681.3.06:809.451.1

PROGRAMRENDSZER MAGYAR NYELVŰ SZÖVEGEK SZAVAINAK TÖVESÍTÉSÉHEZ

Balogh Zoltán
Infelor Rendszertechnikai Vállalat

1. A PROBLÉMA MEGFOGALMAZÁSA

Szöveges információtároló és -kereső rendszerek kialakitásánál gyakori probléma a tárolandó szövegek szakmai szóanyagának a szövegből való kigyüjtése, a rendszer müködéséhez szükséges szótárak, tezauruszok készitéséhez. A már létező szótárakban meg kell keresni a szövegben talált, ott többnyire toldalékokkal ellátott szavakat,hogy megállapitható legyen róluk, az adott szótő szerepelt-e már a szótárban vagy tezauruszban, vagy bőviteni kell-e vele a szótárak valamelyikét. Neheziti a megoldást, hogy a szöveg vizsgált szava ezenkivül hibásan is lehet leirva, vagy röviditett formában szerepelhet. Az eredményre mindenekelőtt a tezaurusz szóanyagának gyüjtésénél, az indexelésnél, kérdés elemzésénél vagy az input ellenőrzésénél lehet szükség. Mélyebb feltáró munkánál a szövegszó pontos szerkezetének, szófajának meghatározása is nagy szerepet játszik. A problémát egy számitógépet müködtető programrendszerrel kivántuk megoldani,amelylyel a magyar nyelvü, vagy ahhoz hasonló strukturáju /morféma szerkezetü/ szövegek automatikus elemzése lehetséges. Közelebbről a cél a szöveg szavai közül a jelentéssel nem biró vagy a szövegben tartalmat kifejezetten nem hordozó szavak felismerése, a kifejezést hordozó szavak nyelvtani besorolása, kötőhangjainak, jelentést nem módositó pre- és szuffixeinek levágása és azonositása, az esetleges tőhangváltozások, hasonulások felismerése és eredeti tövekre való visszavezetése. A felhasznált kódolási és kategória rendszereket nagyrészt az MTA Nyelvtudományi Intézete dolgozta ki.

2. A FELHASZNÁLT PROGRAMRENDSZER

Az IBM 360 OS PL/L nyelv F szintjén megirt programrendszer a fenti feladat végrehajtását az alábbi korlátozó feltételekkel oldja meg:

a/ A szóelemzéshez és felismeréshez közepes méretü, legfeljebb 5000 szavas szótárakat használ, amelyekben igen gyors keresési algoritmusokkal keres.

b/ A prefixeket, tőhangváltozásokat, hasonulásokat még nem kezeli a végcélként kitüzött módon. Az ilyen szavakat szótárak segitségével megkerülve dolgozza fel. A szuffixek előtt a kötőhangokat felismeri.

c/ A sorelválasztásból adódó szótöredékeket nem vonja össze, igy ezeket nem elemzi.

2.1 A felhasznált szótárak

2.11 A toldaléktár

A tövesitéshez, illetve az ezt célzó összehasonlitáshoz és felismeréshez a program szerint müködő számitógép háromféle szótárat használ fel; az un. toldaléktárat, a jelentéshordozó szótövek szótárát, a tőszótárat és a tartalom kifejezésében részt nem vevő szavak /szótövek/ szótárát: az un. nullszótárat.

Szerkezete:

1. toldalék /maximálisan 10 betü/

2. toldalék kódja /decimális sorszám/

3. Jelölés arra nézve, hogy a toldalék milyen szófaju tővel társul a szó végén, éspedig

a/ névmással,

b/ tulajdonnévvel,

c/ melléknévvel,

d/ számnevekkel vagy

e/ a szó elején

Ha a toldalék szócikkében a megfelelő pozicióban megfelelő 1-es áll, akkor társulhat az illető szófaju tővel, ha 0, akkor nem.

2.12 A jelentéshordozó tövek szótára /tőszótár/ és a tartalom kifejezésében részt nem vevő szavak szótára /null- vagy funkcionális szótár

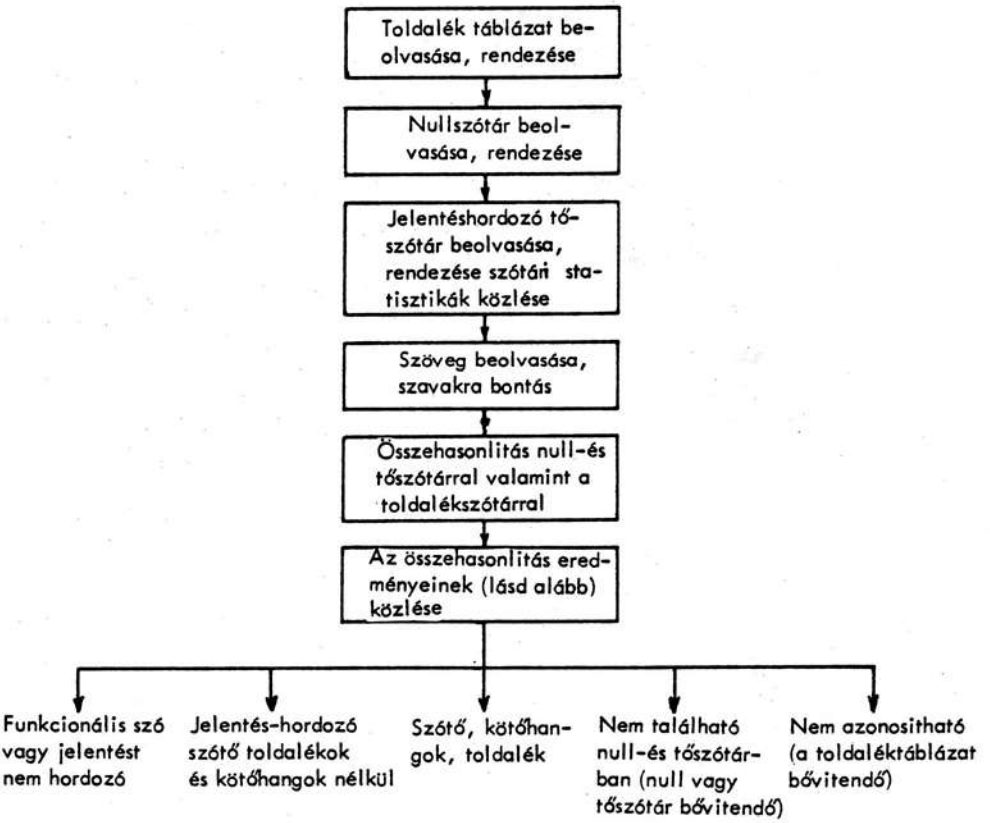
Szerkezete:

1. szótő /maximálisan 26 karakter/

2. szófaji kód /az MTA Nyelvtudományi Intézeténél alkalmazott két karakteres kódrendszer szerint, bővitésekkel/

Jellegét /szerkezetét/ tekintve a nullszótár is tekinthető tőszótárnak. A rövidség kedvéért azonban a továbbiakban a jelentéshordozó tövek szótárát nevezzük tőszótárnak.

2.2 A PROGRAMRENDSZER SZERKEZETE ÉS MÜKÖDÉSE



A fentiekben vázlatosan ábrázolt programok szerint müködő számitógép a folyamat eredményeként minden megvizsgált szövegszóról közli az ábrán feltüntetett "üzenetek" valamelyikét: a szótövet, a szófaji kódot, az esetleges kötőhangokat, a toldalékot, a toldalék összes kódját és megjelölését.

3. A KISÉRLETI ANYAG ÉS A FELDOLGOZÁS EREDMÉNYEI

Kisérleteink anyaga bármilyen szakszöveg lehetett volna, melyhez a megfelelő szótárak rendelkezésünkre állnak. Mivel a korábbi munkáink során jogszabályok szövegeit vizsgáltuk, ezekhez viszonylag egyszerüen el tudtuk készíteni a tő- és a nullszótárt. A program ellenőrzésére néhány soros számitástechnikai szöveget is vizsgáltunk, ehhez azonban a tőszótár nem állt rendelkezésre.

A jogi szöveg 27 soros volt, 80 pozíciós lyukkártyára volt lyukasztva. Ebből

valódi szövegszó	153 db
töredék	29 "
jogszabályszám /jelzet/ és ennek eleme	42 "

A számitástechnikai szöveg 3 soros volt. Ebből

valódi szövegszó	31 db
töredék, rövidités	8 "

Az alkalmazott szótárak tartalma a következő volt:

toldalék táblázat	19	toldalék
nullszótár	18	szótő
tőszótár	57	szótő /tulnyomórészt főnév/

Az elemzés /összehasonlitás/ eredménye a következő volt:

	Jogi szöveg	számitástechnikai szöveg
funkcionális szó	17	0
szótő toldalék nélkül	56	0
szótő toldalékkal	10	0
szótő kötőhangokkal és toldalékkal	55	0
null- és tőszótárban nem szerepel	13	31
töredék, hibás jel	29	0
jelzet része /jogszabályé/, rövidités	42	8
nem azonositható	2	0

Az elemzéshez szükséges programfutás ideje /központi egység idő/ 30.14 s volt.

4. A PROGRAM FELHASZNÁLHATÓSÁGA

A programrendszer fejlesztési stádiumban van, amit a biztató eredmények sürgetnek. A felhasználást egyelőre szükebb szakterületeken ajánljuk dokumentációs célokra, adatbázis kereső file-ok létrehozására, információkereső rendszerek létrehozása során speciális szótár /tezaurusz/ készitésére, szövegek automatikus elemzésének és feldolgozásának céljaira.

5. A KUTATÁSOK FOLYTATÁSA

A 2. fejezetben emlitett, jelenlegi korlátozások megszüntetését tekintjük célunknak, továbbá olyan esetleges ujabb problémák megoldását, amelyek a továbbiakban vizsgált nagyobb szöveganyagnál felmerülnek.

Az a/ korlátozás megszüntetését és a program teljes optimálását, esetleg BAL /Basic Assembler Language/ nyelvre való teljes vagy részleges átirását nem tekintjük fontosnak, amig az algoritmusok nem felelnek meg minden tekintetben a kitüzött célnak.

A szótárkezelés kiterjesztésének megoldása szintén elhalasztható, mivel a program egy-egy szükebb szakterület szövegeinek feldolgozásához, automatikus indexeléséhez kiván segitséget adni, ahol a maximum 5000 szócikkes szótárméretek kielégitőek lehetnek.

További célunk egy optimalizált szótárkezelő program, a korábban elkészült tezaurusz-készitő programrendszer, valamint a fent ismertetett és a kutatás során kidolgozott egyéb programok, programrendszerek összekapcsolása kisérleti szinten, adagolt üzemmódban.

BALOGH Z.: Programrendszer magyar nyelvü szövegek szavainak tövesitéséhez

A magyar nyelv szerkezete a prefixek és szuffixek alkalmazása miatt nem teszi lehetővé az angol nyelvterületen alkalmazható szöveganalizáló módszerek alkalmazását. Az információtároló és -kereső rendszerek, a szöveges információkat tároló adatbázisok szóanyagának elemzésénél gyakori probléma a szövegekben előforduló szóformátumok visszavezetése a szótövekre; a szuffixek előtt gyakori kötőhangok felismerése és leválasztása; a szuffixek és prefixek levágása; a szófajok felismerése; a már létező szótárakban található szavakkal való azonositásuk.

A programrendszer a jelzett problémák megoldását teszi lehetővé magyar nyelvü szövegeknél.

Segédeszközül felhasználja:

a szuffixek és prefixek táblázatát;

a szövegek formaszavainak un. nullszótárát;

a tartalmat hordozó szótövek szótárát.

Mindezek a szótárak automatikusan bővithetők a feldolgozások eredményeként. Ehhez azonban már az emberi kontroll szükséges. További problémát jelent a szuffixek hasonulása, valamint a szótövekből való hangkiesések, tőhangváltások megoldása.

A programrendszer IBM 360 gépen OS PL/1./F/ nyelven üzemel.

A rendszer szerkezetét és kisérleti feldolgozásának eredményét mutatja be a cikk.

BALOGH, Z.: Programme system of word stem forming from words in Hungarian texts

Because of the peculiar use of prefixes and suffixes the Hungarian language structure cannot apply English text analyzing methods. Analyzing word stocks of information storage and retrieval systems, textual information stored in data banks it is a frequent problem to trace back word forms to word stems; to recognize and separate frequently used glides before suffixes; to remove suffixes and prefixes; to recognize kinds of words; to identify them with words in dictionaries.

The programme system solves these problems regarding Hungarian texts. As aids the followings are used:

tables of suffixes and prefixes;

so-called null-dictionary listing form words of texts;

dictionary of content carrying word stems.

All these dictionaries could be automatically enlarged as a result of the processing. Human control is, however, indispensable. Further problem is the assimilation of suffixes and the solution of word stem elisions and phonetic changes.

The programme system is operating on IBM 360 with OS PL/1./F/ language.

Structure of the system and results of an experimental processing are described.

БАЛОГ, З.: Программа на ЭВМ выполняющая отсечение суффиксов и префиксов от слов венгерского языка

Структура венгерского языка из-за применения префиксов и суффиксов не позволяет использовать методы, аналогичные методам анализа текстов английского языка. При анализе слов информационно-поисковых систем и систем для хранения информации, часто встречаются следующие проблемы: сведение разных форм слов к их корням; опознавание и отсечение соединительных гласных, часто появляющихся перед суффиксами; отсечение суффиксов и префиксов; опознавание частей речи; сравнение слов со словами уже имеющихся словарей.

Предлагаемая программа позволяет решить для венгерского языка все вышеупомянутые проблемы.

Программа использует следующие вспомогательные средства:

словарь суффиксов и префиксов;

так называемый нулевой словарь формальных слов текстов;

словарь корней слов, отражающих содержание.

Все эти словари могут быть расширены за счет результатов обработки, но при этом необходим интеллектуальный контроль. Дальнейшими проблемами являются асоимиляция суффиксов и опознание некоторых звуков или же изменение корней при прибавлении суффиксов.

Программа написана на ЭВМ 36о на языке OS PL/1./F/.

В данной статье описывается структура программы и приводятся результаты экспериментальной обработки текста с помощью этой программы.

BALOGH, Z.: Programmsystem zur Wortstammbildung in ungarischen Texten

Die Struktur der ungarischen Sprache schliesst wegen der ihr spezifischen Anwendung der Präfixe und Suffixe die auf dem englischen Sprachgebiet verwendbaren Textanalysemethoden aus. Bei der Analyse des Wortmaterials für Informationsspeicher- und -recherchesysteme sowie für textliche Informationen speichernde Datenbasen bildet es eine Reihe von Problemen, dass die vorkommenden Wortformen auf die Wortstämme zurückgeführt werden müssen; die vor Suffixen häufigen Bindelaute erkannt und abgetrennt werden müssen; die Präfixe und Suffixe abgetrennt werden müssen; die Wortgattungen erkannt werden müssen und diese mit den in bereits vorliegenden Wörterbüchern enthaltenen Worten identifiziert werden müssen.

Das Programmsystem ermöglicht die Lösung dieser Aufgaben in Bezug auf ungarische Texte.

Als Hilfsmittel werden

Tabellen der Suffixe und Präfixe,
das sog. Null-Wörterbuch der Formwörter in Texten,
das Wörterbuch der Inhalt tragenden Wortstämme

benützt. Als Ergebnis der Bearbeitung können all diese Wörterbücher automatisch erweitert werden, wozu jedoch menschliche Kontrolle erforderlich ist. Die Identifizierung der Suffixe, der Lautschwund aus den Wortstämmen und die Stammlautveränderungen bilden jedoch noch weitere Probleme.

Das Programmsystem läuft auf IBM 360 in der OS PL/1./F-Sprache.