

PL/I NYELVEN ÍRT PROGRAMRENDSZER TEZAUROSZKÉSZÍTÉSHEZ

Balogh Zoltán

INFELOR Rendszertechnikai Vállalat

BEVEZETÉS

A program elkészítése és kipróbálása a rendszerszervezés utolsó lépése. Ezt megelőzően rögzíteni kell a problémát, az alapfeladatot és a probléma megoldását behatároló alapelveket, módszerbeli követelményeket. Az 1. és 2. pont vázlatosan azokat a tezauszokra vonatkozó meghatározásokat és módszerbeli elképzeléseket foglalja össze, amelyekre a program felépül.

1. A tezausz szerkezete

A tezausz információtároló és -kereső rendszerekben alkalmazható strukturált szótár, azaz fogalmak szervezett gyűjteménye. Struktúráját a fogalmak közötti szemantikai kapcsolatok adják. Az információkereső rendszer nyelveként kapcsolatot biztosít a természetes nyelv és a rendszerben tárolt, a tezausz nyelvén, vagyis információkereső nyelven leírt, keresendő információk között.

1.1 A fogalmak

A tezausz fogalomállományát a deszkriptorok fejezik ki. Emellett a tezauszban helyet kapnak a természetes nyelvnek az adott ismeretágban használatos, de deszkriptornak nem minősített kifejezései, a nem deszkriptorok is. Ez utóbbiak a deszkriptorokkal - általában - szinonim vagy legalábbis szoros szemantikai kapcsolatban állnak. Az ekvivalens kifejezések csoportjából a tezausz készítése során kiemelnek egyet, amelyet a továbbiakban deszkriptornak tekintenek. A csoport többi tagja nem-deszkriptor.

A deszkriptor olyan önállóan kezelhető nyelvi egység, kifejezés /szó vagy szóösszetétel/, amelynek jelentése, terjedelme általában külön megjegyzés nélkül is világos, és amelynek értelmi összefüggései a többi deszkriptorral egyértelműen meghatározhatók. Másrészt egy olyan véges számú kifejezés-gyűjtemény eleme, amely gyűjtemény elégséges ahhoz, hogy az adott információtároló és -kereső rendszerben az információk indexelésének és keresésének közös nyelve /szótára/ legyen.

A deszkriptorok egy vagy több szóból állnak; igen gyakran összetett szavak.

1.2 A kapcsolatok rendszere

A tezausz alkotási deszkriptorok közötti tartalmi, értelmi /szemantikai/ kapcsolatok egyik elterjedt csoportosítása az alábbi:

a/ nem-faj /generikus - specifikus/ jellegű hierarchikus kapcsolat; pl. mezőgazdaság - növénytermesztés;

b/ affinitív jellegű, vagyis egyéb értelmi rokonságot kifejező kapcsolat; pl. nevelés - tanítás. Ebbe a típusba sorolják gyakran mindazon /eltérő jellegű/ kapcsolatokat, amelyek nem tartoznak az előbbi csoportba.

Szemantikailag jóval részletesebb kapcsolat-osztályozás is megadható és egyes tezauszokban ezeket is használják az igényektől függően.

Gyakran említik a preferenciális kapcsolatot is. A deszkriptorok és a nem-deszkriptorok közötti kapcsolatok kifejezésére szolgál; példája: alkohol - szesz. Ez valójában nem deszkriptorok közötti kapcsolat, hanem a természetes nyelv adott kifejezései és a deszkriptorra "választott" kifejezés kapcsolata.

1.3 A tezausz fejezetei /részei/

a/ A tezausz egyik strukturált része /gyakran főrésze/ a tezauszba felvett kifejezéseket /deszkriptorokat és nem-deszkriptorokat/ tartalmazza, betűrendben, de minden kifejezés alatt a kapcsolat típusának megjelölésével együtt felsorolja az adott kifejezéssel a megjelölt típusú kapcsolatban lévő, a tezauszba felvett többi kifejezést is. Az ilyen önálló szócsoporthoz neve szócsoporthoz. A szócsoporthoz csak olyan kifejezés szerepelhet, amely a betűrend megfelelő helyén deszkriptorként vagy nem-deszkriptorként is megjelenik.

b/ A KWIC vagy más típusú betűrendes permutált mutató a tezauszba felvett összes deszkriptort és nem-deszkriptort tartalmazza az összes lényeges szóösszetételek szókezdőnek megfelelő betűrendi helyén.

c/ A grafikus rész irányított gráfok formájában ábrázolja a főbb fogalmi csoportokat. A deskriptorok közötti kapcsolatok jobb áttekinthetőségét segíti elő.

d/ Egyéb részek lehetnek még: a fogalmi csoportok szerinti /azon belül betűrendes/ mutatók, tetszőleges kódszám /azonosító jelzet/ szerinti mutatók vagy tárgyköri /szak-/csoportonkénti, esetleg kategóriánkénti, különféle mélységű hierarchiával rendelkező mutatók, mely utóbbiakat pl. a sokszintű hierarchiájú tezaurusrészekhez érdemes készíteni.

Az itt ismertetett program alkalmazásával az a/, b/ és a d/ pontokban megjelölt részeket készítheti el a számítógép.

2. A tezauruszkészítés főbb fázisai

Ezek általában a következők:

a/ A szóanyag gyűjtése a legkülönbözőbb források felhasználásával /külföldi vagy hazai tezauszok, osztályozó-rendszerek, lexikonok, szakszótárak, dokumentumok szövegéből kiemelt kulcsszavak stb./.

b/ Ekvivalens kifejezéscsoportok kialakítása a tezauszba felveendő kifejezésekből és a deskriptorok kiválasztása ezek közül. Ez a művelet a későbbiekben még megismétlődhet.

c/ A szóanyag lexikográfiai szerkesztése. Ez a kifejezések "selejtezését", rendszerezését, nyelvi szabványosítását, kódolását foglalja magában.

d/ A tezausz teljes kapcsolatrendszerének /szóbokrainak, kifejezéscsoportjainak, hierarchikus táblázatainak/ kialakítása.

A tezausz strukturált betűrendes részének /esetünkben fő részének/, valamint indexeinek, mutatóinak előállítását történhet számítógéppel. Mivel a többszöri bővítés, selejtezés és ellenőrzés miatt állandóan ismétlődnek a b/, c/ és d/ alatt felsorolt tevékenységek, a program /és a számítógép/ alkalmazására is ismételten szükség van a végső változat előállításáig.

2.1 A tezauruszkészítés munkamenete

A hazai szakemberek véleménye megegyezik abban, hogy a tezausz kialakítását két fő irányban haladva, szinte egymáshoz kapcsolódva célszerű végezni:

a/ a szógyűjteményből /kifejezések gyűjteményéből/ kiindulva kiépíteni közöttük a strukturális /értelmi/ kapcsolatokat, majd kiala-

kitani a tezausz sztrukturált /betürendes/ részét;

b/ a sztrukturált részhez elkészíteni az abban már szereplő kifejezések permutált mutatóját, és ezen betürendes rész felhasználásával folytatni a további munkálatokat /ld. a fentiekben a/ - d/ pontokat/.

Egyes esetekben statisztikailag is elemzik a kapcsolattípusok gyakoriságát és arányát. Jóllehet ez az elemzés hasznos szolgálatot tehet, véleményem szerint az optimális és teljes visszacsatolást csak az információkereső rendszerben való használat, a kapott találati arány, a rendszer "érzékenysége" megfigyelése biztosítja majd.

3. Tezausz készítő programrendszer

Fenti módszertani követelmények messzemenő kielégítését szem előtt tartva készült el az itt ismertetendő programrendszer, az IBM 360-as számítógép PL/1 programozási nyelvén. A rendszer szerkezetét és szolgáltatásait az 1. ábra mutatja be. A programrendszer inputja /bevitt adatai, adatbevitel/ és outputja /kiviteli adatai, adatkivitel/ a következő formátum lehet a 2.1 alatt említett két munkamenetnek megfelelően:

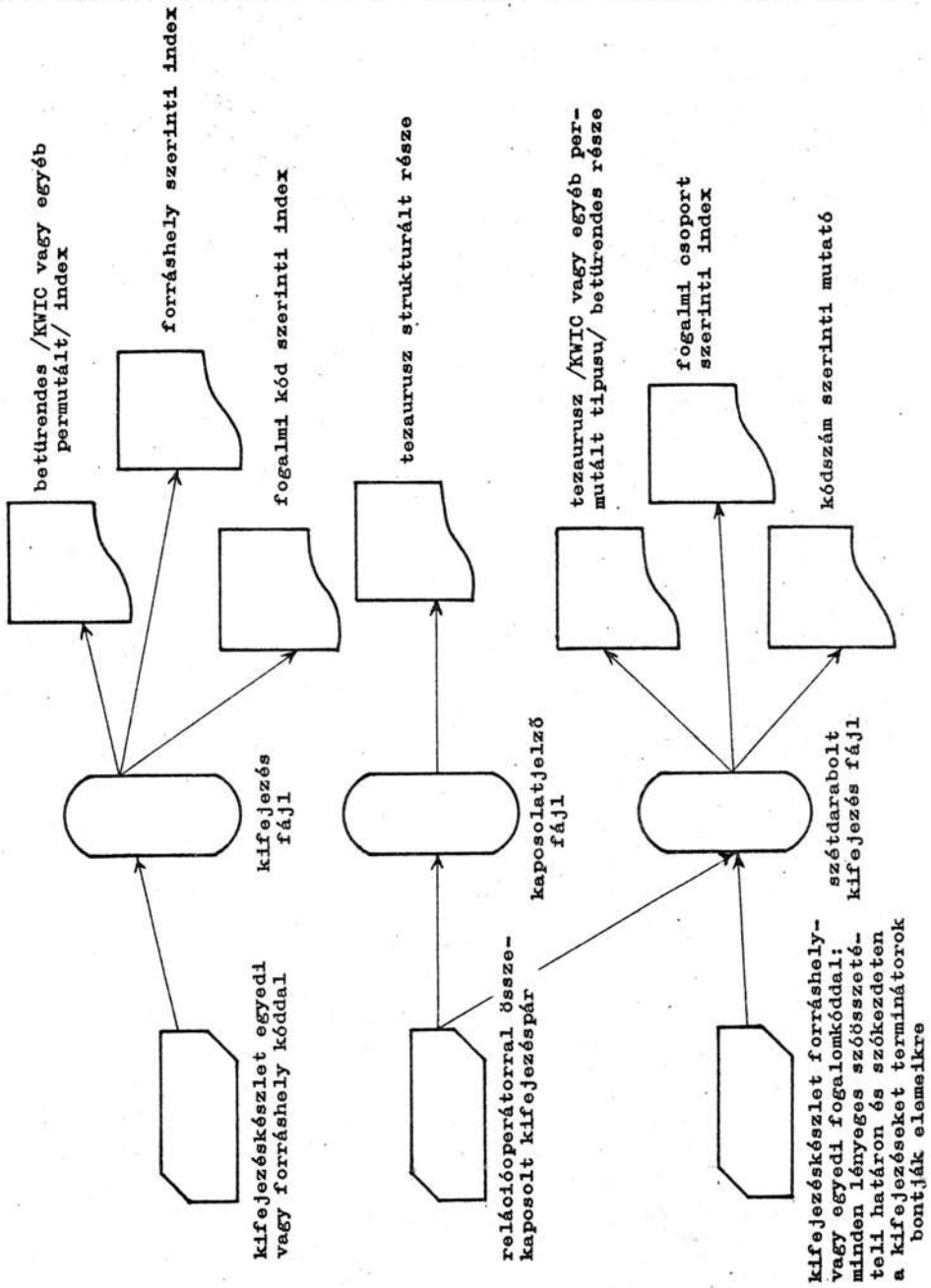
a/ Input: a sztruktúráltan kifejezés gyűjtemény a kifejezések forráshelyének megjelölésével és/vagy azonosító kódszámával van ellátva. Ha a kifejezések összes lényeges szóösszetételi és különirási helyeit is már ebben a fázisban ki akarjuk alakítani, akkor bármilyen terminátor jelet, illetve mesterséges szóelválasztást jelző jelölést alkalmazhatunk.

Az input egy tétele pl. a következő lehet:



Output: a programrendszert alkalmazó számítógép előállítja a kifejezések betürendes jegyzékét és/vagy forráshely, illetve kódszám szerinti jegyzékét is. Ha az összetett kifejezéseket a lényeges szóösszetevők szerint felbontottuk és a felbontás helyét terminátor jellel megadtuk, akkor KWIC rendszerű vagy egyéb permutált index is készíthető. A terminátorokkal elválasztott kifejezés-elemek száma kö-

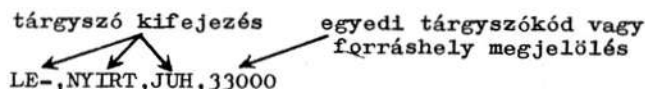
TEZAUROSZ KÉSZÍTŐ PROGRAMRENDSZER VÁZLATA



tetlen, ugyszintén ezek, valamint a forrásmegjelölés vagy /esetünkben az alfanumerikus/ kódszám hossza is. A betűrendezés természetesen minden esetben a magyar ábécé szerint történik.

b/ Input: strukturált kifejezés-gyűjtemény, amely a kapcsolatot jelző jellel, az un. reláció-operátorral összekapcsolt kifejezéspárok-ból áll. A kifejezések után egyedi kódszám vagy egyéb tetszőleges hosszúságú alfanumerikus jelsorozat állhat, előttük pedig a fogalmi csoportot jelölő, szintén tetszőleges hosszúságú jelsorozat. Mindezeket terminátorok választják el, a reláció-operátort pedig szintén külön e célra megállapított jelnek kell megelőznie.

Az input egy tétele pl. a következő lehet:



Output: A fenti inputból a program szerint ismét készülhetnek betűrendes, továbbá kódszámok szerinti vagy csoportosított és azon belül betűrendes kifejezés-jegyzékek. Ugyanezeket a betűrendes jegyzékeket permutált index formájában is készíthetjük, ha a fogalompárban az összes lényeges szóösszetéti helyen vagy szóhatáron terminátor-jellel, illetve elválasztójellel biztosítjuk a szétvágást. A betűrendes jegyzékekben ekkor összeírva jelennek meg a kifejezések, mert az összes terminátort, valamint egyéb segédjelet eliminálja a program.

Ebben a fázisban azonban - felhasználva az inputban szereplő reláció-operátorokat - a gép már olyan jegyzék/ek/et is előállít, amelyekben a kifejezések alatt azok szóbokrait is kiírja, a relációk szükség szerinti megfordításával. Amint az 1. ábrán látható, a tezauszus mindkét fő része gépi információhordozón tárolva továbbra is rendelkezésre áll, így bővítése, ismételt kiírása a fenti formákban, de információkereső rendszerbe való beépítése is bármikor lehetséges.

4. Kísérleti feldolgozás a rendszer kipróbálására

A programok kipróbálásához a mezőgazdaság fogalmi körébe tartozó kifejezéseket használtam fel. Ebben az esetben a tárgyszóanyag és a belőle előállított tezauszus semmi esetre sem tekinthető ellenőrzött, teljességre törekvő és helyesen kiválasztott fogalmi anyagnak, csupán illusztráció célját szolgálja, de az elkészített jegyzékek a tezauszus-készítés módszertanának helyes kialakításához segítséget adnak.

4.1 A programfutas ideje

A futási idő várható nagysága a

$$t = t_e + c \cdot n$$

képlet alapján becsülhető meg, amelybe multiprogramozás esetén még a többi program is beleszól, ezzel együtt adja az ún. falióra-időt /brutto gépidő a program elindulásától befejezéséig/.

A képletben t_e - a programbetöltés, fájl megnyitás; c - egy tétel /kifejezés vagy kifejezéspár/ feldolgozására jellemző konstans időszükséglet; n - a tételek száma.

Reális becslésre ez a képlet akkor használható, ha nem változik a relációféleségek és az egy kifejezéssel kapcsolatban /relációban/ álló más kifejezések átlagos száma, a kifejezések átlagos hossza és átlagos szegmentáltsága.

A mi esetünkben pl. 74 darabból álló kifejezőkészletre 35,86 másodperc a központi egység teljes feldolgozási ideje. Azonos jellegű kifejezőkészlet feldolgozásakor, ha ez a mennyiség pl. 100-szoros lenne, akkor a képletben n értéke 100-szoros lenne. A kifejezéspárok feldolgozási idejére a számítás menete teljesen azonos, de a más típusú feldolgozás miatt ott nyilvánvalóan más t_e és c érték is adódik. A konstansok ismeretében a gépidő-szükségletet a továbbiakban jól lehet becsülni.

4.2 A kísérlet során mért idők

Kifejezések száma:	74 db
Átlag szegmentszám kifejezésenként:	1,8
Kifejezéspárok száma:	78 pár

A központi egység feldolgozási ideje az első típusú /kifejezésenkénti/ feldolgozás esetén:

input	14,10 s
rendezés	10,46 s
kiíratás	<u>11,30 s</u>
összesen:	35,86 s

Kifejezéspáronkénti feldolgozás esetén:

input	17,20 s
rendezés /kifejezéspárookra/	<u>10,98 s</u>
	28,18 s

áthozat:	28,18 s
rendezés /kifejezésekre/	11,78 s
szóbokros rész kiírása	10,56 s
betürendes rész kiírása	11,04 s
összesen:	61,56 s

5. A programrendszer felhasználhatósága és felhasználói környezet

5.1 Hazai helyzetkép

A külföldi és magyar dokumentációs irodalomban nagyszámu publikáció foglalkozik a tezausz-kérdéssel. Az UNESCO, az ISO és a dokumentáció különféle szervezetei szabványjavaslatokat, illetve nemzetközi szabványt is dolgoztak ki. Magyar szabvány mind ez ideig nem jelent meg. Az OMKDK három kötetben közreadta az Általános műszaki fogalmak tezauszát, amely a hazai tezauszkészítésben uttörő szerepet töltött be, valamint kidolgozta a számítástechnikai fogalmak tezauszát. Ezek célja az is volt, hogy módszertani iránymutatóul szolgáljanak a hazai, magyar nyelvű tezausz-munkálatokban.

Egyre növekszik azon intézmények száma, ahol az Egyetemes Tizedes Osztályozás /ETO/ korlátai miatt kénytelenek tárgyszavas osztályozási rendszereket és tezauszokat kidolgozni. Ezek a hazai intézmények /pl. a Központi Statisztikai Hivatal Könyvtára/ hosszabb ideje gyűjtik a kifejezéseket, elsősorban cédulákon. A cédulák gyakori másolása, selejtezése, rendszerezése manuálisan egyre több időt rabol el a tezauszkészítés érdemi munkáitól, a gyűjtés körülményeitől, a helyes fogalom-kijelöléstől, a használhatóság ellenőrzésétől, statisztikai és egyéb elemző munkától. Egyrészt ez indokolja a számítógép fokozott igénybevételét tezauszok készítésekor. Másrészt az, hogy az információkereső rendszerek hatékony működtetésének a számítógépbe beépített és folyamatosan karbantartott tezausz a fő támasza. Mivel a dokumentációval foglalkozó intézmények egyre közelebb kerülnek a számítógépes információkereső rendszerek kiépítéséhez, egyre sürgetőbb feladat a lehetőleg gépi adathordozón kiegészíthetően rendelkezésre álló, de nyomtatásban is bármikor előállítható és használható tezauszok kialakítása. Az információkereső rendszer gyakorlati üzembe helyezésekor már kisebb feladatot jelent kész tezausz alapján indexelt dokumentumok és a hozzájuk tartozó tezausz beépítése a rendszerbe, mint annak a több évre terjedő alapos és nagy szakmai felkészültséget igénylő munkának egyszeri elvégzése, amit egy jó tezausz készítése jelent. Adatfeldolgozó berendezések, illetve software-ek bérlete esetén a tezausz hiányában kihasználatlanság miatt felmerülő költségek a rendszer bukását okozhatják. A hazai készítésű információkereső rendszerek fejlesztését hátráltatja, és ezek hatékonyságát is évekre leronthatja a tezausz-munkálatokban fennálló - a manuális módszerekre is visszavezethető - lemaradás.

6. További módszertani megjegyzések

A tezauszuszkészítés automatizálására a fent leírtaknál magasabb szinten is történtek kísérletek, de ezek eredményei egyrészt nagyon is kétségesek, másrészt nálunk ezek kipróbálására még nincs meg a következő alapvető feltétel: nem áll rendelkezésre olyan korszerű információkereső rendszer, amely nagyszámu, sok deskriptorral indexelt célinformációt tartalmazna, amelyekből statisztikai vizsgálatok útján, de emberi közreműködéssel is /pl. interaktív módban/ kialakíthatók lennének a deskriptor-kapcsolatok.

BALOGH, Z.: Programme system for thesaurus compilation written with PL/I programme language

Description of a programme system for mechanizing essential parts of formally manual operations in thesaurus compilation.

The first chapter includes the definitions of the structural units and concept relations most common in thesauri. These are:

hierarchical relation: generic and specific

preferential relation: "see" reference, forbidden, reference, association trails.

The programme system handles all these, further any other relations occurring during application.

Thesaurus compilation as well as maintenance of thesaurus files stored on magnetic discs could be made possible by the programme system in two directions:

elaboration and maintenance of subject heading lists complemented with concept codes or source indications: on this base KWIC or other permuted index types as well as subject heading indexes by codes are producible according to request;

terms belonging to subject heading lists could be included into a thesaurus arranged in pairs, linking with relation operators and attached with concept codes: programmes work out the structured part

of the thesaurus, produce the term-file as well as the alphabetical KWIC or the concept-code indexes.

Another advantage of the programme-system that terms could be segmented in compound words, parts with individual content meanings could be denoted by terminators and so they are represented in many places in all the indexes /according to all components/.

Method for the estimation of programme time is described, detailed quantitative and time data of experimental processing are given.



БАЛОГ, З.: Пакет программ для составления тезауруса, написанный на языке PL/1

В настоящей статье дается краткое описание пакета программ для ЭВМ, выполняющего автоматически значительную часть работ, связанных с составлением тезаурусов и выполненных до сих пор вручную.

В первой части статьи дается определение основных структуральных единиц тезауруса и смысловых связей, наиболее часто указанных в тезаурусах. Эти связи следующие:

- иерархические связи (вид-род, часть-целое);
- преференциальные связи ("смотри" и "употребляй вместо");
- ассоциативные связи.

Пакет программ обеспечивает ведение этих связей.

Пакет программ обеспечивает

- обработку массивов понятий, оснащенных специальными кодами или кодами, указывающими их место происхождения;
- составление из вышеупомянутого массива указателей типа KWIC или другого типа, и также списка понятий по их кодам;
- сочетание понятий, окончательно выбранных для участия в тезаурусе попарно вместе с их оператором связи и кодами;
- составление основной (алфавитно-иерархической) части тезауруса;
- составление массива понятий, указателей типа KWIC и списка понятий по их кодам.

Дальнейшим преимуществом пакета программ является то, что при отмечении составных частей сложных слов возможно их появление на разных местах (по каждой их части).

В конце статьи дается метод оценки времени, нужного для данного пакета программ и даются результаты его экспериментального применения.

BALOGH, Z.: Programmsystem in der PL/1 - Sprache für die Thesauruszusammenstellung

Der Artikel beschreibt ein Programmsystem, mit dessen Hilfe ein beachtlicher Teil der bisher manuell verrichteten Arbeiten bei der Thesauruserarbeitung mechanisiert werden kann.

Im ersten Kapitel sind die Definitionen der strukturellen Einheiten des Begriffs Thesaurus und sodann jene Begriffsbeziehungen zu finden, die in Thesauren am häufigsten vorkommen. Diese sind:

hierarchische Relationen: generisch, spezifisch

präferentielle Relationen: Hinweise - "siehe" Verbot-Hinweis

assoziative Beziehung.

Das Programmsystem ermöglicht die Handhabung von diesen sowie von beliebigen weiteren Relationen, die bei der Anwendung vorkommen können.

Das Programmsystem ermöglicht aus zwei Richtungen ausgehend den Prozess der Thesauruserarbeitung sowie die Instandhaltung der auf Magnetkarten gespeicherten Thesaurus-Files:

Bearbeitung und Instandhaltung von mit Begriffs-Kode und Quellenangabe versehenem Sachwortbestand: aus diesen können KWIC-Indexe oder permutierte Indexe anderen Typs sowie dem Kode entsprechende Sachwort-Indexe nach Bedarf hergestellt werden;

die in den Sachwortbestand endgültig aufgenommenen Ausdrücke mit Relationsoperatoren paarweise koppelnd und mit Begriffsgruppenkode versehen können diese in die Thesauren eingebaut werden; die Programme verfertigen den strukturierten Teil des Thesaurus, sie stellen weiterhin den Ausdruck-File und die erforderlichen alphabetischen KWIC-Indexe oder aufgrund von Begriffs-Kode aufgebauten Indexe zusammen.

Ein weiterer Vorteil des Programmsystems besteht darin, dass die Ausdrücke an den Wortfügungsstellen segmentiert werden können und die über selbständige inhaltliche Bedeutung verfügenden Teile des Ausdrucks mit Terminatoren bezeichnet werden und so in allen Indexen /ihren Komponenten entsprechend/ an mehreren Stellen vorkommen können.

Der Artikel teilt auch eine Methode zur Schätzung der Zeit des Programmablaufs mit, wobei auch die ausführlichen Mengen- und Zeitangaben einer experimentellen Bearbeitung mitgeteilt werden.