Това преместване на центъра на тежестта на лингвистическите изследвания (което конвергира с основните тенденции на развитие в областта на МП — срв. § 3) върви успоредно с поставянето под съмнение на някои положения на структурализма, като например тезиса за линейността на обозначаващото в устната реч, за пълната немотивираност на езиковия знак (срв. например изследванията на Якобсон за иконическите елементи в естествените езици), менталистките основи на някои концепции на Чомски за разлика от антименталиста на Блумфилд (вж. например [19] и др.)

[3] Тук и по-нататък (ако не е уговорено нещо друго) под формализиране ще разбираме трансформиране на съдържателните лингвистически анализиращи и синтезиращи операции във формални. Под съдържателни лингвистически операции ще разбираме тези, които се основават върху означаваните страни на езиковите знаци и тяхната предметна съотнесеност, а под формални — онези, които се основават върху означаващата страна на езиковите знаци и тяхната синтагматика (по-подробно вж. [13]).

§ 1. Тъй като онова, което нарекохме нулев елемент, изпъква най-ясно при съпоставянето на един естествен език $L_i^N$ с друг $L_j^N$, отначало ще разглеждаме интересуващата ни проблема и ще определим съответното понятие при превода от един естествен език на друг[4] (1.2) и след това ще разглеждаме някои от неговите общи свойства (1.3). Но преди това е необходимо да напомним нашето разбиране на някои основни положения из областта на общата теория на превода (или по-точно на теорията на междусемиотичните трансформации [29]), от които ще излизаме в по-нататъшното изложение (1.1).[5]

**1.1.** Въз основа на мотивираната в [18], [5], [28]—[30] семиотична концепция под п р е в о д ще разбираме преобразуване на символите на входното съобщение в символи от друг код със запазване (дотолкова, доколкото това е възможно поради наличността на „шумове") на информация, инвариантна спрямо дадена система за отчитане. В последна сметка процесът на превода се свежда към подбиране на образи (съответствия) на езиковите средства (в най-широкия смисъл на думата — вж. [17]) на входния текст по такъв начин, че да се получи посоченият по-горе резултат. Като използваме една бинарна релация [25] между средствата на отделните равнища на два езика, под о б р а з (с ъ о т в е т с т в и е) ще разбираме онова езиково средство (или съвкупност от средства или нулево средство), което дава в превода същата информация, както и неговия п р а о б р а з (р е ф е р е н т). Между средството на оригинала — праобраза, — принадлежащо на $L_i^N$, и неговото преводно съответствие — образа — в даден $L_j^N$ (или в самия него — така наречената абсолютна синонимия) при съответни условия могат да се установят обратими релации: ако $B$ е образ (преводно съответствие) на $A$, то $A$ е образ (преводно съответствие) на $B$. При това, както вече знаем, възможни са и такива ситуации, при които на праобраз, принадлежащ на дадено равнище на $L_i^N$, отговаря „празнина", „нулево" съответствие на същото равнище на $L_j^N$. Установяването и класифицирането на типовете образи (съответствия) при превода от $L_i^N$ на $L_j^N$, както и установяването на обратни релации, споменати по-горе, предполага точно описание на всички фактори на еквивалентност и нееквивалентност (по равнища, функции на езика, конотативна натовареност и пр.; вж. например [20], [26]). Само при тази предпоставка, а тя неизбежно трябва да е налице при всички системи на МП, проблемата на превода може да стане дедуктивна проблема. Тъй като образите (съответствията) носят същата информация или, с други думи, възпроизвеждат функционалността на техните праобрази в оригинала, ще ги наричаме ф у н к ц и о н а л н и о б р а з и [27].

Подбирането на образи в процеса на превода от един естествен език на друг обикновено предполага поради спецификата на тези езици (омонимия, полисемия, синонимия, непълна конвенционалност, полиструктурност, известна субективност при използване на конотативните значе-

[4] По-нататък ще видим, че същото се наблюдава както при съотнасянето с която и да било друга семиотична система извън приетия входен език, така и с по-дълбинните равнища на същия този език.

[5] Нашето изложение ще се ограничи с писмената форма на естествените езици и предимно с онези аспекти на проблемата, които се обуславят от нуждите на МП и на автоматичната обработка на тези езици.

ния и пр.) повече информация от онази, която носи самото превеждано средство, т. е. праобразът. Минималното, което преводачът (човек или машина) трябва да „знае“ или да е в състояние да „научи“ за дадено езиково средство, за да го преведе при дадена езикова ситуация от един конкретен език на друг конкретен език, ще наричаме *необходима преводна информация* (ITN) за това средство (вж. [28]). Съвкупността от операциите над текста, въз основа на които се набира ITN, ще наричаме *анализ* (който може да се разглежда и като превод — или по-точно като съвкупност от преводи — от графемичното равнище на входното съобщение, кодирано във формата на даден $L_i^N$, на равнището на смисъла [29]), а съвкупността от операциите, въз основа на които от ITN се генерират образите (преводните съответствия) и следователно изходният текст — *синтез* (който може да се разглежда и като превод — или по-точно съвкупност от преводи — от равнището на смисъла на равнището на неговата морфографемична интерпретация в $L_j^N$ [29]).[6]

**1.2.** Както следва от изложеното, всяко средство, използвано в превода, е образ на нещо, т. е. на някакво средство от входния текст, което наричаме *праобраз* или *референт*.

**1.2.0.** За разлика от това „нормално“ положение на нещата конструктивните примери, които привеждаме по-долу, съдържат и такива образи, които на пръв поглед като че ли нямат референти във входните съобщения (в случаите, в които не се сочи противното, под входен текст или съобщение ще разбираме изолирано изречение):[7]

(1) Этот наш друг очень способный студент. (1′) This friend of ours is a very clever student.

(2) Собаки, которые лают, не кусают. (2′) Les chiens qui aboient ne mordent pas.

(3) Петр был в кино. (3′) Пьотър е бил на кино.

(4) Ich habe einen Brief geschrieben. (4′) Аз написах едно писмо.

В изходните (превеждащи) изречения (1′)—(4′) образите (съответствията), които нямат референти, зададени на същото равнище на входните изречения, или чиито референти са непълни, са: *is* в (1′), *les* в (2′), *е бил* в (3′) и *написах* в (4′).

При това тук не поставяме въпроса, откъде преводачът е набрал ITN, за да използва тъкмо определения, а не неопределения артикъл в (2′), преизказаното наклонение в (3′) и свършения вид в (4′), а просто

---

[6] Между другото такова разглеждане на процесите на анализа и на синтеза при превода потвърждава, от една страна, изразеното в [29] становище, че преводът като междуезикова и вътрешноезикова трансформация трябва да се разглежда като едно от средищните понятия не само на семиотиката, но и на езикознанието, а, от друга — мотивираното в [30] схващане за еднотипността на процесите на декодирането и разбирането на едно езиково съобщение, както и на кодирането, синтезирането и генерирането. Това между другото, разбира се на различни равнища на абстракция, се доближава до тезата, че една формална граматика трябва да бъде неутрална спрямо говорещия и слушащия, т. е. за еквивалентността на разпознаващото и генериращото описание.

[7] Това ограничение води до игнориране на възможните влияния на хиперсинтаксиса и хиперсемантиката, но се налага от обстоятелството, че засега не разполагаме с модели, позволяващи анализ и извън рамките на работното изречение, въпреки наличието на интересни изследвания в тази насока.

приемаме, че тези преводи са верни. Не е мъчно да се види, че тези примери илюстрират различни проявления на общото явление, което нарекохме нулев елемент.

В (1′) спомагателният английски глагол *is*, изразяващ предикативността в това изречение, е образ на съществуваща в руския граматически строй, но в дадения тип конструкция експлицитно не задавана на графемичното равнище категория. В (2′) определеният артикъл е образ не само на незададен в случая на графемичното равнище на входния текст праобраз, но и на такава семантична категория (определеност), която по начало няма експлицитен изразител в граматичния строй на руския език. В (3′) българският образ *е бил* е не само изразител на предикативност и минало време, в каквото значение му отговаря руският праобраз *был*, но е и експлицитен изразител на семантичната категория „несвидетелство", която няма експлицитен изразител в руския граматичен строй. По същия начин в (4′) българският глагол *написах* е не само пряк лексически образ на немския глагол *geschrieben*, но и експлицитен изразител на категорията вид, която не съществува в граматичния строй на немския език.

**1.2.1.** Това твърде обичайно в преводаческата практика положение на нещата може да се представи опростено в обобщен и схематизиран вид така. Нека са ни зададени две изречения, т. е. редици $Z_1$ и $Z_2$, от терминални символи, принадлежащи съответно на $L_i$ и $L_j$.[8]

$$Z_1 = \# \; a \;\; b \qquad c \;\; d \qquad e \;\; f \;\; g \qquad h \;\; \#,$$
$$Z_2 = \# \; a_1 \;\; b_1 \;\; A_1 \;\; c_1 \;\; d_1 \;\; B_1 \;\; e_1 \;\; f_1 \;\; g_1 \;\; C_1 \;\; h_1 \;\; \#.$$

Нека приемем, че $Z_2$ е превод на $Z_1$ в $L_j$ и че $Z_1$ е превод на $Z_2$ в $L_i$, т. е. че $Z_2$ е образ на $Z_1$ в $L_j$ и обратно (вж. например [31]). Тогава, ако разглеждаме $Z_2$, съпоставяйки го с неговия образ в $L_i$ (т. е. със $Z_1$), ще можем да кажем, че $A_1$, $B_1$ и $C_1$ са експлицитно зададени на графемичното равнище на $L_j$ образи, които на равнището на текста имат празни праобрази в $Z_1$. И обратно — ако разглеждаме $Z_1$, като го съпоставяме с неговото отражение в $L_j$ (т. е. със $Z_2$), ще можем да кажем, че в него има нулеви елементи, които имат експлицитно зададени на това равнище образи в $L_j$ (т. е. в $Z_2$). Това може да се представи така:

$$Z_1 = \# \; a \;\; b \;\; \left|\begin{matrix}\not{\varnothing}^{1} \\ A_1\end{matrix}\right| \;\; c \;\; d \;\; \left|\begin{matrix}\not{\varnothing}^{2} \\ B_1\end{matrix}\right| \;\; e \;\; f \;\; g \;\; \left|\begin{matrix}\not{\varnothing}^{3} \\ C_1\end{matrix}\right| \;\; h \;\; \#,$$
$$Z_2 = \# \; a_1 \;\; b_1 \;\; c_1 \;\; d_1 \;\; e_1 \;\; f_1 \;\; g_1 \;\; h_1 \;\; \#.$$

Не е мъчно да се види, че за нулев елемент говорим при наличността на двойки[9] от вида $\left|\begin{matrix}\varnothing \\ X\end{matrix}\right|$; при това от съдържателно и конструктивно гледище елементите на тези двойки могат да принадлежат както на лексическото, така и на морфологическото, синтактичното и семантичното равнище.

**1.2.2.** Въз основа на изложеното, както и на някои съображения които се обсъждат в т. 2.2 на § 2, може да се предложи следната де

---

[8] По традиция, въведена от Чомски, символът # се използва за означаване на начало и край на изречение.

[9] Двойки от този вид и в такава идеална подреденост се получават при такова схематично представяне. При интерпретация със средствата на конкретни естествени езици такова идеално подреждане се среща рядко, но може да се получи чрез съответна реконструкция.

финиция: нулев елемент $\emptyset$ на дадено равнище $l_1$ на даден $L_i^N$ е експлицитно незададен образ (съответствие), чийто праобраз (референт) е експлицитно зададен или на същото равнище в $L_x$, или на по-дълбинно равнище на $L_i^N$.[10]

**1.3.** След като определихме понятието нулев елемент, ще разгледаме някои от неговите общи свойства с оглед на по-нататъшното изложение.

**1.3.0.** За нулев елемент може да се говори само в конкретен съпоставителен план. Там, където ще има нулев елемент в $L_i^N$ при съпоставяне с $L_j^N$, може да се окаже наличен елемент при съпоставянето му с $L_s^N$ (така, ако в примера (4) бихме съпоставяли не с български език, в който има граматическа категория вид, а например с френския, в който тази категория отсъствува подобно на немския, не бихме имали нулев елемент). Следователно наличността или отсъствието на нулеви елементи в даден език зависи не толкова от неговите собствени свойства, колкото от свойствата на системата за съотнасяне, с която съпоставяме[11] (същото се наблюдава например и при установяването на полисемията на езиковите средства на даден език [32]). При това, както следва от нашата дефиниция, като система за съотнасяне при установяване на $\emptyset$ може да се използва не само някакъв друг естествен език, но и всяка семиотична система. Очевидно е, че най-пълна възможност за анализ в тази насока би дало съпоставянето например с езика на така наречената универсална семантика (вж. например [33]) или със записа на смисъла в езика на предикатното смятане (вж. например [34]), или с така наречения основен език (basic-language) на даден естествен език (вж. например [32]), или с така наречените дълбинни структури.

**1.3.1.** По начало в резултат на осъществяването на процеса на превода трябва да се получи текст, който да дава инвариантна (абстрахирайки се от шумовете) информация в сравнение с онази, която дава входният текст. Следователно в превода нито трябва да се появява „допълнителна“ или „нова“, нито пък да се „губи“ някаква информация. Но и в четирите разгледани превода (1)—(4) на пръв поглед, т. е. на графе-

---

[10] Смятам, че не е необходимо да се дават отделни примери, за да се покаже, че референтът на $\emptyset$ може да е зададен и на по-дълбинно равнище на същия $L_i^N$, тъй като това ясно проличава при анализа на някои разпознаващи и пораждащи модели, който се провежда в § 2. Противното би довело до ненужни повторения. Трябва също да се отбележи и следното. Това предварително определение е дедуцирано въз основа на анализ на структурата и лексическата интерпретация само на изолирани изречения. Такъв подход, както бе отбелязано, се дължи на обстоятелството, че засега не сме в състояние да моделираме процеса на широкия (т. е. извън работното изречение) контекстуален анализ. Ако се премахне това ограничение, в определението би трябвало да се отрази и ситуацията, при която референтът на $\emptyset$ е зададен на същото равнище на същия език, но в друго изречение.

[11] С оглед на това първата задача при едно инвентаризиране на случаите на нулевия елемент в даден $L_i^N$ е да се подбере в зависимост от целите на изследването системата за съотнасяне, спрямо която ще се установяват тези нулеви елементи. При това едно такова инвентаризиране, необходимо както за чисто теоретически лингвистически проучвания, така и за редица приложни задачи (например обикновен и машинен превод, програмирано езиково обучение, автоматично индексиране, създаване на транслатори и др.), е възможно само на квантитативна статистико-вероятностна база според типовете на синтактичните структури, тъй като в противен случай то би предполагало анализирането на всички отделни двойки изречения в съответните два езика (вж. например [36]).

мично равнище, като че ли се появява „нова“ информация. Следователно или тези преводи са неточни, нееквивалентни, или онази информация, която в сравнение с равнището на графемичната форма на входните текстове изглежда нова, допълнителна, в същност не е нова, а е зададена по някакъв начин в тези входни текстове. След като сме постулирали, че разглежданите преводи са верни, остава втората алтернатива и се поставя следният въпрос: в превода по начало не може да се появи нищо ново, нищо незададено във входния текст; щом като това е така, трябва да се установи къде и как е зададена във входните текстове онази информация, която се изразява от „новите“, експлицитно зададени образи в изходните текстове, на които отговарят съответните нулеви елементи във входните текстове. Тук са възможни две хипотези: тази информация е зададена някъде в линейния контекст, извън анализираното изречение, или на по-дълбинни равнища на самото анализирано изречение. Тъй като боравим с изолирани изречения, ще разгледаме само втората алтернатива.

**1.3.2.** В обобщен и схематизиран вид тази втора алтернатива може да се формулира така. Ние приехме, че изречението $Z_2$ е превод, отражение на изречението $Z_1$ в $L_j^N$ и, обратно, че $Z_1$ е отражение, превод на $Z_2$ в $L_i^N$ (да отбележим, че при изследвания от този тип винаги се получава магьосан кръг, който, както ще видим, се разрешава на по-дълбинни равнища — срв. също [31]). При този начин на представяне може да се върви, така да се каже, в две посоки: или а) да се излиза от $Z_2$ и да се върви към $Z_1$, или б) обратно. Нека приемем (както това е и в нашите примери (1)—(4)), че в $Z_2$, т. е. в превеждащото изречение, има елемент $A_1'$, който е експлицитно зададен образ (съответствие) на някакъв нулев елемент $\not{o}^1$ от $Z_1$. Сега да се опитаме да анализираме тези два елемента $A_1'$ и $\not{o}^1$), като вървим по двата посочени пътя: от $Z_2$ (т. е. $L_j^N$) към $Z_1$ (т. е. $L_i^N$) и обратно.

а) Да приемем първо, че входното изречение е $Z_2$, например

$$A_1'$$

# This friend of ours is a very clever student #,

а превеждащото, изходното изречение $Z_1$

# Этот наш друг $\not{o}^1$ очень способный студент #,

в което нулевият елемент е празен образ (преводно съответствие) на $A_1'$ от $Z_2$ (т. е. на *is*). Тъй като вървим от $Z_2$ към $Z_1$ и осъществяваме операцията $A_1' \rightarrow \not{o}^1$, не е мъчно да се установи, че нулевият елемент $\not{o}^1$ от $Z_1$, който е образ на $A_1'$ от $Z_2$, носи същата информация като него. Това положение не се нуждае от коментар. Но при обратната хипотеза ще имаме принципно различно положение.

б) Ако вървим от $Z_1$ (което сега приемаме за входен текст)

# Этот наш друг $\not{o}^1$ очень способный студент #

към $Z_2$

$$A_1'$$

# This friend of ours is a very clever student #,

за да може да използваме в него $A_1'$ (т. е. *is*) като експлицитно зададен образ, ние трябва да научим отнякъде, първо, че в $Z_1$ има нулев елемент $\not{o}^1$, и, второ, каква информация носи той, за да може да реализираме $\not{o}^1 \rightarrow A_1'$. Но тъй като $\not{o}^1$ е нулев елемент и сам по себе си не носи никаква определена информация, то за да му подберем образа (преводното

съответствие) $A_1'$, трябва да намерим някаква друга система за отнасяне на същото изречение, или друг език, или друго равнище, в които да съществува експлицитно зададено средство, което да носи онази информация, която имплицитно съдържа нулевия елемент и която експлицитно трябва да изрази $A_1'$.

С други думи, за да реализираме $\not\phi^1 \rightarrow A_1'$, трябва предварително да намерим елемент (или съвкупност от елементи) $A$ от някаква друга система за отчитане, който да е референт (праобраз) на нулевия елемент.

Тъкмо това общо свойство на нулевите елементи в изложеното погоре разбиране ни позволява да направим следния съществен извод: в последна сметка както при теоретическите лингвистически изследвания, така и при алгоритмизирането на процеса на превода и автоматичното третиране на естествените езици проблемата на обработването на нулевите елементи в текстове, формулирани на даден естествен език, се свежда на първо място към идентифицирането на техните референти (праобрази).[12] От това, между другото, следва едно положение, което не бива да се изпуска из очи: разликата между един „наличен" и един „нулев" елемент при превод от един естествен език на друг се заключава не в това, че единият е зададен експлицитно, а другият — не, а в това, че референтът на първия е зададен на същото равнище във входния текст, а референтът на втория не е зададен на това равнище и трябва да бъде открит чрез съответна процедура за идентифициране. На някои основни проблеми, свързани с тези процедури, е посветен следващият параграф.

**§ 2.** Щом като основната проблема при обработването на нулевите елементи се свежда към идентифицирането на техните референти, първо, трябва да разгледаме съществуващите в тази насока процедури (2.2) и, второ, да видим какви възможности се откриват в тази насока от най-новите постижения на структурната и математическата лингвистика в областта на синтактичните (2.3) и семантичните (2.4) модели и въз основа на това да се опитаме да предложим една процедура, която ще наричаме „дълбинна". Но преди това трябва да изложим някои основни положения на съвременната лингвистика, от които ще излизаме (2.1).

**2.1.** Всеки естествен човешки език предполага една система (langue). намираща се в главата на човека, единият от механизмите на която свързва смисъла (значението) със звука, т. е. поражда речта (parole), Между значението и звука съществува определена мрежа от отношения (тъкмо тази мрежа от отношения се означава обикновено с термина структура на езика). Тази мрежа „от дълбинни отношения в езика, свързана със законите на пораждането на езикови единици от всички рангове от най-простите първични елементи на езика" [15, с. 15], е предмет на структурната лингвистика.[13]

---

[12] Трябва да напомним тук едно съществено обстоятелство, което следва от нашата семиотично-функционална концепция на превода: всяко преводно съответствие (образ) носи същата информация, както и съответният референт (праобраз); този референт може да бъде изразен със средство (или средства) било от същото равнище на езика, било от други равнища, било на същото, било на друго място в последователността на текста, било в същото или в друго изречение.

[13] Във връзка с това не може да не се напомни пределно ясното изложение на тези положения, което дава Л. Йелмслев [40, с. 419—420], като резюмира мислите на Ф. де Сосюр: Напротив, реалните езикови единици съвсем не са нито звуците или писмените

Продуцираната от езика реч (parole) е достъпна за непосредствено наблюдение и с оглед на това обикновено се описва с таксономични (класификационни) подходи и вероятностни математически методи. Структурата на езика, механизмът на пораждането, както и самият език (langue) са недостъпни за пряко наблюдение, поради което тяхното изследване предполага евристично моделиране (т. е. по начало създаване на математически модели — вж. например [46]), позволяващо да се възпроизвеждат (и обясняват благодарение на съответните конструктивни описания) повече факти от онези, които са били наблюдавани при съставянето на модела.[14] Такова моделиране, при което моделът обект [12] се описва с невероятностни структури, може да се прилага при изследването на всички аспекти на езика (фонетика, граматика, лексикология и семантика).

Комуникативната функция на езика се реализира чрез произвеждане и разпознаване на елементарни езикови съобщения, за каквито ще смятаме изреченията.[15] В резултат на функционирането на редица конкуриращи механизми на езика носителят на даден език обективира породеното от него (тук е по-правилно да се говори за произвеждане) съобщение (т. е. изречението) от ляво на дясно при писмената реч и от „началото" към „края" при устната реч в съответна временна последователност. При това всеки последващ елемент се фиксира в неговата окончателна морфофонемична или морфографемична форма и не може да бъде изведен от предишния. Така думите *малкото момче* (от изречението # малкото момче чете нова книга #) не могат на това равнище т. е. на равнището на речта (parole), да бъдат изведени нито от някакви по-общи единици (защото ги няма), нито една от друга. Както е известно от съответната теоретична литература, такова пораждане от ляво на дясно на равнището на речта представлява марковски процес, който може да бъде моделиран от така наречения автомат с краен брой състояния или някое друго еквивалентно устройство. Обаче лингвистическата интерпретация на такъв модел предполага и отделни, конкретни, лексически правила за всеки елемент при всяка стъпка на пораждането, което е равносилно на задаване на всички лексически интерпретации на всички изречения от даден $L_i^N$. А това е невъзможно, тъй като броят на

---

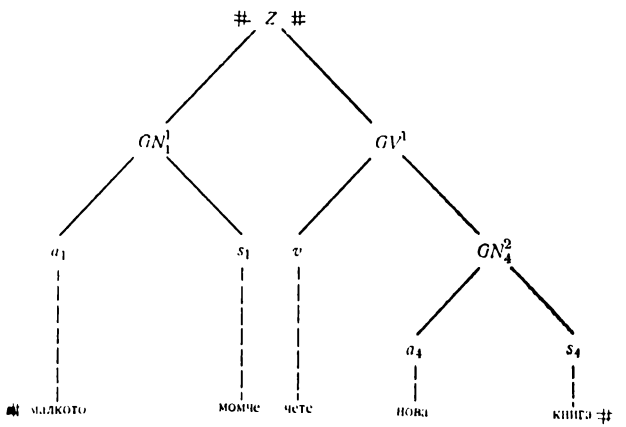
знаци, нито значенията (тъкмо в това е основната разлика с възгледите на традиционната лингвистика, която смята за свой предмет именно звуците и значенията — А. Л.); реалните езикови единици са представените от звуците или знаците и от значенията съотношения... Тъкмо тези съотношения са системата на езика и именно тази вътрешна система е характерна за даден език и го различава от другите езици, докато реализирането на езика в звуци или писмени знаци или в значения е безразлично за самата система на езика... Следователно именно скелетът (т. е. структурата, казано със съвременни термини — А. Л.) трябва да бъде главният предмет на езикознанието, докато конкретното реализиране и манифестиране на скелета на съотношенията са безразлични за определяне на езика в строгия смисъл на тази дума."

[14] Така, като цитира думите на Куайн, Чомски бележи в [38, с. 417]: „Лингвистическата теория дава общо обяснение на онова, което „трябва да бъде" в езика въз основа на това, „което е", плюс простотата на законите, посредством които ние описваме и екстраполираме това, което е."

[15] Тук няма да се занимаваме с проблемите за разликата между пораждане и произвеждане, разпознаване и разбиране, с въпросите за неутралността на генеративното и разпознаващото описание спрямо говорещия и слушащия, с разграничението на областите на компетенцията и перформацията, както и с обсъждането на проблемата за приоритета и оптималността на генеративното или разпознаващото описание [38], [21], [24], [43], [58], [59], [63], [65], [66] и др.

такива интерпретации е огромен. Но дори това и да би било възможно, такова моделиране води до тривиални в научно отношение резултати.

С оглед на това, вместо да се представи като процес на генериране на готови елементи от ляво на дясно само на едно равнище, процесът на пораждането на изречението може да се представи като дедуктивен процес на конструиране на сложни граматически обекти от прости, вървящ от „дълбочинна“ към „повърхностна“ и минаващ последователно през няколко равнища, на всяко от които въз основа на съответни правила в експлицитна форма се разгръщат елементите, имплицитно зададени на по-дълбинно равнище. Така пораждането на взетото по-горе за пример изречение може да се представи в опростен вид в термините на една така наречена безконтекстна граматика по непосредствено съставящи (НС-граматика) по начин, даден на фиг. 1.




Фиг. 1

При такова описание („еднофакторни граматики“) обикновено се обособява пораждането на повърхностната структура, т. е. преходът от началния абстрактен символ $Z$ (изречение) към редицата на терминалните символи (в нашия пример веригата $\# a_1\ s_1\ v\ a_4\ s_4 \#$, където $a$ е прилагателно, $s$ — съществително, $v$ — глагол; индексите означават падежни еквиваленти) и морфофонетичната (лексическата) интерпретация на тази верига (т. е. реалното изречение) на равнището на речта. При това не бива да се забравя, че съответните дълбинни равнища са само конструктивни начини на представяне на процеса на пораждането.

Модели, позволяващи да се опишат процеси от този тип, се наричат понякога формални граматики. Тук ние няма, а и не сме в състояние да обсъждаме същността и предимствата на отделни видове формални граматики (граматики с краен брой състояния, НС-граматики тип CF и CS, трансформационни, предсказващи, апликативни, категориални и стратификационни граматики, граматики на зависимостите, графови и диспозиционни граматики и др.) и ще отбележим само едно основно тяхно общо свойство, което има съществено значение за всичките ни по-нататъшни разсъждения — техния аксиоматично-дедуктивен характер. Благодарение на това тяхно общо свойство във всички подобни граматики, както във

всяка дедуктивна система, в извода (т. е. веригата, последователността от символите на всяко равнище) не може да се появи нищо, което в имплицитна форма да не се съдържа в правилата за извод и в единиците от по-дълбинно равнище и в последна сметка в аксиомите на съответния дедуктивен модел.

Като имаме пред вид тези най-общи положения на съвременната лингвистика, както и някои изводи, които могат да се направят от изследваното множество от текстове и експериментите, проведени от групата по машинен превод и математическа лингвистика при Математическия институт с Изчислителен център на БАН, ще изложим първо същността на съществуващите начини за идентифициране на референтите на нулевите елементи на равнището на речта, а след това ще предложим съответна процедура, основаваща се върху анализ на описанието на подълбинните равнища.

**2.2.** Първата от тези две процедури включва два подхода за установяване на референтите на нулевите елементи — *повърхностно-лингвистичен* (2.2.0) и *извънлингвистичен* (2.2.1).

**2.2.0.** При повърхностно-лингвистичния подход необходимата преводна информация (ITN) за установяване на референта на нулевия елемент (независимо от това, дали става дума за една морфологическа, синтактична, лексическа или стилистическа категория) се набира от тъй наречения лингвистически контекст[16], т. е. от зададените средства на равнището на входния текст както в рамките на анализираното изречение, така и извън тях. Така например, ако входният текст е руското изречение # сын царя Ивана Грозного #, при превод, да кажем, на френски или на български ще се постави въпросът, дали трябва да се използва определеният или неопределеният артикъл като еквивалент на съответния нулев елемент в руското изречение. В рамките само на този входен текст въпросът е неразрешим и неговото решение предполага съобразяване с извънлингвистични данни (срв. по-долу). Но ако бихме имали например и контекста „... *родившийся в ... году*“, то този контекст би дал необходимата информация, за да се установи, че в случая става дума за точно определено лице, което в езика се изразява чрез категорията „определеност“, нямаща формално представяне в граматичния строй на руския език и изразяваща се обикновено с лексически средства. Тази информация би била достатъчна, за да се идентифицира референтът на съответния нулев елемент, а следователно и за да се установи „наличното“ преводно съответствие на $\alpha^1$ в $L_j^N$, т. е. $A_1'$ (във френски превод — артикълът *le*).

Този начин на идентифициране на референтите на нулевите елементи чрез анализ на лингвистическия контекст на равнището на текста е твърде добре познат в преводаческата практика и е бил използван (наистина само в рамките на работното изречение) и в теорията и практиката на машинния превод при създаването на преводни алгоритми, изградени върху лексико-морфологическа база (така наречената H-операция; вж. [28]). Както ще видим, той може (поне в определени случаи) да се използва

[16] Тук терминът *контекст* се използва в смисъла на традиционните работи по обща теория на превода, а не в смисъл на съвкупност от условия, които позволяват да се преведе едно езиково средство (следователно и извънлингвистични).

с успех и при създаване на алгоритми за МП, основани върху синтактико-сематични концепции.

**2.2.1.** Но въпреки неговата относителна (поне от гледището на човека преводач) простота и удобство този подход за идентифициране на референтите на $\phi$ на равнището на текста, при който анализът върви, така да се каже, по повърхността, става безрезултатен, когато лингвистическият контекст на това равнище не дава възможност да се извлече ITN. В тези случаи, за да може да се осъществи процесът на превода, се налага „извънлингвистичен" анализ и съпоставяне с действителността[17] [18], което поне засега (т. е. докато тази действителност не е описана по съответен начин и въведена в паметта на машината) изключва алгоритмизацията на тази част от процеса на превода.

Както следва от изложеното, от двата подхода, характерни за традиционната процедура за установяване на референтите на $\phi$ и за генериране на техните образи (преводни съответствия), при днешното равнище на описанието на семантиката и „способността" на машината да „съпоставя" с действителността може да се алгоритмизира както за целите на МП, така и изобщо за целите на автоматичната обработка на информация, зададена във формата на естествените езици, само повърхностно-лингвистическият подход в рамките на едно изречение. Поради това в редица случаи автоматичното идентифициране на референта на $\phi$ остава невъзможно. С оглед на това трябва да се види дали тази процедура, насочена по „повърхността" на текста и „извън" езика, не може да се замени с процедура, насочена „навътре" в езика, при което първо ще разгледаме тази проблема на синтактично равнище.

**2.3.** Предлагането на процедура за идентифициране на референтите на нулевите елементи, която да се основава не върху анализ на контекста, а върху анализ в дълбочина, навлизайки в по-дълбинни структури, вече не на равнището на речта, а на равнището на езика, стана възможно благодарение на представянето на това равнище чрез моделите на структурното и математическото езикознание.

**2.3.0.** Към изложеното в т. 2.1 на този параграф трябва да се добави следното. С оглед на това, че според основния тезис на структурната лингвистика манифестирането на езика в звуци и значения е ирелевантно за анализа на неговата структура, описанието на ненаблюдаемата част на езика започна с описанието, моделирането на процесите на пораждането и разпознаването на абстрактни синтактични структури. Първите пораждащи модели (граматики с краен брой състояния, граматики по непосредствено съставящи —безконтекстни и контекстни— и трансформационни граматики), разработени от Чомски, бяха построени върху предложеното от него разбиране на граматиката ($G$) на даден език $L_i^N$ като съвкупност от правила, рекурсивно пораждащи (или разпознаващи) всички белязани изречения от дадения $L_i^N$ и само тях и приписващи им деривации, т. е. структурни описания (структури на извода), непротиворечащи

[17] В приведения по-горе пример с руското изречение # сын царя Ивана Грозного # ITN за това, дали в един френски превод би трябвало да се използва определеният (*le*) или неопределеният (*un*) артикъл, би могла да се извлече от исторически данни: ако Иван Грозни е имал само един син, то това обстоятелство е достатъчно основание за използване на определения артикъл, в противен случай и този извънлингвистичен анализ не би довел до натрупване на ITN.

на нашето интуитивно разбиране [38]. Създаването (до към 1960—1962 г.) на този тип граматики, които са опит да се даде положителен отговор и на въпроса, дали граматическите категории могат да се описват, без да се прибягва до значението,[18] отбеляза съществена крачка напред по пътя на моделирането на механизма на езика, въпреки че техните автори съзнателно се абстрахираха от проблемите на семантичната интерпретация на пораждане (разпознаване) от съответните модели синтактични структури.

В последна сметка всички граматики от този вид, както и граматиките на валентностите („зависимостите"), разработени в Ленинградската група по МЛ и МП (вж. например [31]), си поставят за цел да моделират механизма, пътя, по който дълбинните (базисни) синтактични структури последователно се трансформират в повърхностни синтактични структури, в изводи (белязани изречения), представляващи редици от абстрактни терминални символи от дадена „азбука". В процеса на това пораждане (или разпознаване) тези граматики приписват на всяко едно изречение така наречените структурни описания (или структури на извода), съществената част от които се свежда към представянето[19] на реализирането на извода (т. е. на деривацията) чрез така наречения структурно дърво. Всяка съвкупност от върхове на това дърво (или един връх), намиращи се на еднакво разстояние (дълбочина) от „главния" връх, е отделно равнище на структурата на извода. По същия начин една разпознаваща граматика от този тип моделира последователните преминавания от равнището на терминалните символи или словесните класове към базисните синтактични структури.

Като имаме пред вид свойствата на разгледаните типове граматики, нашата задача е да видим дали този начин на моделиране и представяне на механизма на разпознаването и на пораждането дава възможност и да се предложи процедура за идентифициране на референтите на $\varnothing$ на по-дълбинните равнища на езика.

**2.3.1.** С оглед на изложеното по-горе, както и на извода, направен в края на т. 2.1 във връзка с аксиоматично-дедуктивния характер на този тип модели, може да се формулира следното основно за нашето изследване положение: щом като нито в областта на речта, нито на отделните равнища на структурата на извода (област на езика) не може да се появи нищо, никакъв елемент или отношение (и следователно информация), които имплицитно да не се съдържат на по-дълбинните равнища (и съответните правила), то последователните преминавания при

[18] В случая става дума за описание на категориите на езика обект посредством специално създадена за тази цел семиотична система (метаезик), изградена въз основа на синтактични определения. За разлика от семантичните определения, които се основават върху явното съпоставяне с предмета, синтактичните определения, за които става дума, са такива, при които „предметът се различава от другите предмети чрез правилата за боравене с него, чрез начините или целите на неговата употреба" [39, с. 314, 315] (вж. също интересния обзор в [15, с. 120—135]). Следователно тук не става дума за това, дали трябва или не трябва да се описва семантичното равнище на езика, а за това, дали категориите на синтактичното равнище трябва да се описват с метаезик, изграден върху семантични или синтактични определения.

[19] Да напомним, че в повечето работи, посветени както на теорията на пораждащите граматики, така и на автоматичния синтактичен анализ (вж. например [42]), се прокарва последователно разграничение между начина на установяване и начина на представяне на синтактичните структури.

генерирането (или разпознаването) от равнището на базисните синтактични структури на отделните равнища на извода, както и от равнището на езика към равнището на речта и обратно, са превод в дефинирания от нас смисъл. А щом като това е така, то онова, което не е експлицитно зададено на дадено равнище, трябва да има референт на едно от по-дълбинните равнища на структурата на синтактичния извод на даденото изречение. С оглед на това не е мъчно да се види, че може да се предложи една друга процедура за идентифициране на референтите на нулевите елементи, която да се свежда към съответен анализ на структурите на синтактичния извод, приписани от даден модел, т. е. към анализ в „дълбочина" в областта на езика.

**2.3.2.** Следните примери илюстрират предложената възможност. При това ще разгледаме два опростени конструктивни примера (2.3.2.0), а след това — един пример от теоретичната литература (2.3.2.1) и един от даден експериментиран алгоритъм (2.3.2.2).

**2.3.2 0.** Нека наред с изходния символ $\#Z\#$ — изречение — имаме конфигурациите $GN_i$ (именна група) и $GV_i$ (глаголна група) и нетерминални ($P_i$ — местоимение изобщо, и $A_i$ - определителна дума изобщо) и терминални ($pd_i$ — показателно местоимение, $pp_i$ — притежателно местоимение, $a_i$ — прилагателно, $s_i$ — съществително, $v_i$ — глагол, $ad$ — наречие)[20] символи и нека ни е зададен следният фрагмент от една условна разпознаваща граматика ($G_1$):

1) $pd_1 pp_1 \rightarrow P_1$;
2) $ada_1 \rightarrow A_1$;
3) $P_1 s_1^1 \rightarrow GN_1^1$;
4) $A_1 s_1^2 \rightarrow GN_1^2$
5) $v_{\text{aux}} GN_1^2 \rightarrow GV_1$;
6) $GN_1^1 GV_1 \rightarrow Z$.

Да предположим също така, че разполагаме с един модел за лексико-морфологичен анализ, чрез използването на който руското входно изречение, което бяхме взели за пример,

$$\#\,\text{Этот наш друг очень способный студент}\,\#$$

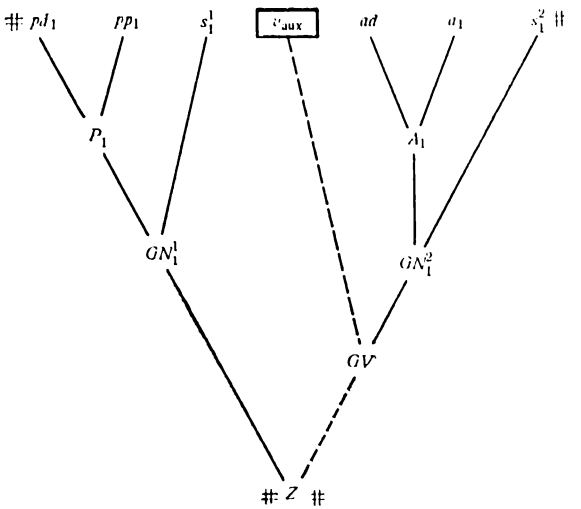
може да бъде сведено към неговата повърхностна структура (т. е. към една редица от синтактични словесни класове, представени чрез съответни терминални символи). Тази повърхностна структура изглежда така:

$$\#\,pd_1\ pp_1\ s_1^1\ ad\ a_1\ s_1^2\,\#$$

При прилагането на правилата на нашата $G$ към тази повърхностна структура се получава следната ситуация: правилата от 1) до 4) се прилагат нормално; обаче ние не сме в състояние да приложим правилото 5), а следователно и правилото 6) и да получим завършена структура на извода (дърво). Правилото 5) $v_{\text{aux}} GN_1^2 \rightarrow GV_1$ не може да се приложи, защото в повърхностната структура не фигурира в експлицитен вид лявата непосредствено съставяща на неговата лява част (т. е. символът $v_{\text{aux}}$), а правилото 6) $GN_1^1 GV_1 \rightarrow Z$ не може да се приложи, защото поради невъз-

[20] Да напомним, че долните индекси означават падежи (например $s_1$ е съществително в именителен падеж), а горните — реда на влизането в изречението на еднакви синтактични класове (например $s_1^1$ означава първото съществително в именителен падеж в работното изречение).

можността да се приложи правилото 5) не може да се реконструира дясната непосредствено съставяща на неговата лява част (т. е. $GV_1$). И така правилата 5) и 6) не могат да бъдат приложени и структурата на извода завършена, защото в повърхностната структура липсва явно зададен $v_{aux}$. Като изхождаме от посоченото по-горе обстоятелство, че моделите от този тип са дедуктивни системи, и от обстоятелството, че липсващият елемент ($v_{aux}$) е зададен в правилото 5), ние условно реконструираме този елемент в повърхностната структура, което ни позволява да приложим правилата 5) и 6) и да получим завършена структура на извода (синтактично дърво) на нашето изречение. Това е представено на фиг. 2, като реконструираният елемент е представен с $\boxed{v_{aux}}$, а страните на подграфовете, получени след неговото реконструиране, са дадени с пунктир.




Фиг. 2

Тъкмо този реконструиран въз основа на анализ на по-дълбинните структури на изречението и на правилата на $G_1$ елемент е референтът $A$ на нулевия елемент от нашия пример.

Същата възможност може да се покаже и при анализ (наистина в конкретния конструктивен пример с допълнителни интерпретации) на структурата на извода, която би ни дала и една граматика на зависимостите, описваща синтактичната структура на изреченията не в термините на НС, а в термините на граматическите зависимости между словоформите на изречението (вж. например [31, с. 105 и сл.]).

Нека при същите терминални символи, които използвахме при нашата граматика по НС ($G_1$), ни е зададен следният фрагмент от една условна разпознаваща граматика на зависимостите $G_2$:[21]

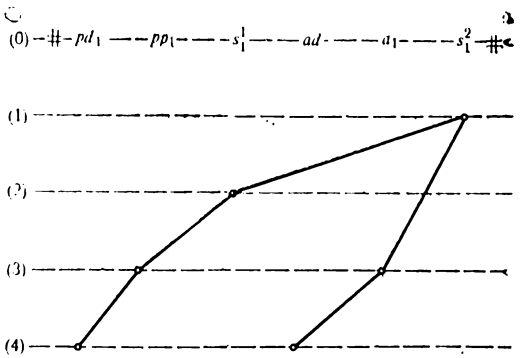
[21] Да напомним, че докато при граматиките от разгледания по-горе вид правилата от типа $X \rightarrow Y$ означават, че вместо $X$ трябва да се запише $Y$, в граматиките на зависимостите правилата от същия вид $(X \rightarrow Y)$ означават, че $Y$ зависи от $X$. Както и при НС-граматиките, горният индекс означава поредния номер на „влизането“ на един и същ клас в последователността на изречението, а долният — падежа.

1) $s_1^2 \rightarrow s_1^1$,
2) $s_1^1 \rightarrow pp_1$;
3) $s_1^2 \rightarrow a_1$;
4) $pp_1 \rightarrow pd_1$;
5) $a_1 \rightarrow ad$.

Като приложим правилата на тази граматика към повърхностната структура на нашето входно изречение

# Этот наш друг очень способный студент #,

получаваме структурата на извода (графа на управлението), дадена на фиг. 3.




Фиг. 3

Като приемем, както това се прави в повечето граматики от този тип, че за руския език „върхът", „господарят" на структурата на извода е сказуемото, което трябва да се намира на равнище (0), получаваме възможност за следната интерпретация (която може да се автоматизира): „върхът" на графа на управлението на всяко дърво (сказуемото) трябва да принадлежи на равнище (0); от това следва, че нашият граф, чийто връх е на равнище (1), е некомплексен. Тази констатация наред с допълнителни правила, които за простота тук няма да даваме, позволява да се реконструира $v_{\text{aux}}$ в повърхностната структура на входното изречение, което от своя страна създава основа за дълбинен анализ за установяване на референта на $\phi$.

**2.3.2.1.** След тези конструктивни примери ще дадем един пример от изследването на американския автор Д. С. Уорс „Трансформационен анализ на конструкции с творителен падеж в руския език" [71]. Авторът констатира, че при традиционните начини на описание на синтаксиса на руския език, които дават само неговата повърхностна картина, разрешението на случаите на граматическата многозначност (или омонимия) на конструкциите се основава върху чисто семантични критерии. При тези критерии изследователите, за да достигнат необходимата степен на общност, се виждат принудени да „боравят с термини, чиято неточност достига почти до безсмислица" [71, с. 679]. Вместо такъв подход авторът предлага и разработва трансформационен анализ на същите конструкции, основан изключително върху формални критерии.

Един от типовете на конструкции с творителен падеж, анализиран от Уорс, е структурният тип $s_1^1 v s_4^2 s_5^3$ ([71, с. 670—676]).[22] Според автора този структурен тип е получен (т. е. е трансформ) от ядрената структура

$$s_1^1 v s_5^2 \quad \begin{pmatrix} \text{он был президентом} \\ \text{он был студентом} \end{pmatrix}$$

чрез прилагане на трансформацията

$$T_1 : s_5^2 \Rightarrow s_4^2 s_5^3,$$

в резултат на което получаваме

$$s_1^1 v s_4^2 s_5^3 \quad \begin{pmatrix} \text{они выбрали его президентом} \\ \text{я знал его студентом} \end{pmatrix}.$$

Структурният тип $s_1^1 v s_1^2 s_5^3$ се смята за предикативен, ако допуска трансформация от вида

$$T_2 : s_1^1 v s_4^2 s_5^3 \Rightarrow s_1^2 \not\varnothing s_1^3,$$

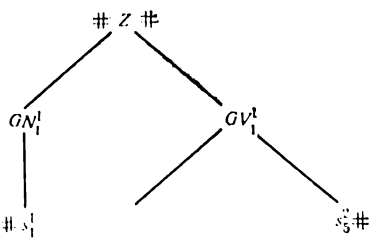
например

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # $s_1^1$ | $v$ | $s_4^2$ | $s_5^3$ | # | ⇒ | # $s_1^2$ | $\not\varnothing$ | $s_1^3$ # |
| # я | знал | его | студентом | # | ⇒ | # он | | студент #. |

Без да има нужда с оглед на интересуващата ни проблема да минаваме през $T_1$ и $T_2$, може да смятаме конструкциите с нулева връзка от вида $s_1^1 \not\varnothing s_1^2$ за трансформ на ядрената структура $s_1^1 v s_5^2$, която се поражда от една НС-граматика по начина, показан на фиг. 4.

Както виждаме, референтът на $\not\varnothing$ от трансформа $s_1^1 \not\varnothing s_1^2$ се намира на равнището на извода на НС-структурата.

**2.3.2.2.** Сега ще приведем така, както е излязъл от машината (електронен комплекс IBM 7044 и 1401), пример на изречение, автоматично анализирано от алгоритъма за синтактичен анализ от системата на руско-френския машинен превод на групата по МП в Гренобъл, Франция (Centre d'Etude pour la traduction automatique — CETA), която засега представлява най-доброто световно постижение в тази област [41]. При този алгоритъм за синтактичен анализ на машината се подава непрекъснат текст. След това думите от входния текст се търсят в речника, сегментират се на основи и окончания и в резултат на работата на алгоритъма за морфологичен анализ на основите се приписват всички възможни в логиката на този модел словесни класове, а на окончанията — всички възможни (при омонимия) синтактични кодове, интерпретиращи техните граматични значения. В хода на морфологичния анализ се разрешават само онези случаи на омонимията на словесните класове и на синтактичните кодове, които могат да бъдат обработени само в



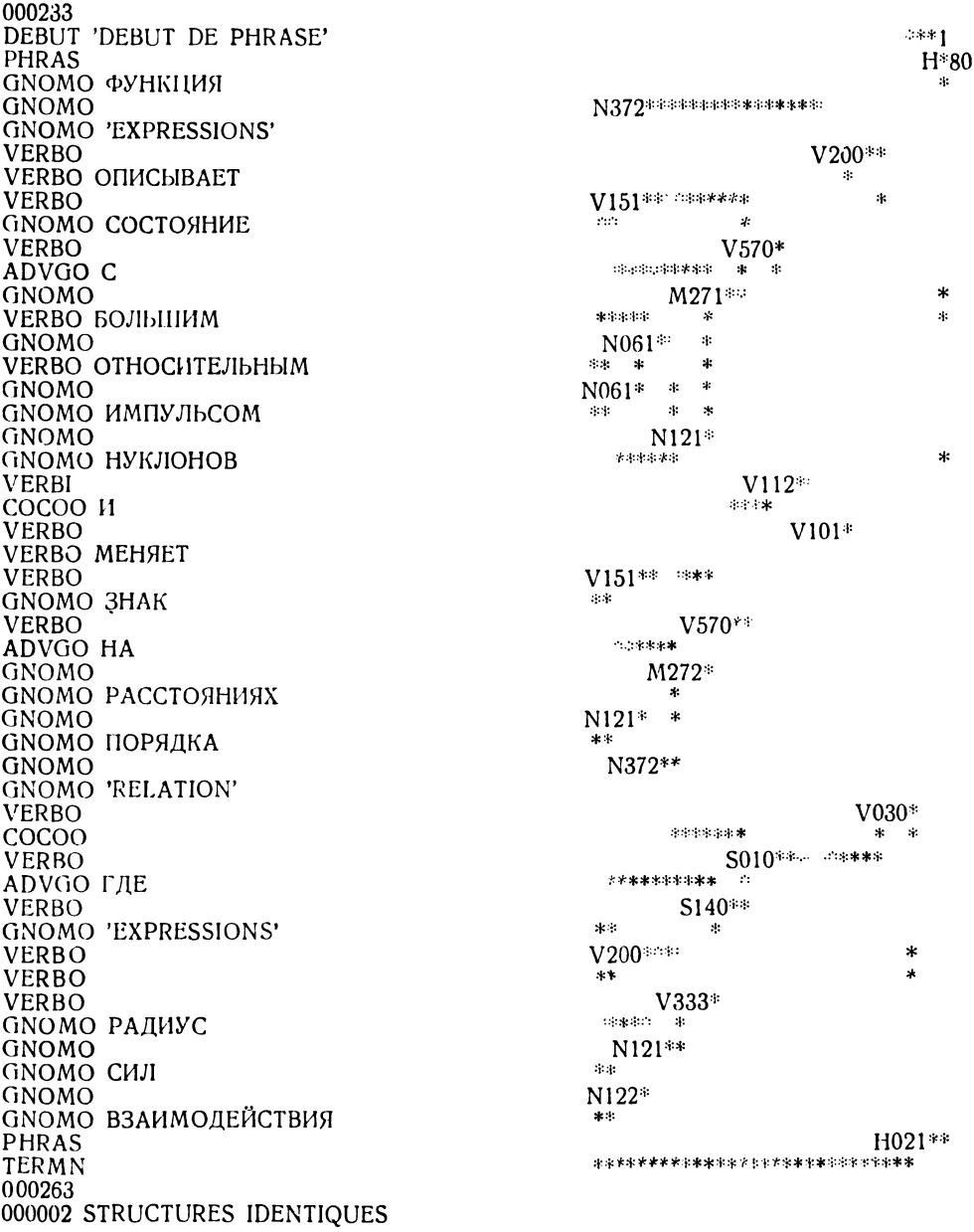

Фиг. 4

[22] Тук се използва почти същата символика, както и в разглежданите по-горе примери; към нея трябва да се прибави следното: $T_i$ е правило за трансформация от вида $X \to Y$, $s_i^2$ — съществително или заменящо го местоимение. Да отбележим също, че в работата си Уорс разглежда само редуцирани конструкции, т. е. такива, от които са изключени за простота на описанието всички елементи, невлияещи върху трансформационния потенциал на конструкцията. Например нередуцираният запис на нашия пример (1) дава следващия редуциран запис: $s_1^1 \varnothing s_1^2$.

рамките на единиците от тези две равнища (т. е. в рамките на словоформите). Ако в изречението няма граматическа омонимия (т. е. омонимия на словесни класове и на синтактични кодове), изходът на морфологичния модел е една верига от терминални символи. В обратния случай морфологичният модел приписва (т. е. дава като изход) толкова различни вериги от терминални символи, колкото са необходими, за да се представи всеки случай на омонимията поотделно (тъкмо на това се дължи полиструктурността, която се появява в хода на синтактичния анализ). Изходната верига (или изходните вериги) (елементарни синтагми) на морфологичния анализ представлява (или представляват) входът на алгоритъма за синтактичен анализ, изграден върху следния модел. Първата част от този модел е една безконтекстна НС-граматика, чието предназначение е да установи границите на отделните изречения и тяхната НС-структура. В хода на осъществяването на тази част от синтактичния анализ, от една страна, се изграждат всички възможни в логиката на дадения модел НС-структури (полиструктурност) и се отстраняват онези, които могат да бъдат отстранени по зададени правила за несъвместимост („насищане", saturation), а, от друга, се „коригират" неконтактните структури (т. е. превръщат се в контактни) чрез трансформиране на поддървета на структурата на НС-извода с помощта на теорията на графовете. Втората част от модела за синтактичен анализ е една граматика на зависимостите, която, от една страна, отстранява въз основа на несъвместимост с нейните правила някои от евентуално останалите случаи на полиструктурност, а, от друга, трансформира НС-структурата във „валентностна" структура. Като резултат от работата на тези две части на алгоритъма за синтактичен анализ[23] машината автоматично печата на изхода на лявата страна във вертикална последователност входното руско изречение (френските думи *expression* и *relation*, отпечатани в последователността на руското изречение, заместват формули и други символи; изразът *structures identiques* означава, че при предишния експеримент е била получена същата структура) и отдясно на нея окончателно приетата в хода на анализа синтактична структура (фиг. 5). Във всеки възел на графичното представяне на структурата на извода е напечатан номерът на приложеното правило (например N372, H180 и т. н.).

С оглед на интересуващия ни въпрос трябва да подчертаем следното. От гледище на традиционния синтаксис входното руско изречение е сложно-подчинено; в главното изречение имаме две прости „налични" сказуеми (*описывает* и *меняет*), а в подчиненото — „нулева" връзка (*где expresions ⌀ радиус...*). Както се вижда от приписаната структура на извода, по същия начин, както последователното прилагане на правилата V151, V570, V112; V151, V570, V101 и V200, изявява две налични еднородни сказуеми и ги обединява в „главен" връх на главното изречение чрез правилото V200, прилагането на правилата V200 и V333 идентифицира референта на нулевото сказуемо от подчиненото изречение и чрез S140 и S010 го свързва с „общия" връх на сложното изречение (V030),

[23] Синтактичният модел на СЕТА включва и една трета част — граматика от трансформационен тип, която семантизира валентностната структура, т. е. преобразува синтактичните отношения, идентифицирани в термините на граматиката на зависимостите, в отношения между семантични категории; например отношението сказуемо — допълнение се трансформира в отношение действие—обект.




Фиг. 5

който чрез H 021 и H 180 се свързва с дясната и лявата граница на изречението.

Приведените примери, колкото и опростени и ad hoc да са използваните в т. 2.3.2.0 $G_1$ и $G_2$, показват, че този начин на описание на механизма на разпознаването (и генерирането) дава възможност структурите на извода, които се приписват на съответните входни изречения, да бъдат

представени така, че на по-дълбинните равнища на извода да се изявява онова, което не е явно зададено на по-повърхностните. А това, както видяхме, осигурява възможността да се реализира предложената процедура за идентифициране на по-дълбинно равнище на референтите на нулевите елементи (които имат съответен начин на изразяване в синтаксиса на даден език). Но в същност основното предимство, което създават синтактичните модели от разгледания тип, се заключава в следното: всички подобни модели могат да се реализират от АСМ. А от това следва, че в рамките на тези модели ще бъде формализирано и алгоритмизирано и установяването на референтите на нулевите елементи, което по този начин се освобождава от неизбежния за първата идентифицираща процедура (вж. т. 2.2.0 и 2.2.1) субективен елемент, така и (и това е много по-важно) от необходимостта от така наречения извънлингвистичен анализ. Но въпреки тези предимства не бива да се забравя следното. Тъй като предложената процедура се основава върху анализ на структурното описание, приписвано на работното изречение от приетия модел, тя ще бъде безсилна във всички онези случаи, в които самият модел не е достатъчно мощен за описание на известни части от механизма на езика (в тази насока вж. например [66]). Тъй като моделите от разгледания тип описват само механизма на разпознаването и на генерирането на абстрактни синтактични структури, като се абстрахират от семантичния фактор, то те очевидно дават възможност да се идентифицират само такива референти на нулеви елементи, които имат експлицитен начин на изразяване в синтаксиса на даден език, но не са в състояние да създадат предпоставки за решаването на проблемата тогава, когато става дума за такива категории, които нямат изразяване на това равнище. Тъкмо това се наблюдава в нашите примери (2), (3) и (4) (вж. с. 8). От това следва, че в тези случаи дълбинната процедура за идентифициране на референтите на нулевите елементи би била възможна само при модели, които вземат пред вид и семантичния фактор.

**2.4.** Тъй като тук излизаме от положението, че на по-повърхностното равнище не може да се появи нищо, никаква информация, която да не е зададена на по-дълбинните равнища, трябва да се приеме, че референтът на един нулев елемент, който няма морфологическо или синтактично изразяване в даден език, неизбежно трябва да бъде зададен на най-дълбинното, т. е. на семантичното равнище. От изложеното в т. 2.2 не е мъчно да се види, че такъв референт, представящ определена единица или категория от семантичното равнище, може да бъде (в повечето случаи) идентифициран посредством първата описана процедура — чрез анализ на лингвистическия и извънлингвистическия контекст. Но в много случаи резултатите от тази процедура оставят да се желае нещо, значително по-добро, тъй като освен всичко друго това е една от онези области, в които субективността на преводача (обем на знания, аналитични възможности и пр.) се проявява най-осезателно. Тъкмо с оглед на това се поставя въпросът, дали и тук не е възможно да се използва процедурата на дълбинния анализ, която би сложила идентифицирането на тези референти на обективна основа.

Както вече бе отбелязано, тази възможност предполага наличността на модели, описващи и механизма на свързването на смисъла (значението) със звука (или писмените знаци), т. е. на формални граматики, от-

читащи и семантичния фактор. Следователно, за да отговорим на поставения въпрос, трябва първо да видим дали изобщо съществуват такива модели и какви са техните основни свойства. На това са посветени следващите пет точки 2.4.0.—2.4.4.

**2.4.0.** Тъй като изследванията, за които ще става дума по-долу, се намират още в първоначалните си стадии, ние ще може да разгледаме проблемата само в общия, ѝ принципен аспект.

Когато се говори за семантични модели, обикновено се подчертава следното различие: модели, които си поставят за цел да възпроизведат механизма на генерирането на смисъла (мисълта),[24] и модели, които си

[21] Тук основната идея се свежда към следното: съдържанието на различните мисли, смисълът на дадено изказване не е заложен в готов вид в главата на човека (както и самият израз), а се генерира в зависимост от ситуациите, които се отразяват в дадения случай. Щом като това е така, може да се предположи, че има нещо общо между генерирането на смисъла и генерирането на езиковите изрази, въплъщаващи този смисъл, има някакви елементарни смислови единици, някакви правила за тяхното комбиниране и някаква пораждаща процедура.

Мисълта, че трябва да съществуват елементарни смислови единици, беше продължение на набелязалата се доста отдавна в теоретичното езикознание тенденция да се намерят атомите, диференциалните признаци (по аналогия с фонологията) на значението, да се сегментира и структурира семантиката — срв. принципа на изоморфизма и на двойната сегментация на Йелмслев [44], при реализацията на който, както казва проф. Мартине, „ще може да получим на равнището на съдържанието, както вече получихме на равнището на израза, единици, по-малки от знака“ [45 b, с. 30]; вж. също работите на Кантино [82] и на Прието [47, с. 134—143] и направените в тях опити да се установят смислови единици, по-малки от знака, и система от смислови опозиции (в тази насока вж. също възгледите на проф. Курилович [48], изследванията на Сьоренсен [81], на Жарден [50] и др.).

Този стремеж, обусловен от общотеоретични съображения, беше значително засилен и актуализиран с оглед на проблемите, които се поставиха при създаването на така наречените информационни езици за автоматизирането на процесите на научно-техническата информация, и на нуждите на смисловия машинен превод (вж. например [51]) и доведоха Пери и Кент [52] до опита да представят смисъла на езиковите изрази (на първо място на думите) като произведения от елементарни смислови единици, т. е. семантични множители (СМ) (semantic factors). Тези идеи бяха развити по-нататък, задълбочени и реализирани в многоезичен аспект от колектива на Първата лаборатория по МП при I МГПИИЯ в Москва (вж. например [53]). Оригинални изследвания, които си поставят за цел да моделират процедурата на пораждането на смисъла, принадлежат на ръководителя на Миланската група по кибернетика и математическа лингвистика проф. Чекато (вж. например [54]). Идеите на италианската школа възникнаха при анализа на интелектуалната дейност на човека. Като изхожда от общоприетото положение, че една от основните задачи на лингвистиката е да установи съотношението между звука и смисъла (значението) и следователно между плана на съдържанието (мисъл) и плана на изразяването (език) и че нито единият, нито другият от тях могат да бъдат разбрани докрай, откъснати един от друг, проф. Чекато смята, че моделирането на механизма на езика трябва да се основава върху моделирането на процеса на мисленето, който според него може да бъде представен операционно (оттук и названието на школата), т. е. че съдържанието на всяка мисъл (смисъл) може да бъде получено чрез реализирането и комбинирането на четири елементарни умствени операции: „Резултатите от операционния анализ на мисленето и възможностите да се моделират съотношенията между мисленето и неговия апарат показват, че всички умствени дейности на човека могат да бъдат сведени към четири основни вида дейности: диференциация (тази операция дава номинантите, свързани с измененията на състоянието, например *тъмен — светъл* — А. Л.); диференциация на пространственото положение или фигурация (тази операция трябва да моделира осъзнаването на пространствените премествания и формите, като например *кръгло, копиеобразно* — А. Л.); категоризация (тази операция позволява да се получат абстрактни номинанти, като например *причина, време* и пр. — А. Л.) и корелация (тази операция се свежда към установяване на съотношения между „мислените неща“ — А. Л.); корелациите могат да влизат в други корелации и да образуват „мрежи от корелации“ [55].

Да отбележим, че освен много други дискусионни неща този подход се изгражда върху принципното откъсване на мисленето от езика и върху представата за езика като

поставят за цел или извличането на смисъла от дадени текстови единици (изречения), или въз основа на породен от някакъв модел смисъл да генерират съответния или съответните изрази на равнището на речта на даден език, обективиращи този смисъл. С оглед на целите на нашето изследване ще разгледаме само втория от тези типове модели.

**2.4.1.** Тук не сме в състояние да разглеждаме онази концептуална еволюция, която доведе до едно по-широко разбиране на формалните граматики като съвкупност от правила, генериращи не само белязани терминални синтактични структури, но и съответните семантични интерпретации (вж. например работите на Й. Кац [58], на сътрудниците на Пражката група по машинен превод и математическа лингвистика [56] и на Ревзин [59], Шаумян [15] и др.), и ще отбележим само (с оглед на интересуващата ни проблема) онова съществено, но не винаги ясно отчитано влияние, което оказваха в тази насока изследванията по МП, както и вътрешната логика на тяхното развитие.

Относителната мощност (до към 70—80%) на бинарните алгоритми за МП, основани върху така наречената лексико-морфологическа база, които интензивно се разработваха след първите сполучливи опити за самостоятелен МП (Ню Йорк 1954 и Москва 1955) до към края на петдесетте години, доведе редица изследователи и на първо място Ингве [60], който развн идеите на Освалд и Флетчер [61], Молошная [62], Мельчук [42], сътрудниците на Ленинградската група по МП (Андреев, Фатиалов, Цейтин), на групата по математическа лингвистика и МП при Ренд корпорейшън в САЩ, на СЕТА в Гренобъл, на Пражката група и др., до мисълта, че решението на проблемата на МП трябва да се търси на базата на синтактичния анализ и синтез. Така възникна така наречената синтактично направление в теорията на МП (вж. например [67])[25]. Основната идея на това направление се свежда към следното.

Чрез прилагането на съответни разпознаващи модели да се установява абстрактната синтактична структура на входните изречения и след това чрез реализирането на съответен пораждащ модел да се генерира адекватната синтактична структура на изходното изречение; автоматичното изявяване на синтактичната структура би позволило според мнението на съответните автори да се решат редица случаи на омонимия и мно-

---

механизъм, обличащ вече готови мисли. В областта на моделирането на процеса на мисленето трябва да се отбележат и твърде интересните принципни съображения, които изказва Шаумян, обсъждайки някои паралели между механизма на пораждането на езиковите изрази от предложения от него апликативен модел и механизма на процеса на мисленето [15].

[25] При това още по-ясно се прояви взаимната връзка и обусловеност между принципните основи на решаването на лингвистическите проблеми при МП и същността на господствуващите през даден период езиковедски теории. Така в резултат на факта, че към началото на петдесетте години проблемата за формализиране на описанието на значенията на езиковите изрази едва се поставяше (а машината може да борави със смисъла само ако той е представен по съответен начин), а традиционният синтаксис се изгражда на семантична основа, създателите на първите системи за МП бяха твърде ограничени при снемането на лексическата и морфологическата омонимия и полисемия — трябваше да се задоволяват с формален анализ на непосредственото окръжение и бяха лишени от възможността да провеждат автоматичен синтактичен анализ, т. е. трябваше да водят превода дума по дума. Тъкмо поради това преводните алгоритми се изграждаха изключително върху лексико-морфологическа база с контекстуален анализ. Развитието на теорията на формалните граматики и изобщо на математическата лингвистика с нейните генериращи и разпознаващи модели, изградени на чисто формални основи и допускащи непосредствено реализиране от машината, позволи да се алгоритмизира и синтактичният анализ.

гозначност от лексическото и морфологическото равнище, които не можеха да бъдат решени от алгоритмите, основани само върху лексико-морфологична база.

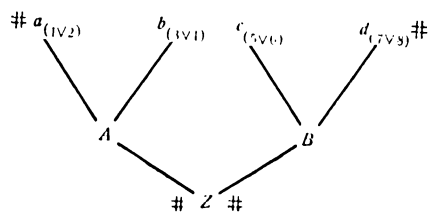
Но въпреки значителния напредък, който беше осъществен в работите на представителите на синтактичното направление в теорията и практиката на МП, той не оправда напълно надеждите, които се свързваха с въвеждането на синтактичния анализ в системите на МП: от една страна, повечето автори, може би неволно, започнаха да разглеждат синтактичния анализ като самоцел, което ги доведе в по-малка или по-голяма степен до игнориране на проблемите на лексиката и на синтеза и до разтваряне на лингвистическата проблематика на МП в общата проблематика на алгебричната лингвистика, което сложи решаващ отпечатък върху цялата съвременна стратегия при МП (по-подробно вж. т. 5.2); от друга страна, практиката показва, че доста проблеми (като оставим настрана редица дори синтактични проблеми, които не можеха да бъдат решени поради недостатъчната мощност на съществуващите модели), като например редица случаи на лексическа многозначност, фразеология, проблеми на актуалното членение и пр., не можеха да бъдат решени и при такъв синтактичен подход. Но не само това. Създаването на модели за синтактичен анализ при МП и тяхното машинно реализиране поставиха и две извънредно сложни взаимосвързани и, както изглежда, неразрешени без привличането на семантически критерии проблеми — проблемата за избора на оптимална граматика и проблемата за полиструктурността.

Както е известно (вж. например [41]), при един основен логически тип на модела може да се създадат няколко конкретни граматики. При това съвременното езикознание и логика предоставят на лингвиста доста богат избор както от основни типове, така и от конкретни варианти (наред с възможността да се създадат нови). Очевиден е стремежът системата на синтактичния МП да се основе върху най-адекватен прост и мощен модел. И тъкмо тук се постави проблемата за критериите на избора на оптималната граматика (както и за критериите на самата оптималност). Тъй като практически засега граматическата теория (в терминологията на Чомски, т. е. теорията на теорията на езика) не е в състояние да даде точен отговор дори на най-слабото изискване, което може да се предяви към нея — да дава процедура за избор между две конкретни граматики[26], изследователите са лишени от логическа основа при този избор.

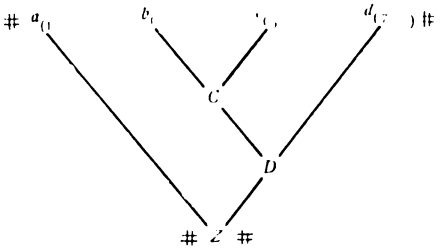
Втората проблема, с която се сблъскаха изследователите при разработването на автоматичния синтактичен анализ при МП, е проблемата за п о-

[26] Към теорията на граматиката могат да се предявяват три различни по мощност изисквания ([38], с. 458): да дава процедура за конструиране (т. е. при наличието на една съвкупност от изрази, принадлежащи на езика $L_i^N$, да дава практичен и автоматичен метод за конструиране на оптимална граматика $G$ за този $L_i^N$); процедура за преценка (т. е. при наличността на една съвкупност от изрази от $L_i^N$ и на дадена граматика $G_1$ на този $L_i^N$ да дава практичен и автоматичен метод за установяване дали тази граматика е наистина най-добрата за съответния език) и процедура на избор (т. е. при наличност на една съвкупност от изрази от $L_i^N$ и на две граматики $G_1$ и $G_2$ на този $L_i^N$ да може въз основа на нея да се реши коя от двете граматики е по-добрата за $L_i^N$).

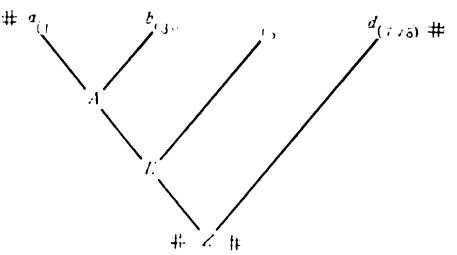
листруктурността: оказа се, че при автоматичен анализ поради омонимичността на входните терминални символи (словесни класове), чиято последователност интерпретира на равнището на повърхностната структура съответното анализирано изречение и представя входа на синтактичния модел, правилата на една и съща граматика приписват на това изречение в редица случаи не една, а няколко структури. Тъкмо това се нарича полиструктурност. Това явление може да се представи чрез следния условен и съзнателно пресилен абстрактен пример.



Фиг. 6



Фиг. 7



Фиг. 8

Нека имаме една верига от терминални омонимични символи (омонимичността е отбелязана в индекса чрез $i \vee j, i = 1,3,5,\ldots$, $j = 2,4,6,\ldots$, което означава, че съответният символ може да принадлежи на класа $i$ или на класа $j$)

$a_{(1\vee 2)}$ $b_{(3\vee 4)}$ $c_{(5\vee 6)}$ $d_{(7\vee 8)}$

и следния фрагмент от една условна разпознаваща граматика:

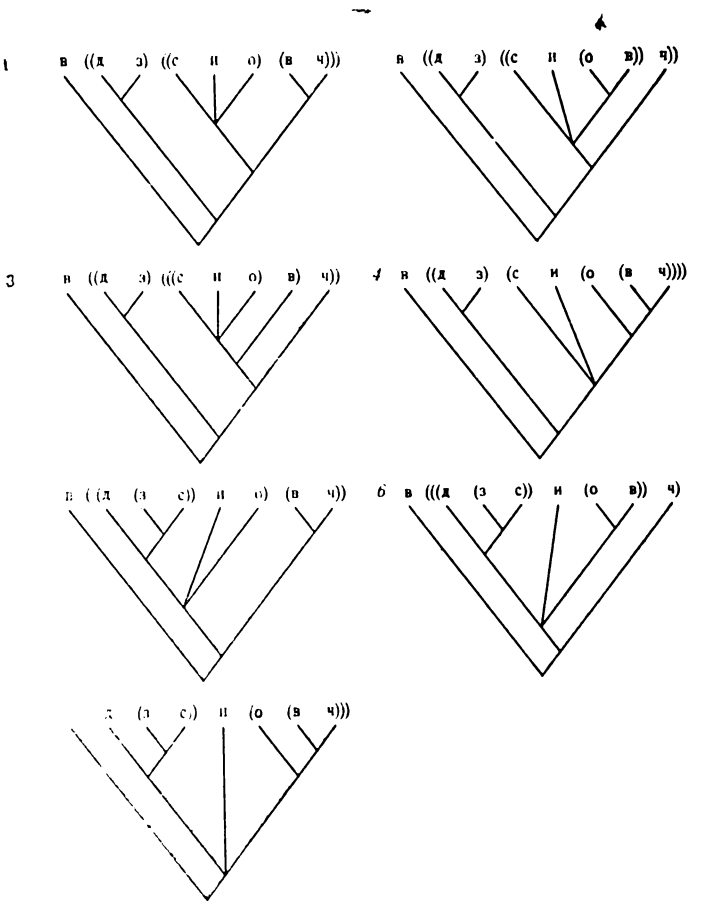
1) $a_1 b_3 \rightarrow A$;
2) $c_5 d_7 \rightarrow B$;
3) $AB \rightarrow Z$;
4) $b_4 c_6 \rightarrow C$;
5) $Cd_8 \rightarrow D$;
6) $a_2 D \rightarrow Z$;
7) $Ac_6 \rightarrow E$;
8) $Ed_7 \rightarrow Z$.

Прилагането на тези твърде опростени правила към нашата входна редица (изречение) дава полиструктурност, представена на фиг. 6, 7 и 8.

Не бива да се мисли, че появяването на тези три различни структури на извода (I—III) се дължи само на абстрактността на входните символи и на преднамереното съставяне на правилата на $G$. Практическата работа по синтактичен МП показва, че полиструктурността се появява много по-често, отколкото може да си представим.[27] И тъкмо тук се постави извънредно мъчно разрешимата (а и в много случаи и нерешимата) в рамките на формалния синтактичен анализ проблема как да се снеме тази полиструктурност, т. е. да се избере „правилната“ в случая структура. Фактически полиструктурността е форма на проявление на синтактично равнище на нерешените на лексико-морфологическо равнище омонимии. Но тъй като едно от първите съображения да се въведе синтактичен анализ в системите за МП беше тъкмо снемането на тази лексико-граматическа омонимия, разрешаването на

[27] Така например синтактичният модел, използван от групата по МП при Унгарската академия на науките [70], приписва 7 структури на извода на руския израз „... *вследствие других законов сохранения и особенностей взаимодействия частиц*“ (вж. фиг. 9).

синтактичната полиструктурност поне в повечето от съществуващите системи не може да използва информация от това по-повърхностно равнище и единственият път е да се търси решение на още по-дълбинно, т. е. на семантично равнище.




Фиг. 9

Тогава съвсем естествено започна все повече да се утвърждава мисълта (която между другото беше доста отдавна изтъквана в някои изследвания, посветени на общата проблема на превода — вж. например [28]), че целта на МП е, както и на обикновения превод, да се извлече смисълът (т. е. информацията) от входния текст, да се представи той в термините на някакъв универсален семантичен матаезик (например в термините на семантични множители или в езика на предикатното смятане — вж. например [34]) и въз основа на него да се генерира изходно изречение със същия смисъл. Но, както не е мъчно да се види, такова разбиране на проблемата на МП, при което анализът може да се разглежда като преминаване от графемичната (или фонетичната) форма на равнището на текста към смисъла, а синтезът — като преминаване от зададен смисъл

към неговото графическо представяне на равнището на текста на някакъв друг $L_j^N$, предполага разпознаващи и генериращи процедури (модели), които да не се абстрахират от проблемите на семантиката, както това беше във всички модели, изградени върху разбирането на граматиката, предложено в първите работи на Чомски, а тъкмо обратното — включващи семантиката като своя органическа част.

Именно така за втори път в продължение на своето едва двадесетгодишно съществуване теорията на МП отново се свързва с най-новите насоки на съвременната лингвистика и оказа съществено влияние върху разработването на семантичните модели. Тъй като създаването на такива модели е твърде сложна проблема, то, както това се наблюдаваше и при синтактичните модели (и то по понятни психологически причини), изследователите се насочиха първо към създаването на пораждащи процедури, чиито принципи ще бъдат накратко разгледани по-долу.

**2.4.2.** Както теоретическите изследвания, посветени на този тип пораждащи модели, така и немногобройните предложени засега конкретни реализации изхождат от следната предварителна хипотеза: „Да предположим, че по някакъв начин е възможно да се изброи множеството от вериги, отговарящи на дадени синтактични условия, което да е модел на множеството от значенията (смислите) на всички граматически правилни изречения в даден език. Такова множество би могло да бъде например множеството от семантичните интерпретации на изреченията, с което борави Кац“ [64, с. 30—31]. При такива изходни данни тези модели си поставят за цел да възпроизведат онази част от механизма на езика,[28] въз основа на която се избира необходимата за дадения смисъл дълбинна синтактична структура, тази структура последователно се разгръща (трансформира, превежда), преминавайки през съответните равнища на извода, в повърхностна структура, на равнището на която се реализира лексическото „запълване“ и съответно морфологическо и фонетично (графично) оформяне.

Тъкмо с оглед на това качествено ново виждане на същността и целта на формалните граматики се засилиха и условията, на които те трябва да отговарят: към двете условия, на които трябваше да отговарят формалните граматики от първоначалния тип — да пораждат (или разпознават) всички граматически правилни изречения от даден $L_i^N$ и само тях и да им приписват структурни описания, непротиворечащи на нашата интуиция, — се прибавя и трето — да описват механизма на свързването на значението със звука (вж. например [64]).

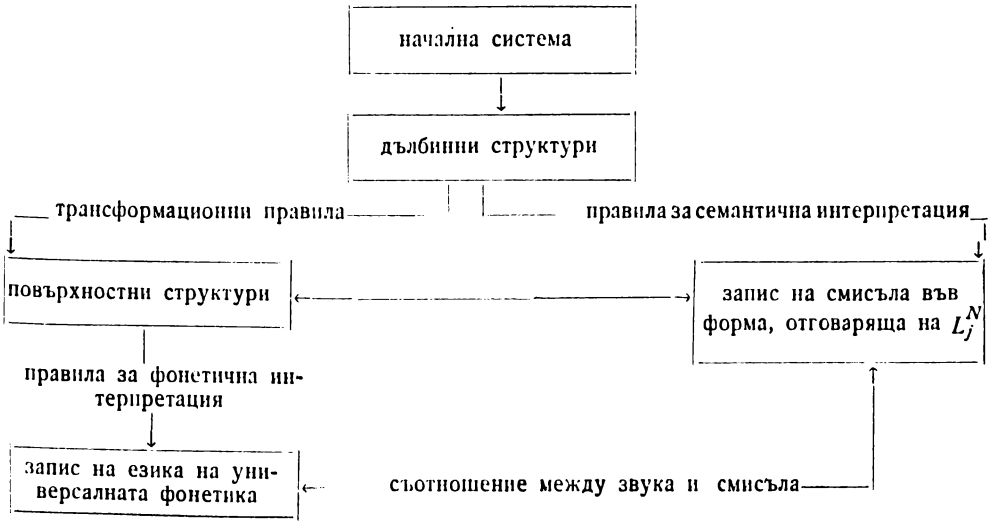
Хипотезата, че разполагаме с някаква процедура за произвеждане и представяне на смисъла, залегна в основата на твърде интересни семантични модели, като например модела, предложен от Сгал (вж. [64]) и си-

[28] Строго погледнато, вътрешният механизъм на езика позволява на носителя на този език да извършва следните четири основни операции: 1) да избере съобщението, което по някакви съображения иска да предаде; 2) да формулира това съобщение, т. е. да преведе неговата семантична интерпретация в едно от съответните фонетични представяния; 3) да разбере съобщението, т. е. да го преведе от фонетичното му представяне на едно от съответните семантични представяния; 4) да подбере най-подходящото представяне както при 2), така и при 3). Очевидно е, че 1) зависи от екстралингвистични фактори, а 4) не може да бъде описано само с чисто лингвистични средства [66]. Следователно задачата на съответните граматики е да опишат механизмите 2) и 3).

стемата за множествен семантичен синтез на Мелчук (вж. [35]). Анализът на структурата на извода на тези модели, например в системата на Мелчук преминаването от дълбинната лексико-синтактична структура (ЛСС) към повърхностната структура и всички възможни перифрази на равнището на речта на даден $L_i^N$, показва, че подобно на разгледаните в т. 2.3 формални синтактични модели те дават възможност да се идентифицира на по-дълбинно равнище референтът на нулев елемент, съответствуващ на някаква семантична категория.

Тази възможност да се използва дълбинната процедура за идентифициране на референтите на нулевите елементи при семантичните модели може да се илюстрира от модела на механизми на езика (отговарящ на трите посочени по-горе условия), предложен от Чомски на Конференцията по семиотика през 1966 г. в Полша.

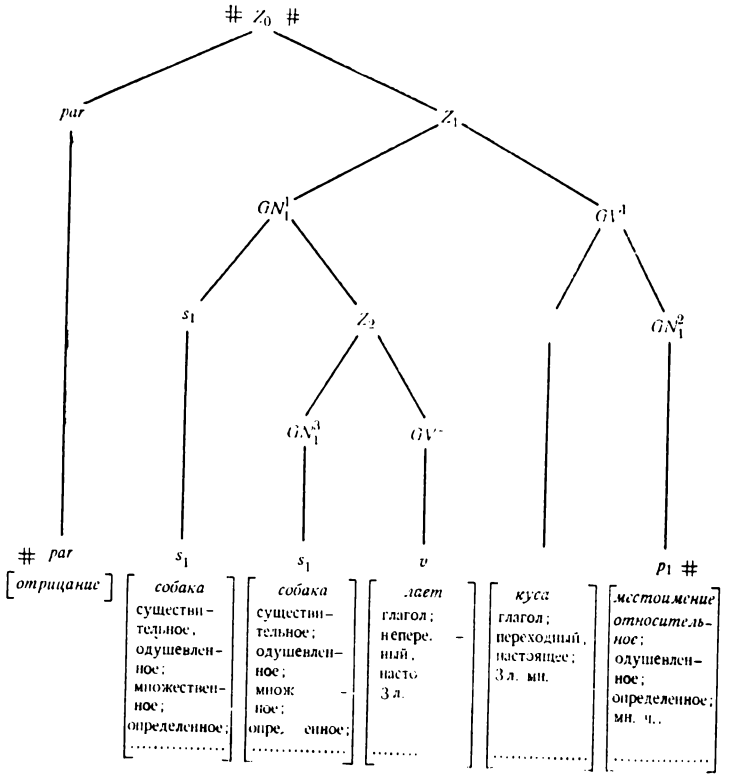
**2.4.3.** Логическата схема на този модел се свежда към следното. Една начална система поражда дълбинните смислови структури на изреченията; към тези дълбинни структури, представящи запис на смисъла на входното изречение и на цялата информация за неговото реализиране във формата на някакъв универсален семантичен метаезик, от една страна, се прилагат правила за семантична интерпретация, които трансформират (превеждат) тези дълбинни структури в съответни записи на смисъла във форма, отговаряща на изходния език, а, от друга — така наречените трансформационни правила, които превеждат синтактичния компонент на дълбинните структури в съответни повърхностни структури, отговарящи на изходния език, към които на свой ред се прилагат правилата за фонетичната интерпретация, трансформиращи (превеждащи) повърхностната структура на изречението и нейното смислово „напълване“ в запис на езика на универсалната фонетика. Графически тази схема е представена на фиг. 10 (вж. [33]).




Фиг. 10

Съотношението между дълбинната структура, която трябва да предопределя за всеки терминален символ от повърхностната струк-

тура всички семантични, синтактични, морфологични и фонетични признаци на езиковите единици от равнището на текста на съответния език, които ще се породят, и повърхностната структура може да се представи със следния пример (вж. [33, с. 38]).




Фиг. 11

Дълбинната структура на руското изречение (а това е нашият пример (2) от т. 2.2)

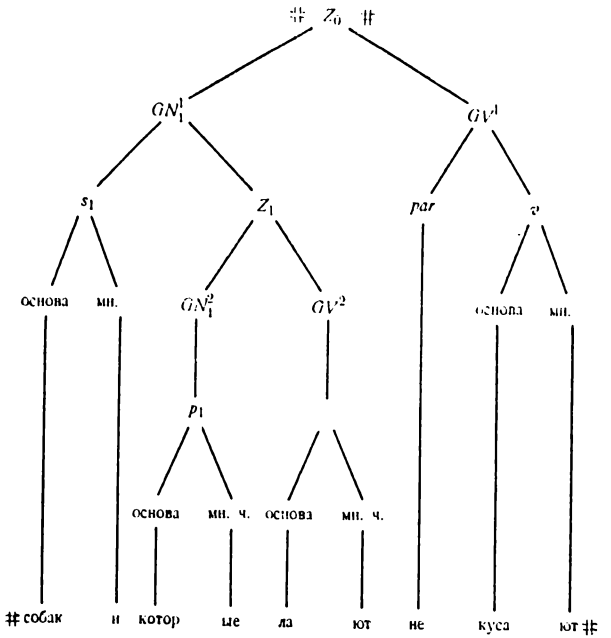
# собаки, которые лают, не кусают #

е представена с известно опростяване на фиг. 11. (Смисълът на думите на семантичното равнище (и съответната допълнителна информация) би трябвало да бъде записан в термините на езика на универсалната семантика, но тъй като такъв език още не е разработен, те се представят във формата на руски език.)

А повърхностната структура на това изречение получава вида, даден на фиг. 12 (при което изводът би трябвало да бъде записан в символите на универсалната фонетика).

Този пример е твърде показателен от интересуващото ни гледище. Да предположим, че разглежданото руско изречение трябва да се преведе на български език. С оглед на това, че в български език съществува постпозитивна членна форма, която служи като формален изразител на семантичните категории определеност и неопределеност, докато тези се-

мантични категории нямат формално изразяване в граматическия строй на руския език, във входното руско изречение се появява нулев елемент, чийто образ може да бъде българският определен или неопределен член. За да може машината или човекът преводач да решат кое от тези две




Фиг. 12

съответствия на нулевия елемент в руското изречение да вземат, те трябва да установят неговия референт на едната от по-дълбинните структури на извода на това изречение. Не е мъчно да се види, че разглежданият начин на представяне дава възможност да се идентифицира търсеният референт на нулевия елемент на най-дълбинното — смисловото равнище: и наистина, в информацията към смисловата единица, получаваща руската интерпретация *собака*, началната система е генерирала и показателя определеност, който е референт на интересуващия ни нулев елемент и дава достатъчно ITN за правилното подбиране на преводното българско съответствие на нулевия елемент от нашия пример.

**2.4.4.** Както виждаме, начинът на представяне на пораждането (и разпознаването) на езиковите изрази от семантичните модели, изградени върху най-новите концепции за същността на формалните граматики, осигурява възможност за прилагането на процедурата на дълбинния анализ за идентифициране на референтите на нулевите елементи. Обаче не може да не се подчертае, че от гледище на практическите нужди на МП и на автоматичното третиране на естествените езици тази възможност ще остава само една възможност дотогава, докато хипотезата, че разполагаме с процедура, пораждаща дълбинни структури, както и със система за записване на този смисъл (език на универсалната семантика или някакъв друг код) не бъде заменена от конкретни ефективни процедури.

§ 3. След като въведохме понятието нулев елемент и обсъдихме някои от неговите общи свойства и след като видяхме, че най-новите концепции и разработки на структурното и математическото езикознание дават възможност да се изграждат такива модели, чиито структури на извода позволяват в резултат на една чисто обективна и допускаща формализиране процедура да се идентифицират референтите на нулевите елементи на по-дълбинни равнища чрез анализ на структурата на извода вместо повърхностния анализ на равнището на текста, сега трябва да видим какви проблеми се поставят с оглед на това за теорията и практиката на МП, както и изобщо за автоматичната обработка на информацията.

От изложеното дотук следва, че в последна сметка при дадена двойка езици или две различни равнища на един и същ език проблемата на нулевите елементи се свежда към проблемата за идентифицирането на техните референти. В този параграф, като имаме пред вид отбелязаното в началото на тази работа обстоятелство, че въпросът за нулевите елементи при МП и автоматичното третиране на естествените езици изобщо не е поставян и обсъждан в обща принципна форма, ще се опитаме, от една страна, да формулираме основните проблеми, които се поставят във връзка с обработката на нулевите елементи при алгоритмизирането на процеса на превода (3.1) и, от друга, да предложим наша концепция за оптимална стратегия при МП (3.2), което ще създаде принципните предпоставки за решение на проблемите, формулирани в (3.1).

3.1. Практически проблемата за нулевите елементи при МП се свежда към това, че машината трябва да генерира в изходния текст техните образи (съответствия). Но за да може да стане това, тя трябва, фигуративно казано, да „знае“ к а к в о да генерира и к о г а да генерира. Машината ще знае какво съответствие трябва да генерира на даден нулев елемент от входния текст, след като е идентифицирала неговия референт въз основа на едната от двете разгледани в предишния параграф процедури. Така стои въпросът с това, какво съответствие трябва да генерира машината при един нулев елемент. Обаче отговорът на въпроса, к о г а трябва да го генерира, е значително, поне на пръв поглед, по-сложен.

Да предположим, че имаме някаква специална процедура за идентифициране на референтите на нулевите елементи, която може да наречем схема „Нулев елемент“. Сам по себе си фактът, че в анализиращия (или синтезиращия) алгоритъм е включена такава схема, още не е достатъчен за автоматичното решение на проблемата за нулевите елементи: както всяка схема от алгоритъм, така и схемата „Нулев елемент“ може да заработи само при съответна сигнализация. Следователно, преди да се решава въпросът, върху какви принципи трябва да бъде изградена схемата „Нулев елемент“ (и изобщо дали тя трябва да се създава — вж. по-долу), спецификата на превода, осъществяван от машина, поставя за разрешение предварителния въпрос за сигнализирането на тази схема. Човекът преводач в повечето случаи изобщо не се замисля над тази проблема, защото, както можем да предполагаме (въпреки че за това няма точни експериментални данни), „човешкият“ анализ протича много по-комплексно, по-многоаспектно, циклично и „зигзагообразно“ както по широчина (линейност), така и на дълбочина и „извън“ езика. Но този

въпрос, който de facto обикновено не представлява трудност за човека, създава ако не принципни, то фактически трудности за машината.

И наистина, как може да „разбере" машината, че в дадено изречение, което тя анализира и превежда, има нулев елемент, на който тя трябва да генерира съответствие, т. е. как да набере съответната информация за сигнилизиране на схемата „Нулев елемент"? По принцип, както това е във всички случаи при автоматичното решение на дадени езикови проблеми в процеса на превода, тази информация може да се получи от машината по два различни начина: да ѝ се зададе „списъчно" или да се набира от самата нея въз основа на някакви константи и правила.

Не е мъчно да се види, че в интересуващия ни случай „списъчното" задаване на информацията за това, че в конкретния случай имаме изречение с нулев елемент, е практически изключено, защото е равносилно на изискването на машината да бъде зададен изчерпателен списък на всички такива изречения в дадения входен език при превод на друг даден език.[29] Следователно остава вторият път: да се набира информацията, необходима за сигнализиране на нулевия елемент, от самата машина въз основа на някакви правила. Но какви могат да бъдат тези правила? Не е мъчно да се покаже, че машината може да набере информация, сигнализираща за наличността на нулев елемент в дадено изречение, само като установи неговия референт (било в $L_i^N$ при анализа, било в $L_j^N$ при синтеза — вж. по-долу). И тук като че ли стигаме до един магьосан кръг: за да се генерира съответствие на нулев елемент, трябва да се установи неговият референт, а неговият референт би могъл да се установи само ако предварително се знае, че в случая има нулев елемент, което пък може да се знае само ако се установи неговият референт.

Както показва работата по съставянето на алгоритъм за превод на математически текстове от руски на български език, провеждана от групата „Машинен превод и математическа лингвистика" при Математическия институт на БАН, този магьосан кръг е привиден: от съдържателно гледище двете процедури — процедурата по сигнализиране на нулевия елемент и процедурата по идентифициране на неговия референт — предполагат набирането на една и съща информация и следователно за машината те са една и съща процедура, която (заедно със съответното генериране) ще обобщим под израза о б р а б о т к а н а н у л е в е л е м е н т. А тъй като тази процедура (независимо от това, какви са нейните принципи и конкретна организация) може да се прилага към всяко работно изречение първоначално като контролна процедура (дори евентуално в някои случаи по сигнализация от лексемния речник), целият въпрос се свежда към това, как да се организира тази процедура.

**3.1.0.** И двете процедури, които обединихме като обработка на нулев елемент при МП, предполагат набирането на съответна информация. От

Да напомним, че още при обсъждането на възможностите на МП в началото на петдесетте години един от пионерите в тази област, проф. Бар-Хилел, казваше, че ако изследователите не биха били ограничени от факторите време, средства и обем на паметта на машината, МП не би представлявал, поне умозрително, никаква трудност и би могъл да бъде сведен към просто търсене в един речник, в който трябва да бъдат зададени всички възможни изречения от $L_i^N$ и на всяко едно от тях да бъде поставено в съответствие превеждащото изречение от $L_j^N$.

практическите ни опити в тази насока и от нашите общи концепции за превода и машинния превод (вж. например [28]), може да се каже, че съществуват следните възможности:

А.1. Да се стремим да организираме преводния алгоритъм така, че информацията, необходима за обработването на нулевите елементи, да се набира от специална схема.

А.2. Да се стремим да организираме преводния алгоритъм така, че нулевите елементи да се обработват в хода на осъществяването на общите анализиращи (или синтезиращи) схеми, без да се създава специална схема.

Б.1. Информацията, необходима за обработване на нулевите елементи, да се набира в хода на анализа.

Б.2. Тази информация да се набира в хода на синтеза.

Б.3. Да се комбинират тези две възможности (Б.1 и Б.2).

В.1. Информацията, необходима за обработването на нулевите елементи, да се набира на равнището на текста чрез първата от разгледаните в т. 2.2 процедури (повърхностна процедура).

В.2. Тази информация да се набира на по-дълбинни равнища на структурата на синтактичния или семантичния извод на равнището на езика чрез втората, разгледана в т. 2.3 процедура (дълбинна процедура).

В.3. Да се комбинират тези две възможности (В.1 и В.2).

**3.1.1.** Сега ще разгледаме по-подробно тези възможности, за да може да изясним по-добре към какво се свежда работата, а, от друга страна, и това е същественото за по-нататъшното ни изложение, да покажем факторите, от които в последна сметка зависи изборът на една от тези възможности (или тяхното комбиниране).

**3.1.1.0.** Следният пример ще ни покаже към какво се свежда същността на първите две алтернативи (А.1 и А.2). Да предположим, че машината трябва да преведе от руски на български разглежданото вече изречение

# этот наш друг очень способный студент #,

в което, както знаем, при съпоставянето му с неговия български образ се появява нулев елемент, чийто референт е спомагателният глагол. Ако нашият алгоритъм е основан върху лексико-морфологическа база, не е мъчно да се види, че нито анализът на всички основи, нито анализът на суфиксите, взети поотделно в рамките на всяка словоформа, не ще може (без специална сигнализация в речника и „трансформиране“ на по-дълбинните трудности в по-повърхностни — вж. § 4) да даде възможност да се набере достатъчно информация, за да се установи, че във входното изречение има нулев елемент, нито пък да се идентифицира неговият референт и да му се подбере необходимото преводно съответствие (образ). С оглед на това се налага създаването на специална схема „Нулев елемент“.

Да предположим сега, че същото входно изречение трябва да се преведе с алгоритъм, основан върху структурно-синтактична база. Както следва от разгледаните в предишния параграф примери, разпознаващият модел на синтактичния анализ ще изяви (след реконструкция) на едно от равнищата на структурата на извода референта на нулевия елемент от нашия пример, а синтезиращите правила ще генерират необходимото преводно съответствие въз основа на получената по този начин IIN. Тук,

както виждаме, обработването на нулевия елемент, който има референт на синтактичното равнище, се извършва от самите анализиращи и синтезиращи правила, без да е необходимо съставянето на специална схема. Същото нещо би могло да се покаже и за един нулев елемент, който има референт само на семантичното равнище: при алгоритми, основани на лексико-морфологична и синтактична база, би била необходима допълнителна схема, а при алгоритъм, основан на семантична база, такава схема не би била нужна.

Както виждаме, изборът на едната от двете възможности (А.1 или А.2) зависи в последна сметка от типа на преводния алгоритъм.

Извършеният при разглеждането на тези примери анализ позволява да се формулира следното основно твърдение (А): специална схема за обработване на нулевите елементи е необходима тогава, когато преводният алгоритъм е основан върху анализ, който се извършва на по-повърхностно равнище от онова, на което принадлежи референтът на съответния нулев елемент, и обратно — специалната схема не е необходима в онези случаи, при които обработваните нулеви елементи имат референти, принадлежащи на същите или на по-повърхностни равнища от онова, на което е изграден преводният алгоритъм.

**3.1.1.1.** Както беше посочено, вторият основен въпрос, който трябва да се реши при организирането на автоматичното обработване на нулевите елементи, се свежда към това, дали необходимата за тази обработка информация да се набира в хода на анализа (Б.1) или в хода на синтеза (Б.2), или пък да се комбинират тези две възможности (Б.3).

Не е мъчно да се види, че този въпрос се поставя само при случая А.1, т. е. тогава, когато преводният алгоритъм е изграден на такова равнище, което налага поради твърдението (А) от т. 3.1.1.0 създаване на отделна схема (или правила) „Нулев елемент", защото при втората от тези две алтернативи (А.2) нулевият елемент се обработва „между другото" в хода на реализирането както на анализа, така и на синтеза на съответния алгоритъм.

Двете възможности се свеждат към следното: необходимата информация за обработването на нулевите елементи може да се набира от специална схема не само в хода на анализа (Б. 1), както това беше изложено в т. 3.1.1.0, но и при синтеза (Б.2). В този случай генерирането на българските съответствия, получени чрез лексико-морфологичен анализ без специална схема „Нулев елемент", на компонентите на нашето руско изречение би дало

\# този наш познат много способен студент #

При такава организация на преводния алгоритъм специалната схема „Нулев елемент", включена в неговата синтезираща част, би трябвало да установи дефектността на получения превод, т. е. в нашия пример да установи факта, че в това българско изречение няма глагол, да определи мястото му и да избере необходимото лице, време, число и пр. Приемането на едната от тези две възможности или тяхното комбиниране (Б. 3) също така зависи от типа на преводния алгоритъм, от това, дали изобщо в него центърът на тежестта е сложен върху анализа или върху синтеза, както, разбира се, и от спецификата на съотношението на езиците, от който и на който се превежда.

3.1.1.2. Последният въпрос, който трябва да се реши при организирането на обработката на нулевите елементи, е за това, на какво равнище се набира необходимата в случая информация (алтернативи В.1, В.2 и В.3). Ясно е, че и този въпрос се поставя също така само при А.1, защото само при тази алтернатива се създава специална схема „Нулев елемент", която трябва да набира съответната информация.

Двете възможности, които съществуват в дадения случай (или тяхното комбиниране), се свеждат към следното. На първо място (В.1), както това беше илюстрирано от нашия пример, разгледан в т. 3.1.1.0, схемата „Нулев елемент" може да набира необходимата информация на същото равнище, на което е изграден анализиращият алгоритъм (в нашия пример това беше на лексико-морфологично равнище). За разлика от това възможен е и следният вариант (Р.2): схемата „Нулев елемент" да набира необходимата информация за обработването на съответните елементи на по-дълбинно равнище от равнището, на което е изграден анализиращият алгоритъм. Така например в нашия случай, докато анализиращият алгоритъм е изграден на лексико-морфологично равнище, схемата „Нулев елемент" би могла да се изгради на синтактично равнище въз основа на един ограничен, насочен синтактичен разпознаващ модел.

Такива са основните въпроси, които се поставят при алгоритмизирането на процеса на превода и изобщо при автоматичното третиране на информация, зададена във формата на естествените езици, във връзка с организацията на обработването на нулевите елементи.[30] Както видяхме, приемането на едни или други от посочените възможности зависи от типа на преводния алгоритъм, при който трябва да се решат тези проблеми. Следователно този избор се предопределя от избора на типа на преводния алгоритъм. А този избор от своя страна се обуславя (наред с редица други фактори, като лингвистическа школа в дадена страна, традиция, достигнато равнище на работата по МП, утилитарна или теоретическа насоченост и пр.) и от следните два основни фактора: спецификата на езиците, с които се борави, и приетата концепция за оптималната стратегия при МП.

Както във всичко изложено досега, така и в по-нататъшното изложение ние правим едни или други предложения, излизайки от спецификата на руския език (и на първо място от неговия синтетичен характер, богата морфология и сравнително постоянен ред на думите поне в научните текстове) и неговото съотношение с българския език. С оглед на това тук няма да разглеждаме въпроса за влиянието на спецификата на езика, който се обработва, върху избора на типа на преводния алгоритъм, а ще се спрем само върху втория основен фактор, предопределящ този избор — концепцията за оптималната стратегия при МП.[31]

Наред с решаването на посочените основни проблеми, предопределящи стратегията на обработването на нулевите елементи, се поставят и редица други твърде интересни теоретически и практически въпроси: дали синтактичните структури с нулев елемент да се разглеждат като ядрени, при което образът (съответствието) на нулевия елемент да се получава в трансформа, или обратно; на кое равнище на извода е най-рационално да се появява (реконструира) референтът на нулевия елемент и пр.

[31] Една концепция за оптимална стратегия при МП се отнася и изобщо до автоматичното третиране на информация, зададена във формата на естествените езици, защото, както се сочи в [73], в последна сметка основните въпроси, които се поставят при автоматизирането на процесите на научно-техническата информация — автоматично класифици-

3.2. Доскоро проблемите за възможните стратегии при МП и за оптималността на тези стратегии не се разглеждаха нито в независим, нито в съпоставителен план въпреки тяхното първостепенно значение както за организирането на цялата работа по МП и за насочването на съответните лингвистически изследвания, така и изобщо за осъществимостта на МП[32] и въпреки обстоятелството, че всяка система на МП почива върху една или друга, осъзната или не стратегия. Това, между другото, се дължи на обстоятелството, че засега теорията на МП все още не е достигнала необходимата степен на обобщеност.

Проблемите на стратегията при МП станаха предмет на задълбочено обсъждане за пръв път на симпозиума „Машперевод — 67" на страните, членки на СИВ, състоял се през октомври 1967 г. в Будапеща (вж. [69]). Включването в дневния ред на симпозиума на тези проблеми не трябва да се свързва с излизането през 1966 г. на доста нашумелия тогава доклад на Комитета ALPAC[33], а се дължи на обстоятелството, че теоретическите

---

ране, селекциониране, индексиране, извличане, реферииране, редактиране и пр., — се свеждат към онези основни лингвистически проблеми, които се поставят и при автоматизирането на процеса на превода.

[32] Да напомним игнорирания понякога факт, че логически е неправилно да се дискутират възможностите на МП априорно и изобщо, както това се прави в [74], [75] и др.; от същия недостатък страдат и изводите на доклада на ALPAC (вж. по-долу). Тези възможности трябва да се обсъждат първо в рамките и с оглед на дадена стратегия и едва въз основа на това да се правят обобщени изводи.

[33] Комитетът ALPAC (Automatic Language Processing Advisory Committee — вж. [76]) беше създаден през 1965 г. от Американската академия на науките, за да проучи състоянието и перспективите на работата по МП в САЩ. Въз основа на анализ на обикновената („човешка") преводаческа практика и на получените от машина преводи комитетът стигна до следните основни изводи:

МП не може да се използва в областта на художествената литература и на особено отговорни дипломатически документи, а въпросът, дали да се използва той в бъдеще в областта на научно-техническата литература, трябва да се реши въз основа на съпоставянето на обикновения и машинния превод по следните три основни показателя: качество, бързина и стойност;

засега (има се пред вид 1964 г. — А. Л.) в САЩ не е достигнат качествен икономичен самостоятелен МП на широки класове научни текстове;

качеството на получените от машината до края на 1964 г. самостоятелни преводи (т. е. без пред- и постредактиране) е такова, че те са с 10 % по-малко точни, четат се с 21 % по-бавно и дават с 29 % по-ниска степен на разбиране, отколкото преводите на същите текстове, направени от преводачи хора; при постредактиране на същите преводи, получени от машината (което прави цялата процедура по-бавна и по-скъпа от обикновения превод), съответните цифри са 3, 11 и 13 %;

МП, който засега не е оправдал практическите надежди, е оказал неоценима услуга за развитието на математическата (computational) лингвистика в най-широкия смисъл на този термин и за задълбочаването на нашите теоретически знания за естествените езици и поради това, въпреки че според някои е съмнително дали изобщо някога ще може да се постигне висококачествен самостоятелен икономичен МП, изследванията в тази област трябва да продължат ако не в името на непосредствената практика, то в името на науката; при това основното внимание трябва да се насочи към широки изследвания в областта на теоретическата и математическата лингвистика, която направи революция в езикознанието, и на теорията на преводa (структурни изследвания, пораждащи модели, семантика).

Като оставим настрана специфичните за САЩ условия и икономическите съображения, които наложиха своя отпечатък върху доклада, не може да не се съгласим с общите му препоръки. Констатацията обаче, че „засега" (т. е. към края на 1964 г.) практически не е достигнат качествен самостоятелен МП, може би вярна за САЩ, противоречи на резултатите, постигнати в Гренобъл; освен това в научно отношение тази констатация не е убедителна, тъй като тя не почива върху анализа на стратегиите, на теоретическите основи и на самите алгоритми, въз основа на които са били получени преводите, разгледани от комитета.

изследвания и практическата работа по МП и МЛ в съответните страни достигнаха такова равнище на развитие, което налага преоценка на някои положения (тук ние не сме в състояние да излагаме и анализираме възгледите, изразени в докладите, изнесени на този симпозиум — вж. критическия обзор в [78]).

Когато се говори за стратегия при МП, обикновено не се прави разлика поне между следните три неща: 1) общи възгледи за МП и лингвистически концепции, от които се излиза, както и редица организационни, икономически и практически съображения; 2) съвкупност от лингвистически съображения и методи за най-рационално, ефективно и просто извличане от текста на необходимата преводна информация и генериране въз основа на нея на изходен текст; 3) математическо представяне на лингвистическите модели, които се използват при анализа и синтеза при МП. В по-нататъшното изложение стратегията при МП ще се разбира главно в смисъла на т. 2).

Проблемите на стратегията при МП са неизбежно свързани с въпроса за нейната оптималност, който пък от своя страна е свързан с проблемата за критериите на оптималността на тази стратегия. Доскоро и тази проблема не е била обсъждана в теоретическата литература по МП. Повечето изследователи и автори на конкретни алгоритми за МП мълчаливо са излизали и излизат от предпоставката, че основни критерии за оптималността на алгоритъма и следователно за стратегията при МП, въз основа на която той е съставен, се свеждат към неговата простота (без да се уточнява това понятие), ефективност и евентуално икономичност. Без да се впускаме в обсъждане на проблемата за критериите на оптималността на стратегията при МП, което трябва да бъде предмет на отделно изследване, ще отбележим, че според нас на сегашния етап от развитието на работите по МП и на лежащите в тяхната основа лингвистически теории към ефективността и простотата (както дедуктивна, така и индуктивна[31]) трябва да се прибави и следният критерий: доколко моделите на анализа и синтеза при МП, изградени върху дадена стратегия, описват езиковата дейност на човека преводач, т. е. имат евристично значение.

Предложената от нас селективна стратегия при МП и изобщо при автоматичната обработка на информация, зададена във формата на естествените езици, чиито основи ще бъдат изложени по-долу в т. 3.2.3, се основава върху общотеоретическите възгледи на автора за превода и машинния превод, върху изградения въз основа на тях селективен модел на превеждащата операция и изводите от теоретическата и практическата работа по МП на групата „Машинен превод и математическа лингвистика“ при Математическия институт на БАН (3.2.1), както и върху един критичен анализ на някои от основните положения (сумирани в 3.2.0) на общоприетите стратегии при МП, направен в тази светлина (3.2.2).

**3.2.0.** Тъй като стратегиите, приети от различните групи по МП, зависят на първо място от съответните лингвистически концепции и общите възгледи за същността и целите на МП (т. е. онова, което обе-

[31] В работите, посветени на логиката на науката, под дедуктивна простота се разбира сравнителната простота на еквивалентни описания, а под индуктивна простота — способността на даден модел да описва с помощта на по-ограничен понятиен код по-широка област от факти и да прониква по-дълбоко в непознатото (вж. например [9])

динихме в т. 1) на стр. 40), изявяването на техните съществени елементи предполага, от една страна, обобщаване и анализ именно на тези концепции и възгледи, а, от друга, съпоставителна преценка на конкретните системи на МП, изградени върху тези стратегии. Такъв анализ се извършва в [10] и [49] и тук[35] ще бъдат изложени само получените въз основа на него обобщения и изводи за съществените елементи на основните съвременни стратегии при МП:

1. Признаване на принципната осъществимост на МП (противно мнение вж. в [37] и [74]) при определени типове ограничения на входните текстове, адекватни на ограниченията, които се срещат в определени класове натурални текстове, на приоритета на така наречения 100% подход и на необходимостта моделите на анализа и синтеза при МП да се изграждат въз основа на моделите на естествените езици, разработени от алгебричната лингвистика.

2. Намаляване на относителния дял на задаваната в речниците и списъците информация за сметка на алгоритмичното набиране на тази информация предимно във фазата на анализа.

3. Преминаване от безконфликтния към „конфликтния" метод и множествения синтез, от алгоритмичната форма на описание на отделните етапи, от комплексни алгоритми (включващи и правилата на съответните граматически модели) към алгоритми като правила за боравене с правила (т. е. „двучастни" и „тричастни" алгоритми — вж. интересния обзор в [79]).

4. Разпространение на схващането, че висококачествен адекватен самостоятелен МП не може да бъде осъществен без последователно пренасяне на значенията на всички компоненти от всички равнища на входното съобщение на семантичното равнище (за разлика от това някои групи, като например в Ленинград, Берлин, Тексас и др., смятат, че е възможен пълноценен „граматически" МП, а пък други, като например проф. Бар Хилел (вж. [37]), намират, че висококачествен самостоятелен МП е невъзможен без извънлингвистичен анализ и без машината да усвои така наречения faculté de langage).

5. Разбиране на езика посредник като инструмент за записване на инварианта, т. е. на смисъла на превода; разглеждане на семантичното равнище на естествените езици като език посредник при МП и неговото представяне като универсален метаезик със свои единици и синтаксис.

6. Стремеж да се разработва „вътрешен", т. е. независим анализ, ориентиран само към езика посредник (но тъй като самият този език посредник се изгражда с оглед на особеностите на определен кръг от езици, независимостта на анализа става относителна).

7. Признаване на необходимостта да се семантизират изявените при синтактичния анализ структури на входното изречение, а, от друга страна,

[35] Тези работи се основават върху анализ на основните публикации и експериментирани алгоритми на следните групи по МП в Европа и САЩ: I Лаборатория по МП при I МГПИИЯ в Москва, Секторът по структурна лингвистика на Езиковедския институт на АН на СССР; групата по МЛ и МП при Изчислителния център на Ленинградския университет; групата по МП при Изчислителния център на Арменската АН; Отделението за математическо и приложно езикознание и машинен превод при Германската академия на науките; групата по МП и МЛ при Изчислителния център на Карловия университет в Прага; групата по МП при Изчислителния център на Унгарската академия на науките; Centre d'Etude pour la traduction automatique (CETA) в Гренобъл; групата по МП при Bunker-Ramo Corporation в САЩ, както и на групите по МП и МЛ при Масачусетския технологически институт, Харвардския и Калифорнийския университет.

да се инвентаризират онези случаи на полиструктурност, които не могат да бъдат разрешени без отчитане на семантичната компонента (вж. например [57]).

8. Представяне на процеса на анализа (а в обратен ред и на синтеза) като съвкупност от последователни междуравнищни преминавания от най-повърхностното равнище към равнището на дълбинните смислови структури.

9. Стремеж процесът на анализа (както и на синтеза) да се възпроизвежда с помощта на взаимосвързани относително обособени (кибернетически) системи (вж. [80]), състоящи се от отделни модели, описващи последователните равнища на езика, при което (с изключение на първия и на последния) изходът на модела от по-повърхностното равнище е вход на модела от по-дълбинното равнище (и обратно). При това сложността на моделите (а следователно и на типовете анализ) нараства успоредно с преминаването на по-дълбинни равнища.

10. Тенденция на единиците от всяко равнище да се приписват всички възможни в логиката на дадените модели сегментации, структури и интерпретации (наблюдава се и противоположна тенденция — процесът на МП да се организира така, че в хода на синтактичния анализ на работното изречение да се приписва само една структура—групите в Ленинград, Берлин и др.).

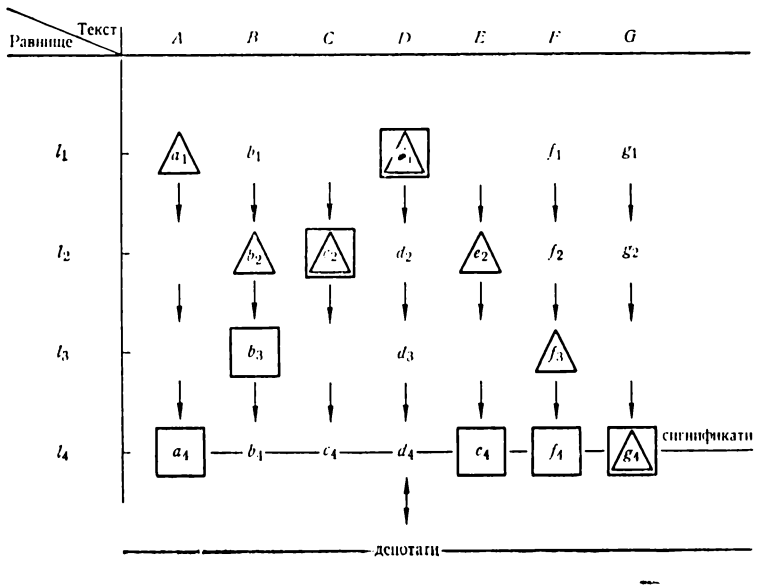
11. На всяко равнище анализът да се води само в рамките на единиците от това равнище и само чрез съответен тип анализ и всички нерешени на това равнище трудности да се пренасят на по-дълбинни равнища (в някои системи, например в Ленинградската, се допуска и контекстуален анализ на всяко равнище).

Основните положения на тези стратегии и на първо място тези, отбелязани в т. 1, 4, 8, 10 и 11 (както между другото и редица отрицателни съждения за възможности на МП — вж. например [37]), са логическо следствие на това, че техните автори, съзнателно или не, излизат от такива концепции за същността на лингвистическия механизъм на обикновения превод, които водят до модел на превеждащата операция, който може да се нарече „тотален модел“. Във връзка с това, без да съм в състояние да разглеждам подробно този въпрос тук (вж. например [30]), ще отбележа накратко следното. В съответствие с традиционните възгледи съответните автори приемат, че процесът на превода протича по схемата текст — смисъл — текст. Извличането на смисъла е разбиране, което според тях предполага, както това е и при едноезичната комуникация, еднозначно установяване не на означаваните страни (на сигнификатите), а на денотатите. Това от своя страна налага извода, че преводачът пренася, както това прави и адресатът на едноезичната комуникация, значенията на всички елементи от всички равнища, на входното съобщение към най-дълбинното, смисловно равнище, разрешава трудностите, възникващи на по-повърхностни равнища, като ги пренася обикновено на най-дълбинното равнище, а в редица случаи трябва да прибягва и до извънлингвистични данни. Тъкмо затова често, и то с основание се твърди, че човекът преводач установява смисъла чрез съдържателни анализи, които предполагат у него опит от действителността, общи познания и култура, познания на специалния предмет и пр.

Схематично това пренасяне на значенията на всички елементи от всички равнища на входното съобщение към най-дълбинното равнище и

разрешаване на трудностите на все по-дълбинни равнища може да се представи така, както е дадено на фиг. 13.

Тъй като при това виждане на проблемата на най-дълбинно равнище се пренасят значенията на всички елементи на входното съобщение, такъв




Фиг. 13

модел на превеждащата операция ще наричаме тотален модел, а стратегиите, изградени върху този модел — тотални стратегии (вж. например [10], [49]).

Този тотален модел, както и редица от основните положения на изградените върху него тотални стратегии (и на първо място отбелязаните в т. 1, 4 и 11) будят редица възражения (вж. по-долу т. 3. 2. 2) в светлината на един друг подход, концепция и модел на превода, обосновани в [5], [17], [28], [30] и [73], чиито основни положения в най-общи линии могат да се сведат към следното.

**3.2.1.** Работата по алгоритмизирането на всеки процес се обуславя в определена степен и от спецификата на този процес. Следователно алгоритмизирането на процеса на обикновения превод (ОП) трябва да се обуславя и от спецификата тъкмо на този процес.

Анализът на езиковата същност на ОП позволява да се формулират следните положения. Процесът на ОП се свежда към разбиране на входния текст и въз основа на това — към генериране на изходно съобщение. От интересуващото ни гледище разбирането се свежда към еднозначно установяване (избор) на значенията (означаваните страни, сигнификатите) на езиковите компоненти на входното съобщение. С оглед на това може да се каже, че във фазата на анализа процесът на превода е избор на актуалните означавани (сигнификатите) при зададени означаващи, а във фазата на синтеза — избор на актуалните означаващи при зададени означавани. Както беше посочено в началото на т. 1.1,

поради спецификата на $L_i^N$ това еднозначно установяване, т. е. този избор, а следователно и преводът (защото приемаме, че значението на един езиков израз е неговият превод с друг или други изрази) предполагат набиране на ITN. Въвеждането в теорията на превода на това основно понятие и разглеждането на процеса на анализа като процес на набиране на ITN за осъществяване на съответните избори позволяват да се установят следните характерни за езиковия механизъм на ОП моменти. Съставът и количеството на ITN са обективна величина, която варира не само при различните двойки езици, но и при различните подкодове и равнища на даден естествен език; при различни двойки езици различни са класовете обекти, за превеждането на които е необходима допълнителна информация; при различни класове обекти както от едно и също, така и от различни равнища на даден $L_i^N$ са различни начините на набирането на ITN — в едни случаи референциални (т. е. посредством съотнасяне с по-дълбинно равнище от равнището, на което принадлежи анализираната единица), а в други нереференциални (т. е. без такова съотнасяне), безконтекстни и контекстуални; ITN за едни и същи класове обекти от едно и също равнище на даден език при превод на друг даден език може да се набира на различни равнища в резултат на различни типове анализ. От тези констатации може да се направи следният основен извод: процесът на ОП (а следователно и на анализа при превода от един естествен език на друг, защото той може да се разглежда като съвкупност от последователни преводи от равнище на равнище) няма, така да се каже, непрекъснат, тотален характер, при който значенията на всички компоненти от всички равнища на входното съобщение се пренасят на най-дълбинното равнище, а селективен, подборен характер, за който е характерно на първо място следното:

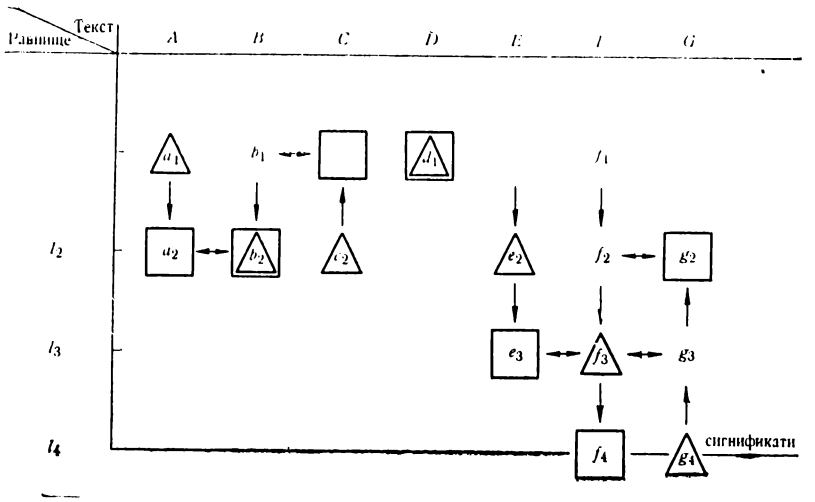
за разлика от адресанта на едноезичната комуникация преводачът не трябва да избира съобщенията, а само значенията на вече зададени означаващи;

при различни двойки езици едни и същи класове обекти от едни и същи равнища предполагат различно третиране с оглед на набиране на ITN;

при дадена двойка езици различни класове обекти от различни равнища предполагат различно третиране с оглед на набирането на ITN;

извличането на ITN, и това е най-важното, не предполага по необходимост пренасянето на значенията на всички компоненти от всички равнища на входното съобщение към най-дълбинното, смисловото, равнище, а само на някои от тях, при което даден елемент се пренася на дълбочина само до това равнище, на което е възможно извличането на ITN. Графически това е представено на фиг. 14.

Тъй като при посоченото виждане на механизма на превеждащата операция между естествени езици на най-дълбинно равнище се пренасят не всички елементи на входното съобщение, а само тези, за които това се налага от нуждите на натрупването на ITN, достатъчна и необходима за еднозначно установяване на сигнификатите, т. е. на избора на съответните значения, то позволява да се предложи за разлика от общоприетите тотални модели селективен модел на превода. Стратегия, изградена върху селективен модел, ще наричаме селективна стратегия (вж. по-долу).




Фиг. 14

3.2.2. Сега в светлината на тези базисни констатации ще разгледаме три от основните положения на общоприетите стратегии при МП.

Разпространеното разбиране на така наречения 100 % подход и становището, че МП предполага завършен математически модел на езика (langue) изобщо, практически води до поставяне на знак на равенство между обема и характера на знанията за езика изобщо, необходими от гледището на теоретическото езикознание, и обема на знанията за езика, необходими за осъществяването на превод от един конкретен език на друг, както и до поставяне на знак на равенство между целта на разпознаващите и генериращите модели и на моделите на анализа и на синтеза при МП (което между другото води до изпускане от очи и на разликата между разпознаването и разбирането, между пораждането и произвеждането, между компетентността и перформацията). Това от своя страна води до поставяне на знак на равенство между описанието на комуникацията изобщо и на процеса на превода по-специално, до игнориране на спецификата на този процес и до разтваряне на лингвистическата проблематика на теорията на превода (обикновен и машинен) в общата проблематика на алгебричната лингвистика и теоретическото езикознание.

Тъй като, както беше посочено по-горе, реализирането на процеса на МП, както и на ОП, зависи от натрупването на ITN, което предполага извършването на някакви (различни в зависимост от спецификата на съотношението на дадена двойка езици и на дадени равнища) операции над някакви (различни пак в зависимост от същата специфика) езикови обекти, то обемът на знанията за езика (а следователно и мощността на съответните модели), необходими за осъществяването на МП при дадена двойка езици, трябва да се определя не с оглед на целите на теоретическото езикознание, а тъкмо с оглед на спецификата на съответните операции и обекти. От това следва и обстоятелството, че моделите на анализа и синтеза при МП трябва да се различават по своя основен тип, мощност,

логика и цел от общите разпознаващи и генериращи модели на алгебричната лингвистика. А това налага съответно изменение на разбирането на .00 % подхода при MII (вж. също [77]) и преодоляване на парадоксалното положение, че почти всички, които искат да моделират процеса на превода, не анализират спецификата на този процес, а спецификата на процеса на езиковата комуникация въобще.[36]

Тенденцията да се пренасят и прекодират, минавайки последователно от равнище на равнище, всички единици от всички равнища на входното съобщение към най-дълбинно равнище, която допринася за решението на основната проблема на теоретическото езикознание — моделиране на механизма на свързването на значението със звука, не е адекватна на механизма на процеса на превода, който, както беше посочено по-горе, не протича по тази тотална схема и при който се пренасят не всички, а някои компоненти, при което варира и дълбочината на равнището, до което те се пренасят при различните двойки езици. Това показва, че вижданото на анализа като строга последователност от прекодирания и пренасяния на значенията на всичките елементи от всички равнища на входното съобщение до най-дълбинното равнище не е обусловено от спецификата на самия процес на превода, а е привнесено отвън.[37]

Извършването на анализ само в рамките на единиците от всяко равнище противоречи на цикличния и подборен характер на анализа при обикновения превод. Не е мъчно да се покаже, че процесът на анализа е съвкупност от процеси на превод. Щом като това е така, то и тези процеси имат също така подборен, цикличен характер, от което следва, че и анализирането на всяка единица от дадено равнище само в рамките на самата тази единица без контекстуален анализ и пренасянето на всички други нерешени трудности на по-дълбинни равнища не отговаря на реалното протичане на тази фаза от процеса на превода.

**3.2.3.** Въз основа на всичко изложено по-горе основните предпоставки на селективната стратегия, която предлагаме (по-подробно вж. [49]), могат да се формулират така:

1. Езиковият механизъм на процеса на превода от един естествен език на друг не е адекватен на езиковия механизъм на „еднозначната" комуникация.

2. Проблемата на превода като междуезикова и вътрешноезикова трансформация трябва да се разглежда като дедуктивна проблема и цялата работа по алгоритмизирането на процеса на превода да се води дедуктивно, излизайки от предварително определения състав и количество

[36] Разбира се, в идеалния случай моделът на процеса на превода би трябвало да се получава дедуктивно от модела на комуникацията, осъществявана от двуезичния субект. Обаче с оглед на съвременното състояние на нашите знания за езика и за процеса на превода този път на алгоритмизиране е невъзможен.

[37] Това между другото може да се илюстрира със следния пример. Почти общоприето е становището, че при MII word by word, основан върху контекстуален анализ на лексико-морфологическо равнище, не се отчитат дълбинните синтактични и семантични връзки. Но това становище не може да се приеме, тъй като анализирането на контекста дори на равнището на речта е синтактичен анализ par excellence, отчитащ и тяхната обусловена от семантиката валентност. Разликата се заключава в това, че в дадения случай тези дълбинни връзки се установяват и описват на по-повърхностно равнище и не в термините на познатите синтактични модели. Обаче не е необходимо да се доказва, че начинът на установяването и описанието на дадени връзки и отношения в никой случай не влияе върху характера на самите тези връзки и отношения.

на ITN, необходима за средствата на дадена двойка езици. Работата по МП над други двойки езици ще води до постепенното увеличаване и уточняване на ITN, а оттук и на категориите на езика посредник.

3. При алгоритмизирането на процеса на превода на всички равнища трябва да се изхожда от неговата езикова селективна същност.

4. При решаването на проблемата на МП, основавайки се върху постиженията на съвременната лингвистика („традиционна", структурна и математическа), да се изхожда не от обема на знанията за езика изобщо необходим от гледището на задачите на теоретическото езикознание, и да не се поставя знак на равенство между генериращите и разпознаващите модели на алгебричната лингвистика и моделите на анализа и синтеза при МП, а от онези знания първо за два конкретни езика и спецификата на тяхното съотношение, необходимостта от които се обуславя от нуждите на организирането на натрупването на ITN при дадена двойка езици и приетите ограничения на класовете входни текстове.

5. Автоматизирането на процеса на превода, както и на всяка обработка на информация, зададена във формата на естествените езици, предполага като conditio sine qua non алгоритмизирането на натрупването на ITN. При това са възможни две принципно различни решения: или машината да се „научи", подобно на човека, да върши съдържателни анализи с отчитане на извънлингвистичната действителност, или чрез специално описание тези съдържателни анализи да се представят в достъпна за нея форма.

6. Реализирането на първата от тези две възможности, към което в същност води приемането на тоталния модел и на тоталната стратегия, по необходимост предполага такова изменение и усъвършенстване на АСМ, че те да са в състояние да боравят с извънлингвистични данни, денотатите, да имат опит от действителността, така наречената faculté de langage и пр., което поне в едно обозримо бъдеще поставя непреодолими прегради пред развитието на МП.

7. Реализирането на втората от тези възможности, т. е. трансформирането на съдържателните анализи във формални, които не предполагат боравене с денотатите, а само със съответни описания на означаващите страни на сигнификатите и тяхната синтагматика, което се предполага от селективната стратегия и селективния модел на превеждащата операция, изисква не невъзможното при днешните условия усъвършенстване на АСМ, а възможно при същите условия подобрение на езиковите описания и на математическото моделиране на езиковите процеси. От всичко това между другото следва, че бъдещето на МП зависи не от невъзможното поне засега „очовечаване" на АСМ и тяхното „снабдяване" с faculté de langage, а тъкмо от усъвършенстването на езиковите описания (по-подробно вж. [13]).

8. Основният тип, логиката и мощността на отделните модели, описващи различните равнища на езика (както и мощността и категориите на езика-посредник), да се определят дедуктивно, излизайки от спецификата на процеса на превода и от нуждите на натрупването на ITN, а да не се пренасят механично в готов вид от алгебричната лингвистика.

Такъв подход постепенно ще оформи предпоставките за създаване на „граматика на преводача“[38].

9. Анализът (и обратно синтезът) да се моделира като система от междуравнищни преводи и да се води не само на „дълбочина“ и на „широчина“, но и „нагоре“, като всеки един от неговите етапи трябва да бъде подчинен на принципа на селективността. При това:

а) трудностите от по-дълбинните равнища да се трансформират дотолкова, доколкото това е възможно, в трудности от по-повърхностни равнища и следователно да се разрешават чрез по-прости типове анализ (вж. § 4);

в) да се пренасят на всяко по-дълбинно равнище чак до най-дълбинните не всички елементи на входното съобщение, а само тези, за които на дадено равнище не може да се извлече ITN при дадена двойка езици, и само в тези случаи, когато тези трудности не могат да бъдат разрешени на даденото равнище нито в резултат на анализ в рамките само на единицата от това равнище, нито чрез контекстуален анализ на това равнище, нито пък трансформирани в трудности от по-повърхностни равнища.

10. Селективното пренасяне на по-дълбинно равнище само на онези елементи или комбинации от елементи, от които не може да бъде извлечено достатъчно количество ITN на дадено равнище и следователно приписването им само на онези „трудности“, които пречат на това набиране, обуславя постепенно „освобождаване“ на входното съобщение от по-голямата част от компонентите от по-горните равнища. Това от своя страна обуславя възможността да се създават за всяко равнище *насочени* вместо цялостни интегрални модели.

Както следва от изложеното, наред с утвърждаването на селективната идея и осъзнаването на необходимостта от трансформиране на съдържателните анализи във формални един от най-съществените моменти на предложената стратегия е трансформирането на трудности от по-дълбинни равнища в трудности от по-повърхностни равнища и тяхното разрешаване чрез по-прост тип анализ, характерен за това равнище. Практическото реализиране на тази възможност при алгоритмизирането на процеса на превода, както и изобщо при автоматично обработване на информация, зададена във формата на естествените езици, осигурява (както това се потвърждава не само от логическо гледище, но и от резултатите, получени от нашата експериментална работа), не само по-големи възможности за осъществяване на МП и за преодоляване на неразрешими при другите стратегии трудности, но и приближава неговите модели до дейността на човека преводач. Илюстрация за практическо реализиране на тази възможност при автоматичното обработване на $L_i^N$ се дава в следващия параграф.

**§ 4.** В този заключителен параграф след съвършено кратко описание на разработваната от групата по МП и МЛ при Математическия институт на БАН система на МП на руски математически текстове на български език (4.1) ще приведем един пример на пренасяне на трудности,

[38] Логически тази възможност (и необходимост) се потвърждава от опитите, посветени на създаването на „граматика за слушащия“. Обаче тази идея трябва да бъде предмет на по-нататъшни изследвания.

свързани с обработването на един клас нулеви елементи, от по-дълбинно на по-повърхностно равнище и съответните начини за тяхното разрешаване (4.2).

**4.1.** Системата на машинен превод на руски математически текстове на български език, която се разработва в Математическия институт на БАН, се изгражда върху предложения по-горе селективен модел и стратегия и се води дедуктивно от предварително установения състав и количество на ITN (вж. например [28]). В най-общи линии тя се характеризира със следното. Основната част от ITN се набира в хода на анализа. Общият зависим модел на анализа включва интегрален семемен лексически и морфологически анализ в рамките на работното изречение (чиито схеми са готови и минават експериментална проверка) и насочени модели на синтактичен и семантичен анализ (чиито схеми се разработват). Входният текст се подава на машината изцяло и автоматично се сегментира на изречения преди началото на анализа в собствения смисъл на думата въз основа на отчитане само на някои формални елементи от равнища $l_1$ и $l_2$. Търсенето в лексемния и морфемния речник осигурява сегментирането на текстовите единици (словоформи) от работното изречение на семи- и контекстуален анализ на същите равнища, отстранява възможните неправилни сегментации. Схемите на алгоритъма за лексически анализ, работещи по сигнализация от съответните кодове на лексемния и морфемния речник, и на информация, натрупана на лексико-морфологическо равнище, разрешават на тези равнища и следните трудности: сложни думи, контактни и неконтактни устойчиви съчетания, лексическа многозначност, омонимия на словесните класове. Схемите на морфологическия модел, изградени и заработващи по същия начин, снемат на равнище $l_2$ вместо на равнище $l_3$ по-голямата част от омонимията и многозначността на граматическите форманти. Схемите на насочения синтактичен и семантичен анализ трябва да решат нерешените в хода на работата на предишните два модела трудности и на първо място да снемат останалата полиструктурност и да решават проблемите на глаголните времена, наклоненията и артикъла.

При тази организация на анализа ние имаме пренасяне на трудности от по-дълбинни на по-повърхностни равнища в следните случаи: при установяването на границите на работните изречения (на графемно и морфологическо равнище вместо на синтактично); при снемането на лексическата многозначност и обработването на устойчивите съчетания (на лексико-морфологическо равнище вместо на семантично); при снемането на омонимията на словесните класове и омонимията и многозначността на граматическите форманти (предимно на морфологическо вместо на синтактично равнище). Това между другото позволява в повечето случаи предварително да се отстранят причините на синтактичната полиструктурност, която в противен случай би се появила на синтактичното равнище. Но тъй като всички тези случаи по начало не са свързани с проблемата за нулевите елементи, ние няма да ги разглеждаме, а ще се спрем по-подробно върху този начин на обработване на един тип нулеви елементи.

**4.2.** В хода на нашето изложение многократно беше споменаван примерът на руските конструкции с нулева връзка, в която при превод на български език се появява нулев елемент, чийто референт при дълбинната процедура може да се идентифицира на едното от равнищата на

структурата на синтактичния извод. Сега накратко ще опишем тази проблема в светлината на предложената селективна стратегия и пренасянето на трудности от по-дълбинни на по-повърхностни равнища.

**4.2.0.** От твърдението (А), формулирано в края на т. 3.1.1.0, следва, че специална схема „Нулев елемент“ е необходима тогава, когато преводният алгоритъм се изгражда върху анализ, който се води на равнище, по-повърхностно от равнището, на което принадлежи референтът на нулевия елемент. От това следва, че обработването на нулев елемент, чийто референт принадлежи на синтактичното равнище (както е в нашия пример с конструкциите с нулева връзка), може да се провежда без специална схема, ако анализиращият алгоритъм включва и модел на цялостен синтактичен анализ. Но тъй като нашият алгоритъм включва само насочен синтактичен анализ, от твърдението (А) следва, че за обработването на нулевите елементи от разглеждания клас в нашата система би трябвало да се създаде специална схема „Нулев елемент“.

Обаче изложената в края на предишния параграф възможност трудностите от по-дълбинно равнище да се трансформират в трудности от по-повърхностно равнище и да се решават чрез анализ, характерен за това по-повърхностно равнище, открива още един път: от трудност от синтактично равнище проблемата, свързана с обработването на разглеждания клас нулеви елементи, да се трансформира в трудност от морфологическо равнище и да се решава на лексемно-морфологическо равнище като снемане на полисемията „непредикативно/предикативно“ значение чрез контекстуален анализ на това равнище. При такъв подход няма да бъде необходимо създаването на специална схема „Нулев елемент“ (тъй като проблемата се решава от общата схема „Многозначност“), а, от друга страна, остава вярно твърдението (А).

**4.2.1.** Това възможно решение на проблемата беше предложено въз основа на инвентаризиране и анализ на всички изречения с този тип нулеви елементи в един corps de texte и на установяване на формалните начини за снемане на съответната многозначност, в която се трансформира проблемата на нулевата връзка.

**4.2.1.0.** Анализът на нашия corps de texte (книгата „Теория чисел“ от А. А. Бухщаб, М., 1964, 360 стр.) показа, че от всичките описани в нормативните граматики на руския език случаи на изразяване на именната част на съставно сказуемо с нулева връзка (при случаите с тире има формален показател и следователно те са особен подслучай от интересуващия ни тип) в него се срещат следните: именната част е изразена със съществително, с кратко прилагателно, с кратко причастие, с пълно прилагателно, с пълно причастие, с прилагателно в проста форма на сравнителна степен, с прилагателно в сложна форма на сравнителна степен и с прилагателно в проста форма на превъзходна степен.

Тъй като представителите на всички словесни класове, които в нашия corps de texte се използват като изразители на именната част на съставно сказуемо с нулева връзка (съществителни, прилагателни и причастия, а това се отнася и за другите възможни класове), освен „предикативно“ значение, в което те функционират в случая, могат да имат и „непредикативни“ значения, установяването на факта, че в даденото изречение има нулев елемент (т. е. процедурата за сигнализиране) трябва да

се основава върху някакви формални критерии от лексемно-морфологическото равнище на съответните изречения.

Анализът на редуцираните структурни типове на тези конструкции показва, че съответните условия или критерии, които трябва да бъдат налице в работното изречение (тук под работно изречение се разбира просто изречение), могат да се формулират така:

1. Структурен тип $X^{39}\not\varnothing s$ (пример: $\varphi$ ***константа***, *которая появляется во всех подобных случаях*; брой на появи — 5).

Условия: а) отсъствие на глагол;

б) две последователни именни групи в именителен падеж.

2. Структурен тип $X\not\varnothing A^2$ (пример: *Число таких столбцов по определению* ***равно*** $\varphi$; брой на появи — 194).

Условия: а) отсъствие на спомагателен глагол;

б) наличност на кратка форма на прилагателно.

3. Структурен тип $X\not\varnothing A^r$ (пример: *В силу свойства транзитивности все числа класса* ***сравнимы*** *между собой*; брой на появи — 165).

Условия: а) отсъствие на спомагателен глагол;

б) наличност на кратката форма на причастие.

4. Структурен тип $X\not\varnothing A_1$ (пример: *В 1739 г. Эйлер показал, что это число* ***составное***, *и тем самым опроверг гипотезу Ферма*; брой на появи — 38).

Условия: а) отсъствие на глагол;

б) постпозиция на пълното прилагателно ($S_1A_1$).

5. Структурен тип $X\not\varnothing A'_1$ (пример: *Эти условия* ***выполненные***; брой на появи — 1).

Условия: а) отсъствие на глагол;

б) постпозиция на пълното причастие ($S_1A'_1$).

6. Структурен тип $X\not\varnothing A^{c1}$ (пример: *Любая четная дробь* ***меньше*** *любой нечетной*; брой на появи — 8).

Условия: а) отсъствие на глагол;

б) постпозиция на прилагателното в сравнителна степен ($S_1A^{c1}$).

7. Структурен тип $X\not\varnothing A^{c2}$ (пример: ***Более удобен*** *способ отсеивания составных чисел, известный еще греческому математику Эратосфену*; брой на появи — 5).

Условия: а) отсъствие на глагол;

б) наличност на словоформите *более* или *менее* преди $A^{c2}$.

8. Структурен тип $X\not\varnothing A^{s1}_1$ (пример: *Если 10 — преобразованный корень по модулю p, то длина периода* ***наибольшая***; брой на появи — 1).

Условия: а) отсъствие на глагол;

б) постпозиция на прилагателното в проста форма на превъзходна степен ($SA^{s1}_1$).

---

[39] С $X$ се бележи частта на изречението преди (т. е. отляво на) $\varnothing$. Тя може да бъде и празна. Словесните класове са означени по следния начин: $S_i$ — съществително, $A_i$ — прилагателно в пълна форма, $A'_i$ — причастие в пълна форма, $A^r$ — прилагателно в кратка форма, $A^{r'}$ — причастие в кратка степен, $A^{c1}$ — прилагателно в проста форма на сравнителна степен, $A^{c2}$ — прилагателно в сложна форма на сравнителна степен, $A^{s1}_i$ — прилагателно в проста форма на превъзходна степен, $\varphi$ — формула.

4.2.1.1. Анализът на изложените по-горе условия, при наличността на които може формално и еднозначно да се определи, че в дадения случай конкретният представител на даден словесен клас има предикативно значение, т. е. изпълнява функцията на именна част на съставно сказуемо с нулева връзка, показва, че те всички се свеждат към наличността или отсъствието на езикови факти от съответни равнища (принадлежност към даден клас, наличност или отсътствие на представител на даден клас, позиция, падеж), които могат да се установяват формално, т. е. на равнището на означаващите чрез непосредствен (въз основа на данни от речниците) и контекстуален анализ в рамките на съответното просто изречение на лексемното и морфологическото равнище. Установяването на факта, че дадена компонента на руското изречение има предикативно значение, позволява да се снеме съответната многозначност и при синтеза да се вземе българското съответствие със спомагателен глагол (чието лице и число се определят от допълнителен анализ на същите факти). Тези две процедури снемането на многозначността „предикативност/непредикативност“ и вземането на необходимото преводно българско съответствие по своя резултат са еквивалентни на процедурите на сигнализирането и генерирането от специална схема „Нулев елемент“, както и на резултата от идентифицирането на референта на съответния нулев елемент на равнището на структурата на синтактичния извод чрез процедурата на дълбинния анализ. Това илюстрира изложената по-горе възможност трудностите, свързани с обработването на $\varnothing$, чийто референт принадлежи на по-дълбинно равнище, да се трансформират в трудности от по-повърхностно равнище, което позволява тяхното разрешаване чрез анализ, характерен за това равнище. При това необходимо е да се отбележи, че както дълбинната процедура, така и процедурата, свеждаща се към пренасяне на съответните трудности на по-повърхностно равнище, са разработени само за научно-технически текстове.

*

Изложеното в § 2 показва какви големи възможности открива представянето на дълбинните равнища на езика чрез моделите на алгебричната лингвистика, включващи и семантичните компоненти, както за дълбинната процедура при идентифицирането на референтите на нулевите елементи, така и изобщо за решаването на проблемите на МП. От друга страна, съображенията, въз основа на които се предлага селективната стратегия при МП, изложена в § 3, показват, че анализът, провеждан на по-повърхностни равнища, е не само по-прост, но и моделира по-отблизо функционирането на езиковите механизми на превеждащата операция. Установяването на рационално съчетание между тези две тенденции е една от основните задачи, към които трябва да се обърне теорията на МП и изобщо на автоматичното третиране на естествените езици.

## ЛИТЕРАТУРА


1. Федоров, А. В. Введение в теорию перевода. II изд. М., 1958.
2. Vinay J. P., A. Darbelnet. Stylistique comparée du français et de l'anglais. Paris, 1958.
3. On Translation. Cambridge, Mass., 1959.
4. Wojtasewicz, O. Wstęp do teorii tłumaczenia. Warszawa, 1957.
5. Людсканов, А. Към въпроса за предмета на общата теория на превода. Год. на Соф. унив., Фил. фак., **56** (1963), 123—196.
6. Ревзин, И. И., В. Ю. Розенцвейг. Основы общего и машинного перевода. М., 1964.
7. Рецкер, Я. И. О закономерных соответствиях при переводе на родной язык. Вопросы теории и методики учебного перевода. М., 1950.
8. Деже, Л. Машинный перевод русских конструкций с глаголами на -ся на венгерский язык. Доклад на V съезде славистов, София, 1963. Славянска филология, **7** (1965), 386—387.
9. Reichenbach, G. Experience and prediction. Chicago, 1938.
10. Людсканов, А. Уровни языковой структуры и проблема оптимальности стратегии при машинном переводе. „Машперевод-67". Будапешт, 1969.
11. Логическая семантика и модальная логика. М., 1967.
12. Bunge, M. Les concepts de modèle. L'âge de la science, **3** (1968), 165—140.
13. Людсканов, А. Някои основни лингвистически и математически проблеми на автоматичната обработка на информация, зададени във формата на естествените езици. Списание на БАН, 1970 (под печат).
14. Падучева, Е. В. Международная конференция по семиотике в Польше. Научно-техническая информация, **2** (1967), 35—44.
15. Шаумян, С. К. Структурная лингвистика. М., 1965.
16. Тезисы конференции по машинному переводу. Ереван, 1967.
17. Людсканов, А. За предмета, мястото и методологията на общата теория на превода. Дисертация, т. I и II. София, 1963.
18. Людсканов, А. Проблема на преводимости в свете современной лингвистики. Научно-техническая информация, **2** (1967), 28—33.
19. Bloomfield, L. Language, II. London, 1955.
20. Mounin, G. Les problèmes théoriques de la traduction. Paris, 1963.
21. Хомский, Н. Логические основы лингвистической теории. Новое в лингвистике, **4**, (1965), 465—476.
22. Хомский, Н. Лингвистика, логика, психология и вычислительные устройства. Математическая лингвистика. М., 1964. 69—100.
23. Braithwaite, R. B. Scientific Explanation. Cambridge. Mass., 1953.
24. Хоккет, Ч. Грамматика для слушающего. Новое в лингвистике, **4** (1965), 139—167.
25. Herault, D. Eléments de théorie moderne des probabilités. Paris, 1967; вж. също Александров, П. С. Введение в общую теория множеств и функций. М. 1948.
26. Ревзин, И. И., В. Ю. Розенцвейг. К обоснованию лингвистической теории перевода. Вопросы языкознания, **1** (1962), 51—59.
27. Людсканов, А. Принципът на функционалните еквиваленти — основа на теорията и на практиката на превода. Език и литература, **5** (1958), 344—361.
28. Людсканов, А. Основи на теорията на машинния превод с оглед на руско-българския МП. Год. на Соф. унив., Фил. фак., **58** (1964), 287—508.
29. Людсканов, А. Междуязыковой и междуярусный перевод. Actes du X[e] Congrès international des Linguistes, Bucarest, 1967, **4** (под печат).
30. Людсканов, А. Превеждат човекът и машината. С., 1967.
31. Фитиалов, С. Я. О моделировании синтаксиса в структурной лингвистике. Проблемы структурной лингвистики. М., 1962. 100—113.
32. Жолковский, А. К., Н. Н. Леонтьева, Ю. С. Мартемьянов. О принципиальном использовании смысла при машинном переводе. Машинный перевод, **2** (1961), 17—47.
33. Хомский, Н. Некоторые наблюдения, касающиеся проблемы семантического анализа естественных языков. [Доклад на Международната конференция по семиотика в Полша през 1966; цит. по Научно-техническая информация, **2** (1967), 37—39.]
34. Bailly, A. I. Rouault. Un essai de formalisation de la sémantique des langues naturelles. CETA, G. 2200-A. Grenoble, 1967.

35. Жолковский, А. К., И. А. Мельчук. О множественном семантическом синтезе. Проблемы кибернетики, **19** (1967), 117—238.
36. Moreau, R. Les équilibres linguistiques. Université de Paris, A-293, 1963.
37. Bar-Hillel, J. Die Zukunft der maschinellen Übersetzung. Sprache im technischen Zeitalter, **2** (1967), 13—19.
38. Хомский, Н. Синтактические структуры. Новое в лингвистике, **2** (1962), 412—527.
39. Горский, Д. П. О видах определений и их значение в науке. Проблемы логики научного познания. М., 1964.
40. Ельмслев, Л. Метод структурного анализа в лингвистике. Хрестоматия по истории языкознания XIX—XX веков, В. А. Звегинцев. М., 1956, 103—110.
41. Vauquois, B. Le système de Traduction automatique du CETA. Grenoble, 1967.
42. Мельчук, И. А. Автоматический синтактический анализ. Новосибирск, 1964.
43. Lamb, S. M. On Alternation, Transformation, Realisation and Stratification. Mon Series on Languages and Linguistics, **17** (1964), 105—122.
44. Hjelmslev, L. La stratification du language. Word, **2—3** (1954).
45. a. Мартине, А. О книге „Основы лингвистической теории". Новое в лингвистике, **1** (1960), 437—462; b) Martinet, A. La double articulation linguistique. Travaux, du Cercle linguistique de Copenhague, **5** (1949), 30.
46. Marcus, S. Aspecte ale modelarii matematice in lingvistica. Studii şi cercetari lingvisticei, **14** (1963), No. 4, 487—501.
47. Prieto, L. J. Contribution à l'étude fonctionnelle du contenu. Travaux de l'Institut de linguistique, Paris, **1** (1956), 18—32.
48. Kurylowicz, J. Linguistique et théorie du signe. Journal de psychologie, **2** (1949), 15—26.
49. Ludskanov, A. Is the Generally Accepted Strategy of Machine Translation Research Optimal? Mechanical Translation, **11** (1968), No. 1 and 2, 14—22.
50. Gardin, J. C. Le fichier mécanographique de l'out llage. Institut français d'Archéologie Beyrout, 1956.
51. Жолковский, А. К. Предисловие. Машинный перевод и прикладная лингвистика, **8** (1964). Москва.
52. Perry, J., A. Kent. Tools for Machine Literature Searching. New York, 1968.
53. Машинный перевод и прикладная лингвистика, **8** (1964).
54. Ceccato, S. Operational linguistics and Translation. Methodos, **12** (1960); вж. също: Human translation and translation by machine. Teddington, I, 1961, National Physical laboratory, Paper 30.
55. Allani, E., S. Ceccato, E. Maretti. Classification rules and code of an operational grammar for mechanical translation. Information retrieval and machine translation, **2** (1961), 18—32.
56. Conečna, D., P. Novák, P. Sgall. Machine Translation in Prague. Prague Studies in Mathematical Linguistics, No. 3, 1966.
57. Agricola, E. Syntaktische Mehrdeutigkeit (Polysyntaktizität) bei der Analyse des Deutschen und des Englischen, Schriften zur Phonetik, Sprachwissenschaft und Kommunikationsforschung, No. 12, 1968.
58. Katz, J. J., J. A. Fodor. The structure of a semantic theory, Language, **39** (1963), 170—210; вж. също: Katz, J. J., P. Postal. An integrated theory of linguistic descriptions. Cambridge, Mass., 1964.
59. Ревзин, И. И. Метод моделирования и типология славянских языков. М., 1967.
60. Ingve, V. Sentence-for-sentence Translation. Mechanical translation, **2** (1955), No. 2, 29—37; вж. също: Синтаксис и проблема многозначности. Машинный перевод. М., 1957.
61. Oswald, O. A., S. L. Fletsher. Proposals for the mechanical resolution of German syntax patterns. The Modern language forum, **36** (1952), No. 34, 81—92.
62. Молошная, Т. Н. Вопросы различения омонимов при машинном переводе с английского языка на русский. Проблемы кибернетики, **1** (1958), 215—221; вж. също: Алгоритм перевода с английского языка на русский. Проблемы кибернетики, **3** (1960), 209—272.
63. Katz, J. J., P. M. Postal. An Integrated Theory of Linguistic Description. Research Monograph, No. 26, M. I. T., Cambridge, Mass., 1964.
64. Сгал, П. Алгебричен модел на езика. Език и литература, **2** (1967), 29—38.
65. Veillon, G., J. Veyrunes, B. Vauquois. Un métalangage de grammaires transformationnelles, CETA, G., 2300-A. Grenoble, 1967.
66. Сгалл, П. Об отношении между порождающей грамматикой и моделированием компетенции переводчика. „Машперевод-67". Будапешт, 1969.

67. Розенцвейг, В. Ю. Машинният превод и лингвистиката. Език и литература, **2** (1967), 78—86.
68. Actes de la II Conférence Internationale sur le Traitement automatique des langues. Grenoble, 1969.
69. „Машперевод-67", Материалы Симпозиума по МП стран-членов СЭВ. Будапешт, 1969.
70. Варга, Д. Стратегия при машинном переводе. „Машперевод-67", Будапешт, 1969.
71. Уорс, Д. С. Трансформационный анализ конструкций с творительным падежом в русском языке. Новое в лингвистике, **2** (1962), 637—683.
72. Smits, Ph. H. English article insertion, General Analysis Technique Russian-English. Research Reports. p. 19. W. D. C., 1959.
73. Людсканов, А. Некоторые лингвистические проблемы автоматизации процессов НТИ. Материалы симпозиума „Комплексная механизация и автоматизация процессов поиска, обработки, передачи на расстояние и выдачи научно-технической информации". М., 1968, 436 -438.
74. Бар-Хилел, И. Будущее машинного перевода. Филологические науки, **4** (1962), 205—213.
75. Шрейдер, Ю. А. Машинный перевод иллюзии и реальность. „Машперевод-67". Будапешт, 1969.
76. Hays, D. Language and Machines. Computers in Translation and Linguistics. New York, 1966.
77. Сепс, Д. Относительно теории грамматики в свете машинного перевода. „Машперевод-67". Будапешт, 1969.
78. Людсканов, А. Въпроси на стратегията при машинния превод и теорията на формалните граматики. Език и литература, **3** (1968), 121—126.
79. Garvin, P. Machine, Translation Fact or Fancy? Datamation Magazine, **3** (1967), 1—12.
80. Греневский, Г. Кибернетика без математики. М., 1964.
81. Sorensen, H. S. Word class in modern English. Copenhague, S. E. C., Gad., 1958.
82. Cantineau, J. Les oppositions significatives. Cas. Fds., **10** (1952), 84—131.

*Постъпила на 6. II. 1968 г.*

# К ВОПРОСУ О „НУЛЕВЫХ ЭЛЕМЕНТАХ" ПРИ МАШИННОМ ПЕРЕВОДЕ И АВТОМАТИЧЕСКОЙ ОБРАБОТКЕ ЕСТЕСТВЕННЫХ ЯЗЫКОВ

Александр Людсканов

*(Резюме)*

Всеизвестно, что при переводе (как обыкновенном — ОП, так и машинном — МП) с одного естественного языка $L_i^N$ на другой $L_j^N$, в выходном тексте (в качестве такого принимается предложение $Z$) появляются элементы, неимеющие формально заданных образов (соответствий) на графемическом (фонетическом) уровне входного $Z$, т. е. элементы, являющиеся образами (соответствиями) чего-то, что не задано на этом уровне входного $Z$. Это незаданное на определенном уровне входного $Z$ „что-то" и будем называть „нулевым элементом" $\phi$. Рассмотрению некоторых общих теоретических аспектов проблемы $\phi$, которая до сих пор, насколько нам известно, не была предметом обобщенных исследований (и имеет и общелингвистическое значение), и обоснованию процедуры их автоматической обработки в свете предложенной автором селективной стратегии при МП и посвящена настоящая работа.

В п. 1.2 на основе мотивированной в других работах автора общей семиотической концепции перевода и определения таких понятий как образ (соответствие), прообраз (референт), функциональный эквивалент, необходимая переводная информация (ITN), анализ и синтез (вкратце излагаемых в п. 1.1) понятие $\not\varnothing$ определяется следующим образом.

Пусть заданы две последовательности (предложения) терминальных символов $Z_1$ и $Z_2$, принадлежащих соответственно $L_i$ и $L_j$:

$$Z_1 \quad \#a \;\; b \qquad c \;\; d \qquad e \;\; f \;\; g \qquad h \,\#,$$
$$Z_2 \quad \#a_1 \;\; b_1 \;\; A_1 \;\; c_1 \;\; d_1 \;\; B_1 \;\; e_1 \;\; f_1 \;\; g_1 \;\; C_1 \;\; h_1\#.$$

Примем, что $Z_2$ является отображением (переводом) $Z_1$ в $L_j$, а $Z_1$ — отображением (переводом) $Z_2$ в $L_i$. Тогда, рассматривая $Z_2$ и сопоставляя его с его отображением в $L_i$, можно утверждать, что $A_1$, $B_1$, $C_1$ являются формально заданными на графическом уровне $L_j$ элементами, имеющими пустые образы (соответствия) в $L_i$; и наоборот, рассматривая $Z_1$, сопоставляя его с его отображением в $L_j$, можно утверждать, что в нем существуют $\not\varnothing^1$, $\not\varnothing^2$ и $\not\varnothing^3$, имеющие формально заданные соответствия $A_1$, $B_1$, $C_1$ на графемическом уровне $L_j$. Эти пары можно представить следующим образом:

$$Z_1 = \#a \;\; b \;\; |\overline{\not\varnothing^1}| \;\; c \;\; d \;\; |\not\varnothing^2 \;\; e \;\; f \;\; g \;\; |\overline{\not\varnothing^3}| \;\; h \,\#,$$
$$Z_2 = \#a_1 \;\; b_1 \;\; |A_1| \;\; c_1 \;\; d_1 \;\; |B_1| \;\; e_1 \;\; f_1 \;\; g_1 \;\; |C_1| \;\; h_1\#.$$

На основании этого вводится следующее определение $\not\varnothing$ в отношении естественных языков: нулевой элемент ($\not\varnothing$) на данном уровне $l_i$ данного $L_i^N$ является эксплицитно незаданным образом (соответствием), референт которого задан или на том же уровне $L_x$, или на более глубинном уровне ($l_x$) $L_i^N$.

В п. 1.3 рассматриваются некоторые общие свойства введенного понятия и в первую очередь проблема прообраза (референта) $\not\varnothing$ в свете выведенного из дефиниции перевода положения, что в выходном (переводящем) $Z$ не может появиться никакой новой информации, незаданной в входном сообщении. Лингвистическая практика показывает, что референт $\not\varnothing$ может быть задан или где-то в последовательности линейного контекста $Z$ на уровне речи (parole) вне этого $Z$, или где-то на более глубинных уровнях этого $Z$. В работе рассматривается только вторая альтернатива. Если при принятом отображении $Z_1$ на $Z_2$ и обратно исходить из $Z_2$, то, очевидно, что $\not\varnothing^1$, появляющийся в $Z_1$, является результатом операции $A_1 \rightarrow \not\varnothing^1$, что представляет собой общий случай и не требует комментариев. Однако, если исходить из $Z_1$ и рассматривать $A_1$ в $Z_2$, который появляется в результате операции $\not\varnothing^1 \rightarrow A_1$, то ясно, что $\not\varnothing^1$ должен имплицитно содержать или „отражать“ информацию, эквивалентную $A_1$, и, следовательно, операции $\not\varnothing^1 \rightarrow A_1$ должна предшествовать операция $A_1 \rightarrow \not\varnothing^1$. Этот элемент $A$, по необходимости заданный на каком-то более глубинном уровне представления $Z_1$ и будем называть прообразом (референтом) $\not\varnothing^1$. Из этого следует, что с предметно-лингвистической точки зрения проблема обработки $\not\varnothing$ сводится в первую очередь к проблеме идентификации их референтов, что предполагает определенную процедуру.

В п. 2.1. дается краткое описание некоторых основных положений математического моделирования $L_i^N$ и теории формальных грамматик, а в п. 2.2 проводится критический анализ существующих процедур идентификации нулевых элементов — контекстуального лингвистического анализа на уровне речи (parole) и „внелингвистического" анализа, устанавливается субъективный и интуитивный характер некоторых из их механизмов и невозможность их формализации и автоматизации. В начале п. 2.3 выявляется одно из их основных свойств формальных синтаксических моделей — их аксиоматический характер, и на основании этого делается вывод, что ни в терминальной цепочке, ни на отдельных уровнях структуры вывода не может появляться ничего, что не задано на более глубинных уровнях или в аксиомах и правилах вывода. Благодаря этому в ходе распознавания предложений данного $L_i^N$ при помощи моделей этого типа в некотором узле на некотором уровне структуры вывода должен выявляться и референт $\not\varnothing$, принадлежащий некоторому более поверхностному уровню. И именно это дает возможность выявить „глубинную" процедуру „реконструкции" и идентификации референта $\not\varnothing$. В отличие от традиционного контекстуального анализа, идущего „вширь" по уровню текста и анализа, направленного на „вне"-лингвистические факты, обсуждаемая процедура направлена на более глубинные уровни языка (langue) и допускает алгоритмизацию и машинную реализацию. Этот логический вывод иллюстрируется анализом структур выводов и графов зависимостей нескольких синтаксических моделей (при НС-грамматике типа CF, грамматике зависимостей (2.3.2.0), при данной Т-грамматике (2.3.2.1), в системе CETA (2.3.2.2), а также перехода от глубинной к поверхностной структуре в семантических порождающих моделей (2.4.3).

Пункт 3.1 посвящен формулировке основных лингвистических проблем, которые ставятся при создании схемы автоматической обработки $\not\varnothing$ при МП. Их можно представить в виде следующих альтернатив: (А) создавать особую схему „нулевой элемент" (НЭ) или организовать анализирующий алгоритм таким образом, чтобы информация, необходимая для обработки $\not\varnothing$, накапливалась в ходе общего анализа; (В) накапливать необходимую информацию в ходе анализа или в ходе синтеза; (С) эту информацию накапливать на уровне текста или на более глубинных уровнях рабочего $Z$. Анализ этих альтернатив приводит к формулированию следующего основного положения: особая схема НЭ необходима тогда и только тогда, когда переводящий алгоритм основан на анализе, который ведется на более поверхностном уровне, чем уровень, которому могут принадлежать самые глубинные референты данного класса $\not\varnothing$. А из этого в свою очередь следует, что выбор между указанными альтернативами зависит в последнюю очередь от общей организации анализа и соответствующей стратегии при МП.

В п. 3.2.0 обобщаются основные элементы общепринятых стратегий ведущих групп по МП в Европе и США, а в п. 3.2.2 проводится критический анализ некоторых из их основных положений и устанавливается их „тотальный" характер, выражающийся в первую очередь в том, что при нем все элементы (всех уровней) входного сообщения переносятся на самый глубинный уровень и ряд других дискуссионных свойств.

Вместо этих общепринятых тотальных стратегий в п. 3.2.1 излагаются лингвистические предпосылки стратегии иного типа, а в п. 3.2.3

автор формулирует основные элементы предлагаемой им выборочной (селективной) стратегии при МП (как и вообще при автоматической обработке информации, заданной в форме $L^N$). Оптимальная стратегия при МП должна обуславливаться спецификой лингвистических механизмов, действующих при осуществлении ОП, которая в первую очередь проявляется в необходимости накопления ITN. Ввиду специфики средств $L^N$ в конечном итоге анализ сводится к выбору актуальных означаемых (синификатов) при заданных означающих, а синтез — к выбору актуальных „означающих" при заданных означаемых. Эти выборы, как и любые другие, предполагают накопление ITN, которую в принципе (ввиду той же специфики $L_i^N$) нельзя извлечь только из каждого рабочего средства в отдельности. В результате анализа лингвистических особенностей процесса этого накопления формулируются следующие положения: ITN является объективной величиной, варьирующей при различных парах $L^N$; при разных парах $L^N$ различны и классы объектов, для перевода которых необходима дополнительная информация и варьирует и количество этой ITN; при различных классах объектов как одного и того же уровня, так и различных уровней данного $L^N$ различны и способы накопления ITN — в одних случаях референциальные (т. е. посредством перехода на более глубинные уровни), а в других нереференциальные (т. е. без таких переходов) анализы, как и содержательные, контекстуальные, глубинные и внелингвистические анализы. Из этого следует, что процесс ОП имеет выборочный, циклический, а не тотальный характер, при котором не все объекты входного $Z$ переносятся на более глубинные уровни, а только некоторые из них, причем глубина перенесения и тип анализа обуславливается возможностями накопления ITN.

Эта селективная (выборочная) модель процесса ОП и лежит в основе предлагаемой выборочной стратегии при МП, для которой в отличие от тотальной стратегии характерными являются следующие положения: языковой механизм, реализующий процесс перевода при естественных языках, не адекватен языковому механизму одноязычной коммуникации (это, с одной стороны, позволяет уточнить соотношение между алгебраической лингвистикой и теорией МП, а с другой — высказать гипотезу, что подобно выделению грамматик для говорящего и слушающего при одноязычной коммуникации должна существовать грамматика (в понимании Н. Чомского и для переводчика); тип, логика и мощность моделей анализа и синтеза при МП должны определяться, исходя из специфики накопления ITN при данных парах языков, а не привноситься механически tel quel из алгебраической лингвистики; анализ при МП следует вести не только „вширь" и „вглубь", но и „вверх", трансформируя трудности более глубинных уровней в трудности (ambiguïtés) более поверхностных уровней; переносить на более глубинные уровни не все элементы и трудности входного $Z$, а только те, которые не могут быть решены ни контекстуальным анализом на данном уровне, ни трансформированы в трудности более поверхностных уровней. Человек накапливает ITN в результате содержательных анализов, а ЭВМ может проводить лишь формальные анализы, следовательно, алгоритмизация накопления ITN — предпосылка осуществления МП и любой автоматической обработки информации, заданной в в форме $L^N$ — предполагает трансформирование содержательных анализов в формальные посредством математического моделирования $L^N$. Из этого

в свою очередь следует, что будущее МП зависит не от „очеловечивания" ЭВМ и приобретения ими „faculté de langage", а от усовершенствования языковых описаний.

В последнем параграфе описывается на базе опыта работы группы „Машинный перевод и математическая лингвистика" Математического института Болгарской академии наук и анализа corps de texte в 360 стр. практическое применение одного из самых характерных элементов предложенной стратегии — трансформирования трудностей более глубинных уровней в трудности более поверхностных уровней при создании процедуры для обработки $\emptyset$ класса „нулевая" связка именного сказуемого.

## ON THE "ZERO ELEMENTS" IN CASE OF A MACHINE TRANSLATION AND AUTOMATIC PROCESSING OF NATURAL LANGUAGES

Aleksandâr Lyudskanov

*(Summary)*

It is known that when translating (both in case of a human translation — HT, as well as in case of a machine translation — MT) from one natural language $L_i^N$ into another $L_j^N$ in the source text (the sentence $Z$ is assumed to be the source text in the present paper) we find elements which have no images (correspondences) formally given on the graphemic (phonetic) level of $Z$, i. e. elements representing images (correspondences) of something that has not been given at this level of the source sentence $Z$. That component which has not been given on a certain level of the source sentence $Z$ we shall call a "zero element" $\emptyset$. The present study deals with the examination of some general theoretical aspects of the problem of the zero elements which as far as our knowledge goes has not been a subject of any investigation or generalization (although it is of a general linguistic importance) as well as with the argumentation of a procedure for their automatic processing in the light of a selective strategy of machine translation proposed by the author.

In paragraph 1.2 on the basis of the general semiotic conception of translation motivated in other papers of the author and on the basis of the definitions of such concepts as image (correspondence), referent, functional equivalent, necessary translation information (ITN), analysis and synthesis (whose concepts are briefly given in 1.1) the concept zero element is defined through the following deliberations:

Let be given two sequences (sentences) of terminal symbols $Z_1$ and $Z_2$ belonging respectively to $L_i$ and $L_j$:

$$Z_1 = \#a\ \ b\ \ \ \ c\ \ d\ \ \ \ e\ \ f\ \ g\ \ \ \ h\ \#$$
$$Z_2 = \#a_1\ \ b_1\ \ A_1\ \ c_1\ \ d_1\ \ B_1\ \ e_1\ \ f_1\ \ g_1\ \ C_1\ \ h_1\#$$

We assume that $Z_2$ is a reflection (translation) of $Z_1$ in $L_j$ and $Z_1$ is a reflection (translation) of $Z_2$ in $L_i$. Next when examining $Z_2$ and juxtaposing it to its reflection in $L_i$ we may assert that $A_1$, $B_1$, $C_1$ are formally given on

the graphemic level of $L_j$, elements which have their free images (correspondences) in $l_i$ and vice versa examining $Z_1$ and juxtaposing it to its reflection in $L_j$ we may assert that it contains $\not\varnothing^1$, $\not\varnothing^2$ and $\not\varnothing^3$ which have correspondences $A_1$, $B_1$, $C_1$ formally given on the graphemic level of $L_j$. These pairs may be represented in the following way:

$$\begin{array}{lcccccccccccccc} Z_1 & \# a & b & \not\varnothing^1 & c & d & \not\varnothing^2 & e & f & g & \not\varnothing^3 & h & \# \\ Z_2 & \# a_1 & b_1 & A_1 & c_1 & d_1 & B_1 & e_1 & f_1 & g_1 & C_1 & h_1 & \# \end{array}$$

On the basis of these considerations the following definition of a zero element is introduced for the natural languages: a zero element $\not\varnothing$ on a certain level $l_i$ of a given $L_i^N$ is the image (correspondence) not given explicitly whose referent is given either on the same level of $L_x$ or on a deeper one $l_x$ of $L_i^N$.

In paragraph 1.3 some general properties of the concept introduced are discussed, especially the problem of the zero element referent. This is done in the light of the statement deduced from the definition of translation which reads that the output sentence $Z$ (the translated one) cannot contain any new information which has not been implied in the input message. The linguistic practice shows that the referent of the zero element can be given either somewhere in the sequence of the linear context of $Z$ on the level of parole but out of $Z$ or somewhere on the deeper levels of $Z$. Only the second alternative is considered in the present paper. If with the accepted reflection of $Z_1$ into $Z_2$ and vice versa we originate from $Z_2$ then obviously the zero element appearing in $Z_1$ will be a result of the operation $A_1 \rightarrow \not\varnothing^1$ which is the normal case and no comments should be made. But if we begin with $Z_1$ and examine $A_1$ in $Z_2$ appearing as a result of the operation $\not\varnothing^1 \rightarrow A_1$, then it is evident that $\not\varnothing^1$ should implicitly contain or "reflect" an information equivalent to $A_1$ and therefore the operation $\not\varnothing^1 \rightarrow A_1$ should preceed the operation $A \rightarrow \not\varnothing^1$. This element $A$ which by necessity should be given on a deeper level of the representation of $Z_1$ we shall call a referent of the zero element. This implies that the problem of processing $\not\varnothing$ is reduced primarily to the problem of identifying their referents which on its turn presupposes a certain procedure.

Paragraph 2.1 contains a short description of some basic assumptions of mathematical modelling of $L^N$ and of the theory of formal grammars. In paragraph 2.2 a critical analysis of the procedures available for identifying the zero elements is carried out, namely a contextual linguistic analysis on the level of parole and an extralinguistic analysis, the subjective and intuitive character of some of their mechanisms is shown and it is established that they are not subjected to formalization and automation. Paragraph 2.3 begins with a discussion of one of the basic properties of the formal syntactic models, i. e. their axiomatic character and on its basis the following conclusion is drawn: anything not implied in the deeper levels or not given by the axioms and rules of derivation is impossible to appear either in the terminal string or on the separate levels of the structure of derivation. As a result of this when identifying the sentences from a given $L_i^N$ using models of this type the referent belonging to a level nearer to the surface should appear on a certain knot of a given level of the structure of derivation. This enables us to formulate the principles of a procedure for reconstruction and iden-

tification of the referents of $\varnothing$. In contrast to the traditional contextual analysis which is carried out "in width" on the text level and in contrast to the analysis based on extralinquistic factors the procedure under discussion is directed „inwards" towards the deeper levels of language and allows an algorithmization and machine realization. The logical conclusion is illustrated through an analysis of the structures of derivation as well as through an analysis of the graphs of dependences of some syntactic models (those of the grammar of immediate constituents of CF type and of the grammar of dependences (2.3.2.0), of a T-grammar (2.3.2.1) and the CETA's system (2.3.2.2)) and in the process of transition from a deeper to a nearer to the surface structure in the case of a given semantic generating model (2.4.3).

In paragraph 3.1 the basic linguistic problems which we face when creating a scheme for automatic processing of $\varnothing$ in case of a machine translation are formulated. They may be represented as the following alternatives: (A) to creat a special scheme "zero element" or to organize the analysis algorithm so that the information necessary for the processing of $\varnothing$ to be collected in the process of the general analysis; (B) the necessary information to be collected either in the process of analysis or in the process of synthesis; (C) this information to be collected on the text level or on the deeper levels of the working $Z$. Analysing these alternatives we come to the following basic conclusions: a special scheme "zero element" is needed when and only when the translation algorithm is based on an analysis carried out on a level nearer to the surface than the level to which the deepest referents of a certain class of $\varnothing$ may belong. This on its turn implies that at the end our choice of an alternative will depend on the general organization of the analysis and on the corresponding machine translation strategy.

In paragraph 3.2.0 the basic principles of the strategies generally accepted by the leading groups working in the field of machine translation in Europe and USA are generalized and in paragraph 3.2.2 a critical analysis of their basic principles is made and their "total" character is established. What characterizes them all is the fact that the elements (all of them and on all levels) of the input message are transferred to the deepest level. They have also some disputable properties.

Paragraph 3.2.1 contains the logical prerequisites for a strategy of new type which could substitute the generally accepted strategies. In paragraph 3.2.3 the author formulates the basic principles of a selective strategy of machine translation (as well as in case of automatic processing of data given in the form of $L_i^N$) proposed by him. The optimal strategy of machine translation should be determined taking into consideration the specificity of the linguistic mechanisms operating in case of a human translation, in the first place the necessity of collecting ITN. With a view to the specificity of the means of $L^N$ it may be said that the analysis is reduced to a choice of actual significants when given the determinants while the synthesis is reduced to a choice of actual determinants when the significants are given. These kinds of choice presuppose a collection of ITN which in principle (again due to the specificity of $L^N$) cannot be derived from a single working element. As a result of the analysis of the linguistic peculiarities of this process of collecting information the following postulates are formulated: ITN is an objective quantity which varies with the different language pairs of $L^N$; with the different language pairs $L^N$ the classes of objects for whose translation addi-

tional information is necessary are different and the amount of this information varies too; with different classes of objects belonging either to one and the same level or to different levels of a given natural language the ways of collecting ITN are different — in some cases referential analyses (i. e. by means of referring them to a deeper level than that to which the unit under consideration belongs), in others — non-referential (i. e. without such reference) or any kind of analysis: formal, contextual, extralinquistic, etc.

It follows from this that the process of human translation has a selective cyclic but not global character and with it not all objects from the source $Z$ should be transferred to deeper levels but only some of them. The depth of transition as well as the type of analysis are determined by the existing possibilities for collection of ITN.

This selective model of the machine translation process serves as a basis for the proposed selective strategy of machine translation and in contrast to the global strategy it is characterized by: the linguistic mechanism by means of which the process of translation is realized in case of natural languages is not identical to the linguistic mechanism of the monolingual communication (this conclusion helps us on the one hand to establish the correlation between the algebraic linguistics and the machine translation theory and on the other hand to make the following supposition: since there exist grammars for the speaker and for the hearer with the monolingual communication there should be a grammar (in the sence of Chomsky) for the translator too); the type, the logic and the applicability of the analysis and synthesis models for machine translation should be determined having in mind the specificity of collecting ITN in case of a given pair of languages and they should not be transferred mechanistically tel quel from the algebraic linguistics; the machine translation analysis should be carried out not only "in width" and "in depth" but also "upwards" thus difficulties of the deeper levels should be transformed into difficulties (ambiguïtés) closer to the surface; not all elements and difficulties of the source sentence $Z$ should be transferred to the deeper linguistic levels; but only those which can be neither solved by means of a contextual analysis on a given level nor can be transformed into difficulties of levels closer to the surface. Man collects ITN as a result of content analysis while a computer may do only a formal analysis. It follows from this that the algorithmization of the collection of ITN which is a prerequisite for a machine translation realization as well as every kind of automatic processing of data given in the form of $L^N$ presupposes a transformation of the content analysis into a formal one through mathematical modelling of $L^N$. It follows from this that the future of the machine translation does not depend on the „humanization" of the computers and on their acquiring a "faculté de langage" but entirely on the improvement of the linguistic description.

In the last paragraph the practical application of one of the most characteristic elements of the strategy proposed, namely the transformation of the difficulties of the deeper levels into difficulties of levels closer to the surface when creating a procedure for processing of $\phi$ from the class of "zero" connection in the nominal predicate is described on the basis of the experience of the group dealing with machine translation at the Mathematical Institute of the Bulgarian Academy of Sciences and on the analysis of a text of 360 pages.