ДВУЕЗИЧЕН ВХОД ПРИ АВТОМАТИЗИРАНИ СИСТЕМИ ЗА ТЪРСЕНЕ НА ИНФОРМАЦИЯ

Александър Людсканов

Както е добре известно (вж. напр. [1, 2, 3]), непосредствено свързаното със започналото в началото на втората половина на XX в. невиждано развитие на науката и техниката и огромно нарастване на обема (както по броя на изданията, така и на техните езици) на научно-техническата литература[1] доведе до следния парадокс: развитието на науката и техниката и разгръщането на научно-техническата революция обуславят нарастването на обема на научно-техническата информация; по-нататъшното развитие на науката и техниката по необходимост предполага все по-голяма осведоменост на творците на научния и техническия прогрес; но колкото повече нараства обемът на научно-техническата литература, толкова по понятни причини достъпът до нея става все по-мъчен и по-бавен, а понякога и невъзможен и следователно осведомеността на специалистите, която би трябвало да нараства успоредно с нарастването на обема на съответната информация, фактически относително намалява.

Не е мъчно да се докаже, че това относително намаляване на осведомеността на учените и специалистите ще се засилва пропорционално на нарастването на обема на съответната научно-техническа информация.[2] И както с основание отбелязва съветският специалист В. В. Косолапов, ако в тази област към 1980 г. не се постигне принципен прелом, учените и специалистите ще могат да използуват само около 1% от съществуващата в съответните области информация.

---

[1] Книги, статии, доклади, стандарти, патенти и пр. — нататък ще ги наричаме просто документи.

[2] Според изследванията на проф. К. Лески [4] товз нарастване има приблизително следния вид:

Година	Брой на публикациите	Брой на печатните страници
1960	2 500 000	16 000 000
1965	3 300 000	22 500 000
1975	9 000 000	56 000 000
1985	20 000 000	150 000 000
2000	57 500 000	370 000 000

След осъзнаването на това положение на нещата от специалистите както в социалистическите страни, така и на Запад, започнаха усилия за излизане от тази задънена улица на първо място по две линии — организационна и методическа. Усилията по първата линия се насочиха към издаване на специални справочници, реферативни списания, анотации, към създаване на информационни служби и центрове (отраслови и национални)[3], а по втората — към въвеждане на принципно нови методи и средства и на първо място към използуване на автоматични сметачни машини (АСМ) и автоматизиране на процесите на научно-техническата информация (НТИ). Една от съществените форми на реализирането на тази тенденция стана изграждането на автоматизирани системи за търсене на информация от библиотечен тип (АСТИ; кратък преглед за изграждането на такива системи вж. в [5]; вж. също [6, 7, 8]). АСТИ се изграждат за дадено хранилище (библиотека, патентно, стандартно бюро и пр.) и са предназначени да дават автоматично на интересуващите се (потребителите) фактологически отговори на техните запитвания относно документите от хранилището.[4]

Създаването на една АСТИ представлява извънредно сложна комплексна проблема (вж. [12]) и предполага решаването на голям брой основни лингвистически, логически, математически, кибернетически, програмистки, технически, организационни, икономически и други въпроси (вж. [5, 9]).[5] Очевидно е, че тук не съм в състояние, а и не си поставям за цел да изброявам и разглеждам тези проблеми (частично вж. [5, 9, 13, 14]). Вместо това ще се спра върху една друга проблема, която, доколкото ми е известно, изобщо не е била поставяна и решавана по принцип в световната теория и практика в тази област, и ще предложа съответно решение. Това е проблемата за създаване на „двуезичен вход“ при АСТИ. С оглед на тази цел на настоящото изложение ние ще видим, че съществуващите АСТИ имат едноезичен вход (I) и след това въз основа на анализ на процеса на реферирането (по-специално в наши условия) при тях ще формулираме постановката на проблемата за двуезичния вход (II). В III ще бъдат разгледани логико-лингвистическите предпоставки, стандартната форма и методът на създаването на един „двуезичен българо-руски дескрипторен речник“, необходим за създаването на АСТИ с двуезичен вход, а в IV ще бъде разгледан въпросът за алгоритмично осигуряване на такива системи.

I

I.0. За да покажем какво означава, че познатите в световната теоретическа литература и практика АСТИ имат еднозначен вход, ще тряб-

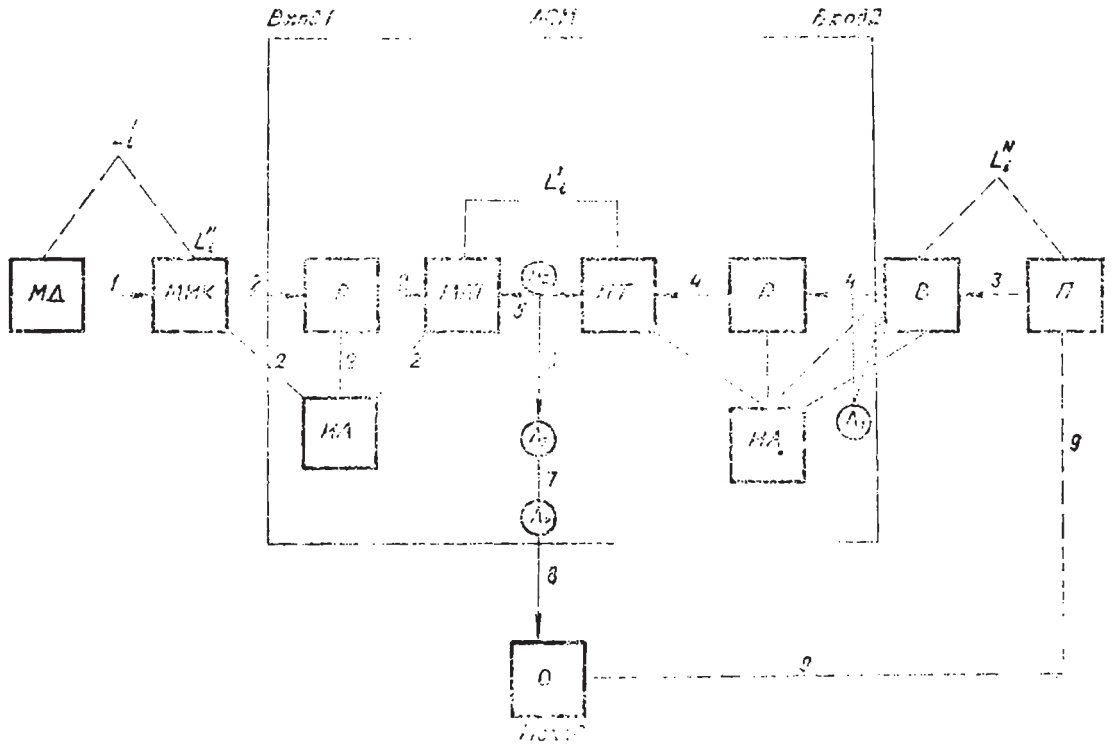
---

[3] Така у нас на съвместно заседание на Секретариата на ЦК на БКП и Министерския съвет от 24. I. 73 г. беше утвърдена програма за изграждане на Единна национална система на научна и техническа информация.

[4] Например потребителят може да задава въпроси от типа „какво е писал авторът X в областта Y“ или „какво и от кого е писано в областта X на езика L“, или „съществуват ли патенти за X“ и т. н. и системата автоматично ще му дава съответните отговори.

[5] Засега у нас са създадени два експериментална модела на АСТИ — първият през 1969/70 г. в МИ на БАН под ръководството на доц. к. ф. м. н. Д. Добрев [10] и вторият — през 1971/72 г. в Респром Радиоелектроника от колектив от представители на това звено и на МИ на БАН под ръководството на автора [11].

ва първо накратко да представим компонентите (възлите) и функционирането на една такава система. На фиг. 1 е дадена опростена логическа схема на условна АСТИ.



Фиг. 1. Автоматизирана система за търсене на информация

Както се вижда от схемата, нейните компоненти са представени в квадратчета и кръгчета, а съответните операции са показани с плътни и прекъснати номерирани стрелки. Ще смятаме, че компонентите и операциите, заградени в правоъгълника в средата на схемата, се намират и протичат вътре в АСМ, а другите — извън нея. За простота и прегледност на изложението лявата част на АСМ ще наричаме вход 1, а дясната — вход 2. Както ще проличи от по-нататъшното изложение, вход 1 има предназначението да въвежда и обработва информационните карти, а вход 2 — въпросите.

Откъм вход 1 е масивът от всички „първоначални документи", които се намират в хранилището, за което се създава дадената АСТИ. Тъй като по понятни причини АСТИ не е в състояние да борави със самите „първоначални документи, чрез операцията $\xrightarrow{1}$, която се нарича реферирание, за всеки такъв документ от МД се създава негов „представител" (информационна карта — ИК), а за целия масив от първоначалните документи респективно се създава масив от ИК (МИК); ИК съдържат два типа данни: фактологически данни за документа (т. е. информационни белези: автор, заглавие, периодично издание, патентна класификация и пр.) и текстова или свободна част, която включва резюмето на съответния първоначален документ (именно тези резюмета ще представляват основен интерес за нас). Цялата информация от ИК се

представя във формата на даден естествен език (L_1^N), който се нарича входен език на системата. Въз основа на едноезичния (вж. по-долу т. III. 0) дескрипторен речник (Р) и индексиращия алгоритъм (ИА) чрез операция $\xrightarrow{2}$, която се нарича автоматично индексиране, за всяка ИК се получава нейният автоматичен превод в информационния език (L_1^I) на системата, т. е. „лик за търсене", а за целия МИК — съответно МЛТ[6].

Системата може да бъде изградена така, че откъм вход 2 да имаме пълна аналогия с вход 1: потребителят (П) формулира ($\xrightarrow{3}$) своя въпрос (В), който трябва да бъде представен във формата на същия L_1^N, както и ИК, т. е. въз формата на същия входен език: въз основа на същия речник (Р) и на същия индексиращ алгоритъм (ИА) въпросът се индексира ($\xrightarrow{4}$), т. е. се превежда автоматично на L_1^I и се получава т. нар. предписание за търсене — ПТ (рационално е преди автоматичното индексиране на въпроса да се включи логическа схема $Л_1$, която предварително да провери неговата допустимост). След това заработват „логическите" схеми: $Л_2$ съпоставя ($\xrightarrow{5}$) по дадена стратегия за търсене ПТ, получено в резултат на индексирането на даден въпрос, с МЛТ и подбира т. нар. релевантно подмножество от ЛТ (и респективно от ИК), които се отнасят към дадени въпрос В. $Л_3$ ($\xrightarrow{6}$) извлича от това подмножество онази и само онази информация, която е необходима, за да се отговори на този въпрос, а $Л_4$($\xrightarrow{7}$) синтезира отговорите (О) и ги извежда ($\xrightarrow{8}$) за печат на изхода. Очевидно е, че отговорите трябва да бъдат представени във формата на същия естествен език, на който бяха представени ИК и въпросите, т. е. във формата на входния език на системата. Отговорите трябва да бъдат формирани така, че да се установи обратна връзка (9) с потребителя, допринасяща за неговото „обучение" и за обучението на системата.[7]

I.1. И така под входен език се разбира онзи естествен език, върху който е изградена съответната АСТИ и на който се задават ИК и въпросите и се формулират отговорите. Всички както експериментирани, така и описани в световната литература АСТИ по начало имат само един входен език — националния (в нашия случай българския). Поради това казваме, че съществуващите АСТИ имат едноезичен вход.

II

Въз основа на изложеното вече може да формулираме проблемата за двуезичния вход (II. 3), като анализираме процеса на реферирането (II. 0) при условията на изграждането на АСТИ в нашата страна (II. 1 и II. 2).

---

[6] МЛТ — масив от ликове за търсене; смятам, че този, получил известна гражданственост у нас термин „лик за търсене", буквален превод на руското „поисковый образ", не е особено сполучлив, но тук няма да се занимавам с терминологични въпроси.

[7] Понастоящем се правят твърде насърчителни опити да се въведе и диалог „машина — оператор" и „машина — потребител".

II.0. Умозрително реферирането може да бъде автоматично или ръчно. Тъй като решението на проблемата за автоматичното реферираме е въпрос на твърде отдалечено бъдеще, в наши дни практически може да се използува само ръчното реферираме. То може да бъде специално и косвено. При първото за всеки документ от първоначалния масив (МД) на хранилището специално за нуждите на АСТИ, която се разработва, се прави по едно резюме. Тези резюмета се включват в съответните ИК. Такова специално реферираме е нерационално, тъй като изисква много време и средства, като се има пред вид, че МД може да включва милиони документи. Затова в световната практика се използува предимно косвеното реферираме, при което необходимите резюмета и други фактологически данни се взимат в готов вид от реферативни списания и други източници. Това е рационално, но изисква наличието на следната предпоставка: тъй като ИК трябва да бъдат зададени във формата на входния език на АСТИ, то и готовите резюмета могат да се взимат от реферативни списания и други източници, излизащи само на този входен език, в случая на български.

II.1. Обаче практиката показа, че у нас поне засега в нашите реферативни списания за дадена специална област може да се намерят готови резюмета за не повече от 20—30% от документите от първоначалния масив на дадено хранилище. Това поставя въпроса, откъде да се вземат останалите 70—80% от резюметата. Тук очевидно пак са възможни две решения: за тях да се използува специалното реферираме или те да се потърсят в реферативни списания, излизащи на други езици. Поради изложените вече съображения първото от тези решения не е за предпочитане. Второто решение е рационално особено у нас, тъй като, както показват нашите проучвания, за почти всички документи от едно специално хранилище, за които не може да се намерят готови български резюмета, могат да се намерят готови руски резюмета. Следователно у нас косвеното ръчно реферираме, което е най-рационално и евтино, може да се осъществи в пълен обем, ако се съгласим да имаме първоначално не само български, но и руски резюмета. Въпросът ще се постави по същия начин и в други страни, например Полша, Унгария, Румъния, ЧССР, отчасти ГДР, а също така и в редица западни страни, където наред с резюметата на съответните национални езици ще се използват и английски резюмета, както и в двуезични страни, като например Канада и др. По понятни причини в нашето изложение ще говорим само за български и руски език. Но по принцип всичко, което ще бъде казано за тях, се отнася в по-голяма или в по-малка степен до която и да е друга двойка естествени езици.

II.2. Както казах, в условията на нашата страна при изграждането на АСТИ е рационално да се използува косвено ръчно реферираме, при което ще се получават в готов вид не само български, но и руски резюмета. Но това поставя нова проблема: тъй като познатите досега АСТИ имат само едноезичен вход и са изградени на съответния национален естествен език, то очевидно е, че получените по този начин руски резю-

мета ще трябва предварително да се превеждат на български език.[8] За тази нова проблема са възможни три решения: а) този превод на руските резюмета на български да се автоматизира; б) да се извършва от съответни специалисти; в) в рамките на една АСТИ да се създадат фактически две (една българска и една руска) със свои отделни речници, списъци от константи, алгоритми, програми и пр.

а) Първото решение, по принцип възможно благодарение на системата за машинен превод (МП) от руски на български, която се разработва в МИ на БАН, в обозримо бъдеще едва ли би било икономически оправдано.

б) Предварителният превод на руските резюмета на български от специалисти оскъпява съответната система с около 30% и обуславя цяла редица допълнителни трудности.

в) Недостатъците на третото решение — вместо една да се правят две системи, са понятни сами по себе си.

II. 3. Незадоволителността на тези три възможни при познатите АСТИ и условията на нашата страна решения ме доведе до следната мисъл: като се създадат някои логически и лингвистически предпоставки, не би ли могло нещата да се организират така, че при един и същ оригинално разработен двуезично-дескрипторен речник и незначителни изменения в „българския" индексиращ алгоритъм, една и съща АСТИ да може автоматично да индексира, т. е. да превежда непосредствено на информационния език на системата не само български, но и руски резюмета или съответно ИК, благодарение на което да се преодолеят всички посочени по-горе недостатъци и икономически нецелесъобразности. Такива системи ще наричаме АСТИ с двуезичен вход (по-точно вж. III, 1. 4).

Ето така се поставя проблемата за двуезичния българо-руски вход, който предлагам. Практическото реализиране на такъв българо-руски вход (по принцип нещата се поставят по същия начин при която и да е двойка естествени езици; при това може да се разработи и триезичен вход, но върху този въпрос няма да се спирам сега) предполага принципно друга организация на дескрипторния речник и съответно алгоритмично обозначение. Тези въпроси ще бъдат разгледани по-долу — в III и IV.

III

За да създадем основа за нашите по-нататъшни разсъждения, тук първо ще опишем накратко структурата на един едноезичен дескрипторен речник (III. 0), ще разгледаме логико-лингвистическите предпоставки, необ-

[8] Ще отбележим, че необходимостта от автоматизирането на процесите на НТИ постепенно налага следната тенденция: авторите на научни книги, статии и пр. не само сами да съставят резюмета на своите работи, но и да ги придружават от списъци на дескриптори. Ако някога тази тенденция се превърне в повсеместна практика, тя би могла да облекчи в значителна степен автоматичното индексиране, но не и да снеме проблемата за многоезичието при реферирането; очевидно е, че никой чужд автор — руски, английски и пр., няма да прави български резюмета и да дава български дескриптори.

ходими за създаването на двуезичен дескрипторен речник (III. 1), ще предложим стандартна форма на такъв речник и ще разгледаме методиката на неговото съставяне, както и възможностите да се автоматизират някои моменти от тази работа (III. 2).

III. 0. Въпреки че това доста често не се осъзнава, дескрипторният речник независимо от метода на неговото съставяне[9] представлява един от най-важните компоненти на всяка АСТИ. Фактически това е преводен речник,[10] който задава: елементите на входния естествен език, с които ще борави системата, т. е. множество дескриптори, информационния език на системата (като множество от кодове), отношенията между тях и информация за омонимия, многозначност, вариантност, синтагматични и парадигматични връзки (йерархични връзки като кодове от тезаурус), въз основа на която работят схемите на индексиращия алгоритъм.

III. 0. 1. За такъв обикновен едноезично-дескрипторен речник (вж. подолу III, 1. 2) на АСТИ с едноезичен вход сме предложили следната стандартна форма[11].

---

[9] Обикновено по аналогия с практиката в областта на МП и на автоматичния анализ на текстове дескрипторният речник се разработва предварително. При този подход предполагаемите в дадена специална област дескриптори се подбират въз основа на достатъчно репрезентативен corps de texte от документи и резюмета от съответната област, от терминологични речници и пр. Тези дескриптори се обработват в лингвистическо отношение (омонимия, синонимия, многозначност, вариантност, анализ „извън", синтагматични, йерархични връзки и пр.) и се попълват графите на стандартната форма на едноезично-дескрипторния речник. Специален алгоритъм въвежда този речник в машината, сегментира дескрипторите на „начала" и „опашки" и ги записва в съответните зони на паметта (вж. също III. 2. 2. 0). Такъв метод е използуван в ръководената от автора разработка на АСТИ в ЦНИРД Респром. Той осигурява възможност за напълно автоматично индексиране при вход 1.

За разлика от това в разработения в МИ на БАН експериментален вариант на АСТИ доц. Добрев предложи друг оригинален метод, който обаче налага ръчно индексиране при вход 1. Първоначално в паметта на машината не е записан никакъв речник. Един специалист в областта, за която се разработва дадената АСТИ, взема последователно всяка ИК от масива и от нейното резюме обособява съответните дескриптори и ги подава на машината която въз основа на тях създава сама словника на своя речник и приписва на всеки дескриптор някакъв код (съвкупността на тези кодове ще представлява информационният език на системата). След това този словник се въвежда за печат и друг специалист-лингвист дава необходимата лингвистическа обработка. И двата посочени метода имат свои положителни и отрицателни страни.

[10] Тези речници, които се използуват в АСТИ, са преводни, защото осигуряват езиковите предпоставки за индексирането, т. е. автоматичното превеждане на резюметата от ИК и на текстовите части на бланките — въпроси от входния естествен език, във формата на който те са зададени, на информационния език на системата. Те във всички случаи (т. е. независимо от това, какъв тип информационен език задават) са дескрипторни, защото включват и поставят в съответствие с кодовете на информационния език не всички компоненти на съответния естествен език, а само дескрипторите от дадена област на науката или техниката. Дескрипторите (които биват прости или сложни) са в същност термини или терминологични съчетания (напр. *кондензатор; генератор на честота*), които най-добре и най-пълно отразяват понятийния апарат на дадената област на науката или техниката. Тези речници са основни за всяка АСТИ. От тях трябва да отличаваме тезаурусите, в които се задават йерархическите или асоциативните връзки между значенията, понятийното съдържание на дескрипторите (вж. напр. [6, 7]). Във вид на съответни кодове информацията на тези връзки се включва в съответната графа на дескрипторния речник.

[11] За АСТИ, която разработваме за областта на радиоелектрониката, сме създали поотделно два едноезично-дескрипторни речника — един български и един руски. Всеки от тях съдържа по 2659 дескриптора (прости и сложни). Всички приведени в тази работа примери на речникови единици са взети от тези два речника.

1	2	3	4	5	6	7	8	9

Тук съдържанието на графите е следното: 1— адрес; 2 — брой на пълнозначни думи в дескриптора; 3 — основа на дескриптора (прост или сложен); 4 — структура на дескриптора; 5 — многозначност; 6 — омонимия; 7 — синтагматични връзки; 8 — парадигматични връзки (в нашия речник тази графа не е дадена); 9 — код (съвкупността от тези кодове представлява информационният език на системата).

Но такъв едноезично-дескрипторен речник, необходим и достатъчен за АСТИ с едноезичен вход, е очевидно недостатъчен за АСТИ с двуезичен вход, за която ще бъде необходим един принципно нов двуезично-дескрипторен речник (вж. също III. 1.4).

III. 1. Изграждането на такъв двуезичен (в случая българо-руски) дескрипторен речник става възможно благодарение на изявяването на следните логически и лингвистически теоретични предпоставки.

III. 1. 0. Както е известно, едноезично-дескрипторният речник приписва на всеки дескриптор (графа 3) от дадена специална терминологична област на даден естествен език някакъв цифров код в графа 9 (напр. *автомат* — 0020), а на многозначните дескриптори толкова кода, колкото значения те имат, например *база* — 0 092
— 0 093.

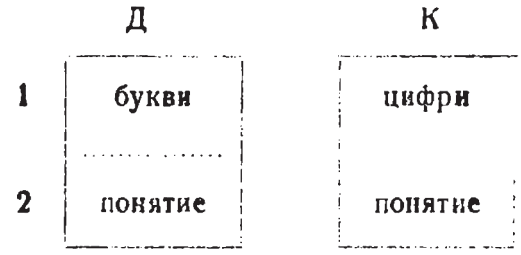
Следователно може да се каже, че в случая имаме двойка от вида Д — К (дескриптор — код). Да анализираме по-подробно двойките от този вид.

III. 1. 1. Дескрипторът (Д) представлява израз от даден естествен език, а кодът — свързан с него „израз“ от даден информационен език. На какво се основава тази връзка?

Като израз от даден естествен език или по-точно като езиков знак Д се характеризира с основното свойство на всеки знак от дадена семиотична система — да има означаваща и означавана страна. Означаващата страна на всеки Д е някаква последователност от графеми (например *т-о-к*), а неговата означавана страна или значение — дадено понятие. Но това понятие, представляващо значението на Д, е нещо идеално, което се намира в главата на човека и за да може машината да борави с него, то трябва да бъде „извадено“ от нея и представено в някаква експлицитна материална форма. Във връзка с това ще използуваме въведеното в езикознанието от Р. О. Якобсон (вж. [16]) определение, че значението (или по-точно описанието на значението) на един езиков знак е неговият превод с друг знак или други знаци. Следователно, за да „извадим“ значението на Д от главата на човека и за да го представим в някаква материална експлицитна форма, ние трябва да го изразим чрез друг знак. В двойките Д — К кодът представлява точно такъв знак.

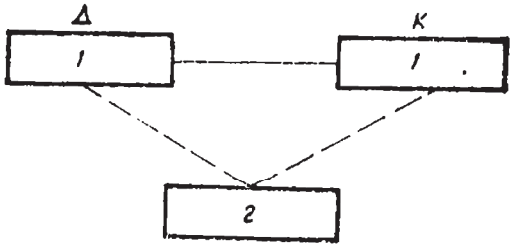
Както знаем вече, кодът принадлежи на някакъв L_1^I и също така има своя означаваща и означавана страна. Означаващата страна на кода е

обикновено някаква четиризначна десетична цифра, а неговата означавана страна или значение — съответно понятие. Следователно семантичната структура на Д и К могат да се представят по аналогичен начин



където 1 е означаваща, а 2 — означавана страна. От това следва, че за да можем да поставим в съответствие на дескриптора Д кода К, т. е. да смятаме, че К или по-точно неговата означаваща страна представлява описание на значението на Д, между тях трябва да има нещо общо, някаква инварианта. Не е мъчно да се покаже, че инвариантата не е нищо друго освен общото понятие, което те представят. Графически това може да се изрази, както на фиг. 2.

Това общо понятие (2), тази инварианта, благодарение на отчитането на която става възможно съпоставянето на даден Д с даден К в речника, ще наречем условно ядро (Я).



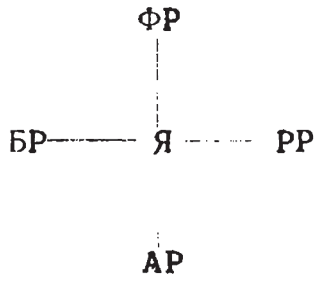
Фиг. 2

III. 1. 2. И така Д и К и по-точно техните означаващи страни са представители в съответния L_i^N и L_i^I на едно и също ядро. Тъй като Я (т. е. съответното значение — понятие) по необходимост трябва да съществува във формата на някаква означаваща страна (тук не сме в състояние да въвеждаме някакъв семантичен метаезик) и тъй като, както ще видим по-долу, между означаващата страна на кода и ядрото съществува еднозначна връзка, за простота на изложението условно ще ги отъждествим и ще говорим засега само за ядра. От друга страна, тъй като Д представлява реализацията на дадено ядро в даден естествен език, ще го наречем р е п р е з е н т а н т (Р). С оглед на това двойките Д — К могат да се презапишат като двойки от вида Р — Я.

III. 1. 3. Да анализираме по-подробно компонентите на двойките от този вид — репрезентанта (III. 1. 3. 0) и ядрото (III. 1. 3. 1).

III. 1. 3. 0. Очевидно е, че компонентът Р, т. е. репрезентатът на дадено ядро, представено чрез код от L_i^I, принадлежи на равнището на

израза (вж. напр. [17, 18]) на даден естествен език, има така да се каже, „национална“ форма и следователно едно и също ядро ще има отделни (понякога еднакви, а понякога различни) репрезентанти в различните естествени езици. Графически това може да се представи така:



т. е. на едно и също ядро (код) от даден информационен език отговарят български, френски, руски и английски репрезентанти.

III. 1. 3. 0. 0. Освен националния характер на репрезентантите да отбележим и друго тяхно съществено свойство — като изрази от даден естествен език те не са и не могат да бъдат изградени върху принципа „едно към едно“ (вж. [15]), от което следва, че те могат да бъдат омонимични, синонимични, многозначни и пр.

III. 1. 3. 1. Докато репрезентантите принадлежат на равнището на израза на даден естествен език и имат своя национална форма и редица „езикови“ особености, ядрата принадлежат към равнището на понятийното съдържание, абстрахирани са от национални (и езикови) особености и имат универсален характер, което се дължи на следното: както видяхме, ядрата са значения на дескрипторите, които в същност представляват основните термини от дадена област на науката или техниката. А термините са репрезентанти в естествените езици на дадени научни понятия. Не е мъчно да се покаже, че като оставим настрана степента на нейното развитие в една или друга страна, всяка област на науката и техниката има интернационален характер и борави с приблизително един и същ понятиен апарат, т. е. с едно и също множество от ядра. Тъкмо на това се дължат универсалният характер и независимостта на специалните науки от особеностите на естествените езици.[12]

Да отбележим, че тази констатация за универсалността на ядрата[13] е вярна не само за тях, но и за логическите и парадигматическите (йе-

[12] Означаващите страни на ядрото — кодовете, имат също универсален характер, тъй като се задават чрез цифри.

[13] Констатацията за универсалността на ядрата, на техните връзки и пр. е напълно вярна само за терминологическите области на естествените езици. Както е известно, терминологическите области са изкуствено и рационално създадени фрагменти в естествените езици, при които по принцип е осигурена взаимно еднозначна връзка между означаващите и означаваните им страни (принцип „едно към едно“). Наистина и нетерминологичните елементи на естествените езици имат свои значения, свое понятийно съдържание и следователно и за тях биха могли да се изградят ядра. Но при тези елементи поради обстоятелството, че те не са изградени върху принципа „едно към едно“, се явяват и такива свойства като т. нар. „различна сегментация на света“ вж. напр [20, 21]), непълната конвенционалност, влияние на прагматиката и пр., които изключват пълното съвпадане на ядрата и следователно тяхната универсалност.

рархически, асоциативни, т. е. „тезаурусови") връзки между тях. От това следва, че нашата констатация за универсалността на ядрата, техните логически и йерархически връзки обуславя възможната универсалност и на трите познати основни типа информационни езици от първа (чисто дескрипторни), от втора (отчитащи някои логически връзки) и от трета (отчитащи и йерархични връзки) степен (вж. напр. [19]), тъй като във всички случаи те представляват множества от кодове, еднозначно описващи (вж. по-долу III. 1. 3. 1. 0) универсалните ядра и евентуално техни универсални връзки.

III. 1. 3. 1. 0. Както видяхме, и означаващите страни на кодовете (четиризначни десетични цифри), и техните означавани страни (ядрата) са универсални. За разлика от елементите на естествените езици те са изградени върху принципа „едно към едно", благодарение на което между въпросните две страни има винаги взаимно-еднозначно съответствие. С оглед на това елементите на информационните езици са винаги еднозначни, неомонимични, несинонимични и пр.

И така репрезентантите могат да бъдат нееднозначни (в широк смисъл), а кодовете не. Тъкмо поради това на един многозначен (или омонимичен) репрезентант се приписват толкова кода, колкото значения той има; на два синонимични репрезентанта се приписва един код и пр.

III. 1. 4. Въз основа на изложеното може да се направи следният основен за нашето изложение извод: универсалният понятиен апарат (значенията на термините) на дадена специална област от науката или техниката може да бъде описан еднозначно чрез кодовете на даден информационен език L_1^I, всеки един от които представлява едно ядро. Следователно даден L_1^I представя ядрата например в областта на радиоелектрониката независимо от това, дали става дума за радиоелектроника, разработвана в България, СССР, Англия и т. н. За разлика от това репрезентантите дават реализациите на кодовете (ядрата) в съответните естествени езици. Тъй като нас ни интересуват българските и руските резюмета, които трябва автоматично да се индексират, и тъй като всяка АСТИ се основава върху множество дескриптори в съответните области, ние можем да кажем, че корелативните дескриптори в двата езика (репрезентантите) и съответните им универсални значения (ядра) могат да се описват чрез тройки от следния вид:

БР — К — РР,

където БР е български репрезентант, К — код и РР — руски репрезентант. Практически това изглежда така

вериг — 0 167 — *цеп*.[14]

Очевидно е, че биха могли да бъдат привлечени и репрезентанти не от два естествени езика, а, да кажем, от три или четири. Тогава бихме

[14] Във всеки дескрипторен речник — едноезичен или двуезичен — се задават не цели дескриптори (изходни словоформи — *верига, цепь*), а сегментирани по един или друг начин техни основи (*вериг-, цеп-*).

имали четворки от следния вид: *вериг — цеп — chain* — 0167. Това поставя въпроса за „триезичен“ вход, с който тук няма да се занимаваме.

Дескрипторен речник, в който на всеки код от информационния език се поставят в еднозначно съответствие репрезентанти само от един естествен език, ще бъде очевидно е д н о з н а ч н о - д е с к р и п т о р е н р е ч н и к, а ако се поставят в съответствие репрезентанти от два естествени езика — д в у е з и ч н о - д е с к р и п т о р е н р е ч н и к. С оглед на това АСТИ, функциониращо въз основа на еднозначно-дескрипторен речник, ще бъде с и с т е м а с е д н о е з и ч е н в х о д, а АСТИ функциониращо въз основа на двуезично-дескрипторен речник — система с д в у е з и ч е н в х о д.

И така, ако имаме дадена определена област на науката или техниката, нейният понятиен апарат може да бъде „покрит“ от кодовете на съответния информационен език, а термините, представящи общи ядра (в български и руски език) — чрез множествата от БР и РР.

Предложеният от автора двуезичен българо-руски дескрипторен речник е изграден тъкмо върху изложените по-горе логико-лингвистически предпоставки. Преди да преминем към въпроса за стандартната форма на този речник (III. 2), трябва да разгледаме още две твърде съществени обстоятелства (III. 1. 5) и (III. 1. 6).

III. 1. 5. Ако анализираме с оглед на морфографемичния им строеж, репрезентантите на тройките от разгледания по-горе вид (БР — К — РР), ще видим, че едни от тях са еднакво оформени, а други — различно оформени.

Е д н а к в о о ф о р м е н а тройка е тази, чиито репрезентанти в двата естествени езика имат еднакъв морфографемичен строеж, например

БР	К	РР
ток	2341	ток

За разлика от това р а з л и ч н о о ф о р м е н а тройка е тази, чиито репрезентанти в двата естествени езика имат различен морфографемичен строеж, например

БР	К	РР
мреж	1179	сет

Разграничаването на еднакво оформените и различно оформените тройки дава възможност цялото множество тройки, които покриват дадени терминологични области в два естествени езика и съответната им научна понятийна област на дадена дисциплина, да се раздели на две съответни подмножества. Това от своя страна открива следните две възможности:

1. Еднакво оформените тройки могат да бъдат редуцирани в д в о й к и. И наистина, щом като в тройка от този тип БР и РР имат еднакъв морфографемичен строеж (ток — 2341 — ток), то тези два репрезентаната могат да бъдат сведени към един общ репрезентант (ОР) и съответната тройка да се запише в следния вид: ОР — К (ток — 2341).

2. В резултат на такава редукция първото подмножество ще се превърне в подмножество от двойки с ОР, а второто ще си остане подмножество от различно оформени тройки. Въз основа на тези подмножества в речника ще може да се обособят две части: обща (задаваща подмножеството двойки с ОР) и диференциална (задаваща подмножеството на различно оформените тройки).

Тъй като разработването на двойките и главно машинното търсене при тях са значително по-прости (а е необходим и по-малък обем на паметта), отколкото при тройките, за предпочитане е общата част на речника да бъде по-голяма от диференциалната или поне с достатъчно голям обем. Обемното съотношение на тези две части на двуезично-дескрипторния речник зависи първо от степента на близостта на съответните два естествени езика, както и от степента на реализирането в тях на научното сътрудничество между съответните две нации (терминологични заемки). С оглед на това очевидно е, че руският и българският език дават прекрасни предпоставки за оформяне на доста голяма по обем обща част на двуезичния речник (например при 2659 дескриптора в българския и също толкова в руския речник, създадени за областта на радиоелектрониката, 1223 се оформят като двойки с ОР; тези констатации са направени от н. с. к. ф. н. К. Илиева). По принцип също такива възможности дават и другите двойки славянски езици, както и английски и френски, испански и италиански и др.

При безспорните предимства на двойките с ОР понятен е стремежът да се увеличи общата част на двуезично-дескрипторния речник. Върху тази разрешима до известна степен проблема не съм в състояние да се спирам тук и ще отбележа само следното: Различната оформеност на редица БР и РР се дължи на „правописни", „буквени" причини. Например *еталон — эталон; барометър — барометр; разпределение — распределение* и пр. С оглед на това е напълно възможно при съставянето на речника да бъдат въведени общи правописни правила, уеднаквяващи различията от посочения тип и следователно увеличаващи общата част, а българските и руските резюмета, които ще бъдат индексирани с помощта на този речник, ще може предварително автоматично да се коригират от помощен алгоритъм, отчитащ посочените общи правописни правила.

III. 1. 6. Сега да отбележим едно твърде съществено ограничение, което в никакъв случай не бива да се изпуска. Всеки речник на една АСТИ, бил той еднозначно-дескрипторен или двуезично-дескрипторен, трябва не само да задава дескрипторите, кодовете и техните връзки, но и в съответни графи да дава цялата информация от лингво-семантичен характер, въз основа на която ще работи индексиращият алгоритъм (структура на дескрипторите, омонимия, многозначност, вариантност, парадигматични връзки, синтагматични връзки и пр.). От това следва, че фактически една речникова единица е представена не само от двойката ОР — К или тройката БР — К — РР, но и от съдържанието на всички други информационни графи. Така еднакво оформената двойка *ток* — 2341, която приемаме за такава само с оглед на морфографемичното съвпадане на думите *ток* в българския и руския език, фактически ще бъде такава само ако освен

това поотделно речниковата информация за всяка една от тях съвпада по всички графи.

Графически това условие може да се представи така. Да предположим, че вместо един двуезичен българо-руски дескрипторен речник имаме два едноезично-дескрипторни — български и руски — речници. Във всеки един от тях дескрипторът *ток* ще бъде представен така:

Български речник[15]

1	2	3	4	5	6	7	8	9
x	1	ток	0	0	0	0	0	2341

Руски речник

1	2	3	4	5	6	7	8	9
x	1	ток	0	0	0	0	0	2341

Следователно двойката *ток* 2341 може наистина да се смята за еднакво оформена, тъй като речниковата информация за двата езика съвпада напълно по всички графи. В противен случай тя би трябвало да се представи като различно оформена тройка. Това е едно много силно ограничение, което поне умозрително би трябвало значително да ограничи обема на общата част на двуезично-дескрипторния речник. С други думи, поставя се въпросът, доколко често може фактически да имаме такова пълно съвпадане на речниковата информация, защото ако този процент е много нисък, безсмислено е да се обособява една обща част в речника. Аз не съм в състояние да навлизам и в тази проблема и ще отбележа следното: предварителните изследвания по нашия експериментален модел, разработван в ЦНИРД на ДСО Респром, показват, че този процент е достатъчно висок и може „изкуствено" да се повишава. Накратко тук нещата стоят така: казано в обща форма, речниковата информация отразява два типа данни — едни, които зависят от особеностите на съответните национални естествени езици (структура на дескриптора, омонимия, многозначност), и други, които се обуславят от фактори, лежащи на общото равнище на съдържанието (елементи, снемащи многозначността, логически и парадигматични връзки и др.). Очевидно е, че различия в речниковата информация може да има по начало само при първия тип данни, защото вторите имат универсален характер. Но и тук, както показват нашите предварителни наблюдения, от една страна, те не са толкова големи (например в нашите български речници за експерименталния вариант на АСТИ за радиоелектрониката, всеки от които, както вече казах, има по 2659 речникови единици, многозначни са само 13 български

[15] За съдържанието на графите вж. III. 0.

и 14 руски дескриптора, при което 7 са общи), а от друга — могат изкуствено да се намаляват.

След всичко изложено може да представим една стандартна форма на двуезичен българо-руски дескрипторен речник (III. 2).

III. 2. Въвеждането на една такава стандартна форма е рационално от няколко гледища: за нагледно представяне на резултатите от всичко изложено досега; като форма, която се попълва въз основа на дадена инструкция, ако разработката се приеме като типова и ще трябва да се съставят подобни речници за различни отрасли; като основа, на фона на която да се представят методите за създаване на такива речници и да се обсъждат възможностите за автоматизиране на някои етапи на това създаване и др. На част от тези въпроси са посветени III. 2. 0. — III. 2. 2.

Обща част на речника

1	2	3 ОР	4	5	6	7	8	9 К
		ток						2153

Диференциална част

1	2	3 БР	4	5	6	7	8	9 К	2	3 РР	1
		вериг						0153		цел	

ОР — общ репрезентант; БР — български репрезентант; РР — руски репрезентант; К — код от L_1^1; графите 1 — 9, както и в еднозначно-дескрипторния речник имат следното съдържание: адрес, брой на пълнозначни думи, основа на дескриптора, структура, многозначност, омонимия, синтагматични и парадигматични връзки; информационен език като съвкупност от кодове.

III. 2. 0. Представената по-долу стандартна форма на един двуезичен дескрипторен речник може да се използува за която и да е двойка естествени знаци; при това очевидно е, че броят и съдържанието на информационните графи може да се изменя.

Разбира се, тази стандартна форма е предназначена за представяне на двуезично-дескрипторния речник, така да се каже, „извън" машината. В машината разположението на информацията придобива съвсем друг вид (вж. по-долу III. 2. 2. 0). След като представихме предложената стандартна форма, можем да се спрем накратко върху методите на съставянето на речник от този двуезично-дескрипторен тип.

III. 2. 1. Трябва да се отбележи, че речник от този тип подобно на едноезично-дескрипторните речници може да се съставя по двата начина — посочени в забел. 10 — предварително и в хода на функционирането на съответната система, така да се каже, като нейно самообучение. Тук ще обсъдим възможните реализации само на първия от двата посочени подхода. Умозрително са възможни следните три метода на работа:

III. 2. 1. 0. Въз основа на нашите специални познания за дадена област, за която се разработва съответната АСТИ (с използуване на терминологични речници и пр.) се съставя списък на ядрата и съответно на кодовете и евентуално се определят някакви логически и парадигматични връзки между тях (ако имаме намерение да разработим за системата информационен език от втора или трета степен). По този начин първо се създава множеството от кодовете, т. е. информационният език на системата. След това се установяват руските и българските репрезентанти на всяко ядро (код), разработва се речниковата информация за тях, въз основа на съпоставянето на морфографемичната им форма (а и на речниковата информация) за всеки един от тях се обособяват еднакво оформените двойки и различно оформените тройки и се попълва стандартната форма на двуезично-дескрипторния речник.

III. 2. 1. 1. Въз основа на определен, достатъчно репрезентативен, corps de texte от български резюмета се обособяват българските дескриптори (прости и сложни) в дадената област. Чрез приети правила за „изкуствена" сегментация се обособяват основите. Всяка основа ще представлява една речникова единица. Основата се обработва в лингвистическо отношение: установява се дали тя е еднозначна или многозначна, омонимична, синонимична (вариантна) и пр. По този начин се подготвя цялата „българска" информация. След това на всяко значение се приписва код. По този начин се създава информационният език на системата. След това за всеки български дескриптор и съпоставеното му ядро се търси руското съответствие, което ще представлява руският репрезентант на ядрото, обработва се по същия начин, съпоставя се с речниковата информация на българския репрезентант по всички графи и въз основа на това се оформят двойките и тройките и се попълва стандартната форма на речника.

III. 2. 1. 2. Въз основа на два достатъчно репрезентативни corps de texte от български и руски резюмета от дадена специална област се разработват независимо един от друг два едноезично-дескрипторни речни-

ка — български и руски. След това се сравняват речниковите редове с еднакви по значение дескриптори, установява се, от една страна, морфографемичното тъждество или различие на репрезентантите, а, от друга, тъждеството или различието на съдържанието на информационните графи и ако е възможно, те се уеднаквяват. Въз основа на това се оформят двойките и тройките и се попълва стандартната форма на двуезичния българо-руски дескрипторен речник.

Въпреки че на пръв поглед третият начин може да изглежда най-сложен или поне изискващ най-голям обем предварителна работа, ние смятаме, че той е за препоръчване, тъй като осигурява най-пълна предварителна подготовка на речника, което от своя страна ще намали „шумовете" при функционирането на системата. Нещо повече, със същия алгоритъм за въвеждане на речника (вж. по-долу III. 2. 2. 0) може поотделно да се въведат двата едноезични речника и въз основа на един експериментален масив например от 500 български и 500 руски резюмета да се проведат подобни индексирания както на едните, така и на другите резюмета с общия допълнен индексиращ алгоритъм (вж. по-долу IV). Едва след това отделно експериментиране може да се пристъпи към сливането на двата едноезични речника в един двуезично-дескрипторен речник.

III. 2. 2. Сега да видим някои възможности за автоматизиране в тази област. Тези възможности се откриват в две насоки: по отношение на въвеждането на речника в машината (III. 2. 2. 0) и по отношение на самото му състояние (III. 2. 2. 1).

III. 2. 2. 0. Н. с. И. Мицев и н. с. С. Станков са разработили, програмирали и експериментирали алгоритъм за въвеждане на речника, разработен за областта на радиоелектрониката, в машината (в случая става дума първо за едноезичния българо-дескрипторен речник). Този алгоритъм осигурява следното: алфабетизиране на дескрипторите по зони (практически в ръчно съставения речник дескрипторите могат да бъдат дадени заедно с техните информационни редове в произволен ред; алгоритъмът осигурява тяхното подреждане по азбучен ред и записването на всички дескриптори, които започват от една и съща буква в отделни зони на паметта); подреждане във всяка зона на дескрипторите по намаляваща дължина (например *автоматизиран- производство-; автоматизиран-; автомат-;* това е необходимо, за да се избегне неправилно влагане); сегментира дескрипторите на „начала" и „опашки"[16], записва ги в

[16] Накратко това се свежда към следното: под „начало" в случая се разбират първите четири символа на дескриптора независимо от това, дали той е прост или сложен (например квад ще бъде „началото" на всички дескриптори, които имат за първа дума думата *квадрат* или някоя от нейните производни, както и на всички прости дескриптори, представляващи такива производни); онова, което остава след отделянето на „началото", се нарича „опашка". В зоните на речника са подредени по азбучен ред първо „началата" и автоматичното търсене, осъществявано от схемата „търсене в речника" от индексиращия алгоритъм, се води тъкмо по тези „начала": ако „началото" не бъде намерено, това значи, че дадената дума от текста не представлява дескриптор, включен в речника, или не е първа дума на такъв сложен дескриптор, и търсенето за нея се прекратява; ако „началото" се вложи, по съответен адрес се минава към онази зона на паметта, в която са записани съответните „опашки", които могат да следват след даденото „начало", и търсенето продължава по тях. Този начин на представяне на дескрипторите (и на съответната информация) в речника опростява и ускорява в значителна степен търсенето при автоматичното индексиране.

различни зони и ги свързва с определена система от адреси; отстранява евентуално допуснати повторения.

Този алгоритъм за въвеждане на едноезично-дескрипторен речник може с незначителни изменения да се използува и за въвеждане на предложения двуезично-дескрипторен речник. Ще отбележим, че беше проведен експеримент с машина Минск-22, при който с този алгоритъм беше въведен както фрагмент от българския дескрипторен речник, така и фрагмент от руския дескрипторен речник.

III. 2. 2. 1. Както показват нашите предварителни изследвания, при самото създаване на двуезично-дескрипторния речник би могла да се автоматизира фазата на обединяването. Ако, както беше препоръчано погоре (вж. III. 2. 1. 2.), при съставянето на двуезично-дескрипторния речник се използува третият възможен подход, т. е. предварително се разработват два отделни едноезични речника, съдържанието на информационните графи на които се уеднаквява в максимално възможна степен, автоматичното оформяне на двойките и тройките, както и самото „компилиране" на друезично-дескрипторния речник според предложената стандартна форма, не би представлявало никакви особени трудности.

*

Както вече беше посочено, създаването на една АСТИ с двуезичен вход предполага не само разработване на оригинален двуезично-дескрипторен речник, но и съответно алгоритмично осигуряване; някои проблеми, свързани с това осигуряване, ще бъдат разгледани по-долу.

IV. Едното от няколкото твърде съществени предимства, които осигурява разработването на АСТИ с двуезичен вход, основани върху предложения тип двуезично-дескрипторни речници, се заключава в това, че тяхното функциониране може да се осигури от същите алгоритми, които са необходими за АСТИ с едноезичен вход с внасяне в тях само на твърде незначителни изменения и допълнения и смяна на списъците на константите. За да изясним тази възможност, ще видим най-напред какви алгоритми са необходими за една едноезична чисто българска АСТИ (IV. 0), а след това ще набележим необходимите изменения и допълнения (IV. 1).

IV. 0. Очевидно е, че тук не мога да описвам подробно алгоритмите, които осигуряват функционирането на една едноезична (българска) АСТИ, и затова само ще ги изброя, като условно с оглед на целите на това изложение ги подразделям на три групи: помощни алгоритми, лингвистически алгоритми и логически алгоритми.

IV. 0. 0. Помощните алгоритми или схеми[17] осигуряват на първо място следното: въвеждане на речника; въвеждане на масива от ИК (както и на въпросите) и кодиране на названията на периодичните издания, страните и пр. (според зададени таблици), които фигурират в резюмето или във въпроса на потребителя; извеждане на отговорите и др.

[17] Тук и по-нататък по аналогия с терминологията, приета при машинния превод, под схема ще разбираме блок или група правила, решаващи определена проблема, а рамките на даден алгоритъм.

IV. 0. 1. Лингвистическите задачи се решават предимно от схемите на индексиращия алгоритъм. Неговото общо предназначение е да превежда автоматично както ИК (информационните карти от масива), така и въпросите на потребителя от естествения език, във формата на който те са зададени (входния език, в случай българския) на информационния език на системата и да записва в съответни зони от паметта резултатите от този превод, т. е. „ликовете за търсене" и „предписанията за търсене". Тъкмо този превод представлява автоматичното индексиране. Обаче, както това е осигурено и в двата експериментални модела, автоматичното индексиране е разделено на два типа: „кодиране" и „собствен" превод. Накратко тук работата се свежда към следното.

Както знаем (вж. I), и ИК, и стандартната бланка-въпрос се състоят от две части: фактологическа и свободна или текстова. Фактологическите части са разделени на еднакъв брой еднакви рубрики, задаващи еднакви информационни белези (т. е. фактологически данни като име и презиме на автора, заглавие, название на периодичното издание, година на издание, език и пр.). В текстовата част на ИК е записано резюмето на първоначалния документ от МД, а в текстовата част на въпроса потребителят има свободата да запише по избран от него начин областта, за която се отнася неговото записване, както и всякакви допълнителни уточнения.

Информацията в рубриката на фактологическите части на ИК и на бланката-въпрос се записва по регламентиран в Инструкцията за използването на системата начин (важно е да се отбележи, че в случая е регламентиран начинът на записването, а не самият български език; това се дължи на приетото от нас решение на проблемата за свободата на езика на съставителите на ИК и на потребителя — вж. [5]). Благодарение на създадения по описания по-горе начин изоморфизъм между ИК и бланката-въпрос, както и благодарение на еднаквия регламентиран начин на записване на съдържанието на еднаквите рубрики от техните фактологически части, което по начало изключва омонимията и многозначността, оказва се възможно тези информации да не се превеждат, а просто да се кодират[18] по един и същ начин. Това кодиране се осъществява при перфорирането на съответните данни. Тези кодирани по еднакъв начин данни се записват на съответните места на „лика за търсене" и на „предписанието за търсене" и тъй като те вече са представени във формата на един и същ еднозначен език (машинния), могат направо да се съпоставят (чрез пълно влагане) при работата на логическата схема 2 ($Л_2$ на фиг. 1) по съпоставянето на даденото предписание за търсене с

[18] Тук се налага едно терминологично уточнение: в същност тук под превод ние разбираме онова, което съветските автори (вж. [27]) разбират под нетавтологично кодиране, а под кодиране — онова, което те разбират под тавтологично кодиране. Тавтологично кодиране имаме, опростено казано, тогава, когато се изменя само графемичната (или фонемичната) форма на съобщението (например при перфорирането, при представянето на едно езиково съобщение във формата на кода Морз и т. н.). Нетавтологично кодиране имаме тогава, когато се изменят и репрезентантите на други равнища на езика на съобщението — морфология, синтаксис, равнището на означаваните и пр. (например превод от един естествен език на друг, превод от естествен език на информационен език и пр.).

даденото множество от ликове за търсене, без да е необходимо никакво превеждане.

За разлика от това, тъй като текстови части на ИК и на бланката-въпрос (резюмето и посочването на областта и допълнителните данни) представляват не отделни фактологически информации, а свързани текстове, представени във формата на свободен български език, тук вече се налага превод от този свободен естествен език на информационния език на системата, т. е. автоматично индексиране в истинския смисъл на думата.

Очевидно е, че при „простото" кодиране пред индексиращия алгоритъм не се поставят никакви лингвистически проблеми (той изобщо не се занимава с него), но тези проблеми се поставят при превода в истинския смисъл на думата. Общо казано, тук се поставят следните въпроси: първо, всяка текстова единица (всяка последователна словоформа от резюмето или свободната част на въпроса) трябва да бъде идентифицирана в речника, а след това да бъдат снети всички възможни свързани с нея неедноозначности, за да се стигне до еднозначен избор на кода от информационния език. Тези задачи се осъществяват от следните схеми на индексиращия алгоритъм: „Търсене в речника", „Проверка на сегментацията", „Проверка на структурата на дескриптора и вариантност" „Омонимия" „Многозначност", „Анализ извън ЛТ и ПТ", „Създаване и свиване на матрицата" и евентуално от схемите „Синтагматични връзки" и схемата „Парадигматични връзки".

IV.0.2. Логическите схеми (на фиг. 1 са дадени четири такива схеми, но само по себе си се разбира, че техният брой и предназначение може да варират) решават, общо взето, следните задачи: проверяват предварително логическата допустимост на въпроса и ако той не е коректно зададен, го връщат на потребителя с посочване на допуснатите грешки ($Л_1$); като се ръководят от определена стратегия и тактика на търсене,[19] въз основа на дадения въпрос (или по-точно на отговарящото му предписание за търсене) подбират чрез установяване на т. нар. критерий на смислово съответствие подмножеството от релевантните ИК (т. е. тези, които по зададените във въпроса информационни белези се отнасят към него — това при нашата организация се върши от $Л_2$); извличат от това релевантно подмножество онази и само онази информация, която е необходима за отговаряне на дадения въпрос ($Л_3$), и формулират съответните отговори.

Въз основа на тези алгоритми и схеми функционира една „българска" едноезична АСТИ.

IV.1. Да си зададем следните два въпроса: ако изброените „български" алгоритми се свържат не с един едноезичен български-дескрипторен речник, а с един двуезичен българо-руски дескрипторен речник, т. е. ако се пренесат от една едноезична „българска" в една двуезична „българо-руска" АСТИ, къде по принцип ще се наложат изменения и допълнения на тези алгоритми (IV.1.0) и какви ще бъдат те (IV.1.1).

---

[19] Тук в твърде голяма степен допринася за рационализирането, оптимизирането и съкращаването на търсенето оригиналното предложение на доц. Д. Добрев последователностите от дескрипторите, които се получават при автоматичното индексиране, да се записват във вид на матрица, която да се свива по определен начин.

IV. 1. 0. Не е мъчно да се покаже, че условията на функционирането на всички помощни алгоритми, схеми и подсистеми (само с добавяне на „руски" букви), както и на всички логически алгоритми няма да се изменят, от което следва, че при пренасянето на тези „български" алгоритми в една АСТИ с двуезичен българо-руски вход в тях няма да се наложат никакви особени изменения и допълнения.

Същото трябва да се каже още и за две области: за въвеждането, обработката и съпоставянето на фактологическите данни и за схемата „Търсене в речника", както и за схемите „Анализ извън ЛТ и ПТ" и „Парадигматични" връзки.

Обстоятелството, че при обработването на фактологическите данни няма да настъпят никакви изменения в условията на функционирането на съответните схеми, е очевидно: та нали тук цялата работа се свежда към просто кодиране и съпоставяне до пълно влагане. По същия начин за схемата „Търсене в речника" е напълно безразлично дали тя проверява съдържанието на „български" или „руски" графи, след като информацията в тях е зададена по същия начин. При другите две схеми — „Анализ извън" и „Парадигматични връзки" — няма да настъпят изменения в условията на функционирането им поради следната причина: тези схеми работят с данни, които принадлежат на универсалното равнище на значенията, на ядрата, и техните връзки от понятийния апарат на дадена научна област, които, както вече беше показано (вж. III. 1. 1. 1), не са свързани със спецификата на дадени естествени езици и имат универсален характер.

И така по начало изменения в условията на функционирането ще се явяват само при работата на „лингвистическите" схеми, които работят въз основа на онази част от речниковата информация, която отразява и свойства, свързани със спецификата на двата естествени езика на равнището на израза, а това са схемите: „Омонимия", „Многозначност" и проблемата за вариантността.

Както виждаме, при предложената организация на двуезично-дескрипторния речник използуването на „българските" алгоритми в една АСТИ с двуезичен руско-български вход ще наложи изменения и допълнения само в една твърде незначителна част от техните схеми. Сега да видим какви именно ще бъдат тези изменения.

IV. 1. 1. Най-напред ще видим какви езикови изменения се явяват в съответните „руски" области и във основа на това ще посочим накратко какви допълнения ще се наложат в съответните схеми.

IV. 1. 1. 0. Проблемата за омонимията се поставя така. По аналогия с теорията и практиката на машинния превод при автоматичното индексиране трябва да се различават два типа омонимия: текстова и речникова. Към текстовата омонимия спадат онези случаи, при които даден дескриптор представлява дума, която по начало е омонимична в даден естествен език, без тази омонимия да се дължи на някакви допълнителни „изкуствени" причини. Така българският дескриптор *ток* е омонимичен в този смисъл, защото в българския език тази дума е омонимична (ток на обувка и електрически ток). За разлика от това речниковата омонимия се дължи не на общоезикови причини, а на онзи начин на сегмен-

тиране, който е приет при създаването на словника на речника. Тук накратко нещата се свеждат към следното. Както беше посочено, в дескрипторния речник (едноезичен или двуезичен) като речникови единици не се включват словоформи (т. е. всички граматически форми на една и съща основа: *квадрат, квадрата, квадратът, квадрати, квадратите*), а се включват сегментирани по даден начин основи.[20] Така например за отглаголното съществително *автоматизиране* (-ето) и адективираното причастие *автоматизиран* (-, -а, -о, -и, -ият, -ата, -ото, -ите) може да се включи в речника една основа — *автоматизиран-*. Може да се предложи и друга сегментация, например да се дават не основи, а условно казано, „лексеми", състоящи се от приставките и корена. Тогава в речника би се включила лексемата *автомат-*, но вече като представител не само в двата посочени по-горе класа и техните граматически форми, но и на всички глаголни производни и техните форми. Това значи, че лексемата *автомат-* ще представлява в речника не само формите като *автомата, автоматът, автоматите*, но и производните класове като *автоматизиран, автоматичен, автоматически, автоматизиране, автоматизирам* и др.

Както казах, самото явление речникова омонимия се дължи на приетата в речника сегментация и се свежда към това, че една и съща речникова единица (основа или лексема) ще се влага по същия начин в различни части на речта и техните форми и машината, след като е идентифицирала, т. е. вложила речниковата единица в текстовата, не е в състояние „да разбере" непосредствено с коя част на речта има работа в случая. Така ако в речника е включена основата *автоматизиран-*, тя ще се влага еднакво в отглаголното съществително (*автоматизиране*, -ето) и в прилагателното (*автоматизиран-*, -а, -о, -и,..., -ите) и ще се получи речникова омонимия съществително/прилагателно — S/A. Но ако в речника се включи само лексемата *автомат-*, по-очевидни причини ще се получи вече тройна омонимия — съществително/прилагателно/глагол — S/A/V. Между другото се набелязва такава закономерност (по-подробно вж. [23]): колкото морфематичният състав на речниковата единица е по-малък, толкова повече намалява обемът на речника и толкова повече се увеличава броят на случаите на речникова омонимия и евентуално на многозначността. По начало подобни типове речникова омонимия се снемат (поне за славянските езици) доста просто — чрез проверка на остатъка. Така посочената омонимия съществително/прилагателно, която се получава при включването в речника на основата *автоматизиран-*, ще се разреши чрез следната проста проверка: ако остатъкът е *е-* или *-ето*, имаме съществително, а във всички други случаи прилагателно. Обаче трябва да се отбележи, че в единични случаи такива проверки са недостатъчни и се налага контекстуален анализ в рамките на сегмента.

---

[20] В предложените от нас речници се използува принцип на сегментиране, доста различен от традиционния, който осигурява, от една страна, максимално възможно намаляване на обема на речника, а, от друга, минимална омонимия и максимално преодоляване на случаите на т. нар. редувани основи (*военен-, военна*). Но независимо от принципите на това сегментиране в речника обикновено се записват основи, в повечето случаи съвпадащи, но в редица случаи и отличаващи се от общоприетото им разбиране.

Ето такива са накратко двата типа омонимия — речникова и текстова — и задачата на схемата „Омонимия“ е да установява кой от възможните омоними имаме в дадения случай. Веднага трябва да се отбележи, че практически при съставянето на дескрипторни речници обикновено не се срещат случаи на текстова омонимия. Така докато за целия български език съществителното *ток* е омонимично, дескрипторът *ток*, включен в един речник за АСТИ, която се разработва за областта на радиоелектрониката, радиотехниката и пр., очевидно няма де бъде омонимичен, тъй като в тези области едва ли някога може да става дума за ток на обувка. Онези единични случаи на текстова омонимия, които поне умозрително биха могли да се появят при един или друг дескриптор, ще могат да се обработват като случаи на многозначност (изобщо при автоматичната обработка на естествените езици, като машинен превод, автоматичен анализ на текстове, автоматично индексиране и пр., начините на обработка и на снемането на двете явления са еднакви и следователно с оглед на тези утилитарни цели двете явления обикновенно не се разграничават).

И така при съставянето на едноезични или двуезични дескрипторни речници остава само речниковата омонимия. Сега да видим дали ще има различия между речниковата омонимия в един български и в един руски дескрипторен речник, посветени на една и съща област. Такива различия ще има и те ще се дължат на следното: както видяхме, речниковата омонимия възниква в резултат на приетия начин на сегментиране; но колкото и да е „изкуствен“, този начин се обусля обективно от морфематичните особености на даден естествен език; и тъй като морфематичният строеж на думите в българския и руския език е все пак различен, различни ще бъдат до известна степен и типовете омонимия, които ще се получат в единия и в другия речник. Обаче трябва да се има пред вид, че благодарение на близостта на двата езика и на една допълнителна лингвистическа обработка тези различия могат да се сведат към минимум: така например в нашия български речник, разработен за областта на радиоелектрониката, има 8 типа омонимия, а в руския, разработен за същата област — 6, при което 4 съвпадат. Но все пак остават различия и ако обемите на двата речника се увеличат значително, не е изключено да се появят и други типове различна речникова омонимия.

Тъкмо тези различия обуславят обстоятелството, че схемата „Омонимия“ от българския индексиращ алгоритъм трябва да бъде модифицирана и допълнена, за да може да снема и „руската“ омонимия при една АСТИ с двуезичен българо-руски вход. Но както показват нашите предварителни изследвания, тези допълнения са твърде незначителни.

Практически схемата „Омонимия“ се състои от толкова блока, колкото типа омонимия има. Очевидно е, че при еднаквите типове омонимия при двата репрезентанта — руския и българския — ще могат да работят същите блокове. При онези 2—4 типа, които евентуално ще се различават било по своя морфематичен състав, било по възможните омонимични класове, било най-сетне по снемащите компоненти, ще трябва да се включат допълнителни руски блокове „Омонимия“, а в речниковата графа, задаваща информация за омонимия, да се включи допълни-

телен индекс, показващ, че ако машината работи в „български“ режим (вж. по-долу IV. 1), трябва да се мине към блок еди-кой си, а ако работи в руски режим — към блок еди-кой си. Но дори и това решение може да бъде опростено, ако блоковете се отделят от константите.

VI. 1. 1. 1. Проблемата за многозначността се поставя така. Отделни както български, така и руски дескриптори могат да имат не по едно, а по две или в единични случаи и по три значения, т. е. да бъдат свързани не с един, а с два или респективно с три кода от информационния език на системата. Например *баз* 0092
0093

Задачата на схемата „многозначност“ е да установи кой от възможните два или три кода трябва да се вземе в дадения случай. Функционирането на схемата „многозначност“ от един индексиращ алгоритъм на една АСТИ може да се основава върху на пръв поглед два различни принципа, които, въпреки че това много често не се съзнава, имат за основа едни и същи дълбоки логико-лингвистически свойства. Първият от тях може да наречем условно „тезаурусов“ а вторият — „валентностен“.

Тезаурусовият принцип на снемането на многозначността на дескрипторите се основава върху следното. Тезаурусът представлява фактически една йерархична класификация (която може да бъде представена дървообразно) на ядрата и въз основа на тях на самите дескриптори. Всеки тезаурус е подразделен на области, а областите — на низходящи йерархични вериги (например, условно казано, областта „наука“ ще се раздели на две вериги — „математически“ и „нематематически“; математическите ще се подразделят на „алгебра“, „геометрия“, и т. н.). Очевидно е, че ако един дескриптор има две или три значения, всяко едно от тях ще се намира на различна верига или подверига в тезауруса. И тъкмо фактът, че дадено значение се намира на дадена верига, се използува за снемането на многозначността.

Валентностният принцип за снемане на многозначността на дескрипторите се свежда към следното: в значението, в което се използува тук, терминът „валентност“ означава способността на дадена дума (или комбинация от думи) да се съчетава в текстове със строго определен и ограничен брой думи (или комбинации), ако е използувана в едното от значенията си, и да се съчетава с друга строго определена и ограничена по брой група думи или комбинации, ако е използувана в другото си значение. За простота да вземем един условен пример. Нека имаме дескриптора *писта* и нека приемем, че той има две значения — самолетна писта и писта на магнетофонен диск или лента. Очевидно е, че в първото му значение той ще се свързва в дадени текстове с такива думи като *самолет, авиация, летище* и пр., а във второто си значение с такива думи като *диск, лента, магнетофон* и пр. Думите (или комбинациите), с които може да се свързва даден дескриптор в дадено свое значение, ще наричаме е л е м е н т и, с н е м а щ и м н о г о з н а ч н о с т т а, и тъкмо върху тях се основава валентностният начин. Без да навлизам в проблемата (въпреки че по нея има много неясноти и ненужни спорове), ще посоча само, както вече отбелязах, че и двата подхода имат една и съща дъл-

бока логико-лингвистическа основа — универсалността на връзките на равнището на ядрата.

Тъй като, както показва нашият опит, броят на многозначните дескриптори в дадена специална област на науката или техниката е обикновено твърде ограничен[21] — в нашия български речник за областта на радиоелектрониката 13, а в руския 10, — то създаването на тезаурус само за целите на снемането на многозначността едва ли е оправдано (напротив, това би било оправдано, ако тезаурусът се използува и за други цели: установяване на семантични връзки между дескрипторите в текста, „обогатяване" или „обедняване" на въпросите и пр.). Следователно при сегашния етап на развитието на АСТИ по-рационален е „валентностният" път на изграждане на схемите „Многозначност".

Очевидно е, че българската схема „Многозначност" ще може да работи и при онези случаи на руската многозначност, при които имаме еднакви типове многозначност (т. е. многозначни са не само репрезентантите на едни и същи ядра, но са еднакви и снемащите елементи). В повечето от случаите на българската и руската многозначност това е именно така. За незначителния брой случаи, в които се наблюдават различия, българската схема „Многозначност" трябва (подобно на схемата „Омонимия") да бъде допълнена в краен случай с 4—6 блока.

IV. 1. 1. 2. Доколкото ми е известно, проблемата за „вариантността" на дескрипторите и нейното решаване в една АСТИ бе за първи път поставена в такъв аспект в нашата работа по създаване на информационната система за областта на радиоелектрониката (вж. [11]). Тук нещата се свеждат към следното. Известно е, че под синонимия (абсолютна) се разбира онова явление в естествените езици, при което две различни думи назовават едно и също понятие. Така дескрипторите *стъпало* и *каскад* са синоними. Този тип синонимия на простите дескриптори се решава твърде лесно: в дескрипторния речник двата синонима се свързват е един и същ код от информационния език. Но както показва обработеният от нас езиков материал, такова явление в българския език се наблюдава не само при простите дескриптори, но и при сложните, при което във всеки случай синонимичността на сложните дескриптори се дължи на „варирането" на служебните думи, които влизат в техния състав. Така сложните дескриптори *генератор на импулси* и *генератор за импулси; ниво на бяло* и *ниво бяло; регулиране на движението* и *регулиране движението* са синонимични в резултат на редуването (варирането) на служебните думи — *на* (*за; на*). Тъкмо поради това вариране ще говорим за вариантност и за варианти на сложни дескриптори като вид техни синоними. Ако не се даде някакво решение на тази проблема за вариантите на сложните дескриптори, функционирането на системата ще има следния недостатък: ако в речника е включен единият от вариантите и ако в текста на резюмето или на въпроса е използуван другият (невключеният в речника) вариант, той очевидно няма да бъде иденти-

[21] И това е напълно понятно, тъй като фактически дескрипторите са термини, а при създаването на терминологията една от основните грижи е избягването на многозначността.

фициран като дескриптор, неговият код няма да бъде включен в матрицата на дескрипторите, отговарящи на даденото резюме или въпрос, в резултат на което ще се намали вероятността на включването на съответния документ в релевантното подмножество и следователно ще се увеличи процентът на шумовете, които се получават на изхода на системата.

По начало вариантността може да се реши по същия начин, както и случаите на синонимичността на простите дескриптори: всички варианти на сложните дескриптори да се включат в речника и да се свържат с един и същи код от информационния език на системата. Това решение има обаче три основни недостатъка: първо, поради това, че в българския език посочената вариантност е свързана до голяма степен с нерегламентирана индивидуална употреба, тези случаи са твърде мъчно предвидими и не може да има никаква сигурност, че речникът ги е обхванал напълно; второ, включването на всички тези варианти би довело до едно значително увеличаване на обема на речника (особено като се има пред вид, че това са винаги сложни дескриптори); трето, по причини, които ще бъдат изложени по-долу, включването в речника на българските варианти би довело до съществено намаляване до общата част на двуезичния руско-български речник.

Поради това предложих друго твърде просто решение, което позволява да се преодолеят и трите посочени недостатъка: всички сложни дескриптори се записват в българския речник без служебните думи, които влизат в техния състав (т. е. в речника ще се запише *генератор импулс-; нив- бял; регулиран- движен-*). При това в графа 2 на речника се дава информация за броя на пълнозначните думи, от които се състои даденият сложен дескриптор (това е необходимо за определяне на дясната граница на търсенето), а в графа 4 се дава информация, дали в състава му могат да фигурират служебни думи. По тази информация при интересуващата ни ситуация, т. е. когато в речника е записан сложен дескриптор, без служебни думи, а в текста на резюмето или на въпроса той е даден с вариантни или невариантни служебни думи, схемата „Търсене в речника", след което е идентифицирала първата дума, установява най-напред дали това е прост или сложен дескриптор. При първата алтернатива започва неговата обработка, а при втората машината взима всяка последваща дума (докато стигне дясната граница) и се опитва да я вложи в специален списък на служебните думи (ако в графа 4 е посочено, че в дадения дескриптор има такива). При положителен резултат от тази проверка тя я заличава, а при отрицателен минава към следващата. По този начин текстовият сложен дескриптор, в който има служебни думи, се „освобождава" от тях, с което схемата „Търсене в речника" едновременно постига два резултата — идентифицира дескриптора и благодарение на заличаването на служебните думи премахва причините за вариантността и с това eo ipso решава и тази проблема.

Така е решена вариантността в нашия български едноезичен речник и в нашия български индексиращ алгоритъм. Но как се поставят нещата в руския дескрипторен речник. Поради исторически лингвистически причини, които тук не съм в състояние да обсъждам (отдавна утвърден

предимно синтетичен строй на съвременния руски език и незавършен процес на преминаването на българския език от синтетизъм към аналитизъм, който се съпътствува от колебания и по-малка диференциация на значенията на предлозите, което води до тяхната синонимия и широки възможности за индивидуална употреба), почти във всички случаи, в които имаме сложни български дескриптори с вариращи служебни думи, в руския език имаме твърди, невариращи синтетични падежни конструкции без предлози: *генератор импульсов; регулирование движения; уровень белего.* От това следва, че докато в един едноезичен български дескрипторен речник ние сравнително често ще се срещаме с явлението вариантност, в един руски дескрипторен речник то по начало е изключено. Това ни налага следните изводи: ако двойките, които оформяме за един двуезичен българо-руски дескрипторен речник, ще трябва да отчитат и вариантността, то трудно ще може да се получи една вариантна двойка, защото българският дескриптор ще бъде вариантен, а руският — не. Поради това и поради обстоятелството, че по начало руските сложни дескриптори твърде рядко могат да бъдат вариантни, за препоръчване е не само при двойките, но и при тройките информацията за вариантността да се отнася само за българските дескриптори. От това ще следват две неща: първо, тази информация няма да пречи за създаването на еднакво оформени двойки, а, от друга — блоковете, обработващи вариантността от схемата „Търсене в речник" на индексиращия алгоритъм, ще влизат в действие само когато този алгоритъм работи в български режим (вж. по-долу).

При това положение на нещата е очевидно, че в българските блокове, обработващи вариантността, няма да е необходимо да се правят каквито и да било корекции и допълнения.

*

Ето такива са схемите, които би трябвало да се допълнят, и такива са самите допълнения, които ще трябва да се направят в „българските" алгоритми, за да може въз основа на включения в системата двуезичен българо-руски дескрипторен речник те да са в състояние да индексират и руски резюмета. Както видяхме, тези поправки и допълнения са твърде незначителни.

IV. 2. В заключение искам да опиша с няколко думи как ще функционира една АСТИ с двуезичен руски и български вход. Очевидно е, че тя трябва да бъде снабдена със съответен двуезичен българо-руски дескрипторен речник и „допълнени" алгоритми. При това тя ще бъде свързана вече не само с масив от български ИК, но и с масив от руски ИК, чиито резюмета ще бъдат взети в готов вид чрез косвено реферирание Българските ИК ще имат един код, да кажем 1, а руският — друг, да кажем 0. По този код ще се определя българският или руският режим на работа.

При български режим на работа схемата „Търсене в речника" ще бъде насочена към общата част на речника и българската страна (колонката БР) от диференциалната част на речника (вж. стр. 161). След

това въз основа на правилата от тази схема машината взима първата дума от текста на резюмето (или на свободната част на въпроса) и се стреми да вложи в нея първото „начало" от съответната зона на общата част на речника. Ако „началото" е вложено, следва идентифициране на опашката с установяване дали това е прост или сложен дескриптор, проверка на съдържанието на всички речникови графи, по сигнализация на които евентуално работят другите схеми на индексиращия алгоритъм („Омонимия", „Многозначност", „Връзки" и пр.), установява се еднозначно кодът на дескриптора и се записва в матрицата на дескрипторите, която представлява част от „лика за търсене", т. е. от превода на ИК в информационния език на системата. Ако никое „начало" от съответната зона от общата част на речника не се вложи в първата текстова единица на българското резюме (въпрос), машината минава към българската колонка на диференциалната част на речника и следват същите операции по идентифицирането и анализирането. При положителен резултат дадена дума (или съчетание) от българския текст е идентифицирана като дескриптор и неговият код се записва в матрицата, а при отрицателен, след като е прегледана цялата зона от диференциалната част, която отговаря на първата буква на обработваната текстова единица, се стига до извода, че тази българска дума не е дескриптор и процесът започва отначало с втората дума от българския текст и т. н.

Абсолютно по същия начин, но по индекс 0 (т. е. когато работи с руски резюмета или въпроси, машината ще функционира в руски режим; разликата е в това, че схемата „Търсене в речника" ще се насочи към общата част на речника, а след това не към българската колонка от диференциалната част, а към руската.

В резултат на всичко това благодарение на предложения двуезичен българо-руски дескрипторен речник и незначителни допълнения само на една малка част от българските лингвистически схеми от „българския" индексиращ алгоритъм машината ще може автоматично да индексира не само масив от български информационни карти, но и масиви от руски информационни карти, с което се осъществява практически двуезичният вход на системата и се преодоляват посочените в началото на това изложение недостатъци, присъщи на съществуващите АСТИ с едноезичен вход. Според нашите предварителни данни всичко това позволява да се намали с около 30% стойността на създаването и експлоатирането на една АСТИ в нашите условия.

Предложеният метод дава и други възможности: от една страна, да се помисли и за включване на трети вход — например английски, което според мен в нашите условия е една задача, заслужаваща голямо внимание, а, от друга, открива възможността чрез несложни допълнения към алгоритмите да се позволи на потребителя по негово желание да задава въпросите на български или руски и пак по негово желание да получава отговорите на български или на руски. Но поне засега тези възможности едва ли имат практическо значение.

ЛИТЕРАТУРА

1. Gardin, J. C.: Etat et tendances actuels de la documentation automatique. Traduction automatique, Paris, 1968, Nr. 1, p. 1—12.
2. Сб. Горизонты науки и техники. М., 1969 (превод от англ.).
3. Coyaud, M.: Linguistique et documentation, Larousse, Paris, 1970.
4. Leski, K., Kopinski, J.: Problems of technical progress. Warszawa, 1969.
5. Людсканов, А.: Автоматизированная система поиска информации в области радиоэлектроники. Материалы Первой национальной конференции по поисковым системам, Варшава, 1973 г. (под печат).
6. Михайлов и др.: Основы информатики. М., 1970.
7. Мидоу, Ч.: Анализ информационно-поисковых систем. М., 1970.
8. Heys, D. G.: Applied Computational Linguistics. Proc. of Second Intern. Congr. of Appl. Ling., Cambr. Engl., 1969.
9. Людсканов, А.: За някои лингвистически и математически проблеми на автоматичната обработка на информация, зададена във формата на естествените езици. Сп. на БАН, № 4, 1970; вж. също: Некоторые лингвистические проблемы автоматизации процессов ИТИ, Труды Симпозиума „Комплексная механизация и автоматизация. . .", М., 1966; вж. също: O pewnych linguisticznych i matematycznych problemach przetwarzania informacyi zakodowanej w językach naturalnych, ODIN, Buletyn, Warszawa, 1970, Nr. 16, 1—31.
10. Мицев, И. А.: Об одной ИПС библиотечного типа. III конф. на бълг. мат., Варна, 1972, Соб. резюмета, ч. I, с. 125.
11. Ljudskanov, A. et al.: Automatic indexing in Documentary Information Retrieval Systems. III КБМ, Варна, 1972, сб. Резюмета, ч. II, с. 280.
12. Илиев, Л.: Математика в современном обществе. II. Единный центр для науки и подготовки кадров по математике и механике, III КБМ, Варна, 1972, 6.—15.
13. Ljudskanov, A.: Mensch und Maschine als Übersetzer. VEB Niemeyer Verlag, Halle, 1972, 1—270.
14. Ludskanov, A.: Is the Generally Accepted Strategy of Machine Translation Research optimal?, Mechanical Translation, New York, v. 11, 1968, No. 1—2.
15. Людсканов, А.: Машина и значение, сб. Проблеми на логиката, С., т. V; 1973.
16. Jakobson, R. O.: On Linguistic Aspects of Translation. Sammelband "On Translation", New York, 1959.
17. Saussure, F. de: Cours de Linguistique générale. 5. édition, Paris, 1960
18. Helmslev, L.: Prologomena to a Theory of Language. International J. of Amer. Linguistics, No. 1, 1953.
19. Coyaud, M.: Introduction à l'étude des languages documentaires. TA-Documents, No. 1, 1966.
20. Абаев, В. И.: Отражение работы сознания в лексико-семантической системе языка. Сб. Ленинизм и теоретические проблемы языкознания, М., 1970, с. 232—263.
21. Mounin, G.: Les problèmes théoriques de la traduction. Paris, 1963.
22. Кибрик, А. Е.: Лингвистические вопросы автоматического кодирования. Сб. Теоретические проблемы прикладной лингвистики, М. 1967.
23. Людсканов, А.: Основи на теорията на машинния превод с оглед на руско-българския МП. Год. на Филологическия фак. Софийски университет, т. LVIII, ч. 1.

Постъпила на 15. XII. 1974 г.

BILINGUAL INPUT IN AUTOMATED INFORMATION RETRIEVAL SYSTEMS

A. Ludskanov

(SUMMARY)

This study includes an introduction and four parts.

I. In the introduction is underlined that in the scientific-technical revolution circumstances the size of the relevant information continuously increases and its processing by the traditional hand methods hampers the due information to the scientists and specialists about all the novelties and achievements in their fields. In connection with this the need is underlined of automatization in the information activity and computing technique as well as of a new type optimization of the automated retrieval systems of scientific technical information (AIRS) through the establishment of bilingual (polylingual respectively) inputs. With this problem the present study is dealing.

II. In the first part a simplified AIRS logical diagram is offered as well as characteristic of the principal topics and their interrelations. The concept "input language" is introduced, i. e. such a natural language on which an AIRS is built. Usually it is a national language and on it the input data are recorded, questions are put to AIRS and its answers are formulated. The automated processing systems of data entered in a natural language are defined as unilingual input systems.

III. In the second part the systems with bilingual input are considered in connection with the summarizing process and particularly the use of ready summaries of publications which are published in other languages (for example Russian). But to be available for use they must either be translated in the approriate national language (for example Bulgarian) by specialists or by a machine system or build up a different retrieval system, which will process the data entered in the natural language. Going through all those solutions the author finds them not enough rational and for this reason he offers each national AIRS (particularly those of the socialist countries) to have a bilingual (polylingual) input. Such input function on the ground of a bilingual (polylingual) descriptor's dictionary and lists containing the constants of the input natural languages. Using them the indexing algorithm (program) indexes automatically, e. g. translates the summaries and questions of these input languages into the AIRS information language.

IV. In the third part are stated the needed logical-linguistic prerequisites for the establishment of a bilingual descriptor's dictionary. Its standard form and methodical instructions are offered. All the possibilities for automatization of some of the parts of the dictionary are also given. A bilingual dictionary consists of pairs and triads of descriptor's basis, which are of the following type: — 1. CR-C (where CR — common representative, C — code) in the case when the descriptors are with the same graphics, for example: TOK-2341. 2. BR-C-RR (where BR — Bulgarian representative, RR — Russian representative), in the cases when the descriptors are with different graphics, for example ВЕРИГ— 0167— ЦЕП. Each of these pairs

and triads is given in a separate dictionary article which also contains a numeric information directing the indexing program to the lists of constants and diagrams[1] eliminating the polysemy, homonymy and other linguistic difficulties as well.

V. In the fourth part is pointed out that AIRS with bilingual input functions on the basis of the same algorithms and programs. The author divides these algorithms into three types: "auxiliary", "linguistic" and "logical". 1. The auxiliary algorithms provide the introduction of a dictionary, information massives, countries' code catalogues, periodical publications etc. 2. The linguistic problems are solved by the indexing algorithm diagrams. The indexing represents the following: the word forms of the summary (question) are compared symbol by symbol with the dictionary's descriptors. Where the symbols coincide the descriptor's code is recorded in the image or prescription for search which in turn could be organized as matrices. In the indexing process the polysemy, homonymy, synonymy and others are eliminated. With a special arrangement of the dictionary the search time could be decreased as at the begining a specific number of initial symbols are compared and only after their coincidence the rest of the descriptor's symbols are compared. 3. The logical diagrams perform the logical control over the entered user's questions, compose the relevant data subset and derive from it only such information which is necessary to the user.

VI. In the conclusion it is described the work of an AIRS with bilingual input. It is underlined that when designed in this way AIRS can continuously be enlarged and enriched by new lingual inputs depending of the needs of its users. The carried out experiments are described and it is underlined that AIRS estabiished on the original principle offered leads to savings of about 30%.

[1] Diagrams — a group of rules for solving given linguistic problems in the frame of a given algorithm.