

PAPP FERENC

Budapest, az MTA Nyelvtudományi Intézete

A MAGYAR FŐNÉV ESETRAGOS ALAKJAINAK AUTOMATIKUS SZINTÉZISÉRŐL *

O. Az alábbiakban megismertetem az olvasót azokkal a legfontosabb nyelvi tényekkel, amelyek a címben foglalt feladat megoldásához kellene. Pontosabban: ezeket a tényeket minden egyes magyar anyanyelvű olvasó jól, "von Haus aus" ismeri: csupán e tények pontos-szigorú számbavétele, algoritmikus rendbe rakása szükségeltetik, épp az automatikus, tehát esetenkénti (ad hoc) beavatkozást nem igénylő szintézis céljából. Az alábbiakra (a gyakorlatban) alig lesz szükség, legfeljebb gépi fordítás, kivonatolás stb. magyar kimeneténél kellene, egyebek mellett, az automatikus szintézis adatai és szabályai. Itt a gép azért kell, hogy rajta ellenőrizzük ismereteinket: vajon valóban mindent tudunk-e a magyar morfológia e fejezetéből? Hiszen a gép ilyen szempontból könyörtelen, nincs nyelvérzéke: ha valamit nem vagy nem a kellő módon közöltünk vele, akkor persze rossz alakokat fog kiadni.

1. Az első, amit létre kell hoznunk, a magyar betűk -- nevezzük őket a továbbiakban karaktereknek (ezzel e szavunk új jelentést kap, angol szemantikai kölcsönzés). Így elkerültük a magyar *betű* szó kétértelműségét: míg a köznyelvben durván szólva egy-egy "leütést" jelöl (tehát ott a *csók* szó négybetűs), addig a helyesírási szakirodalomban a betű egy fonéma jele (a *csók* tehát hárombetűs). Látnunk kell, hogy a latin alapkarakterek az angolban annak hieroglifikus írásrendszere miatt elegendők, vö. a *read* karaktorsor (string) különféle olvasataival, a *blood--foot* stringek *oo* szakaszának fonémamegfelelésével, az *enough* string fantasztikus fonémamegfelelésével stb. Úgyhogy nem olyan biztos, hogy az angol írásbeliség oly ideális, amilyennek tisztán számítógépes szempontból első pillantásra látszanék -- nem minden fenéig számítógép.

Az alapkaraktereken kívül tehát létre kell hoznunk az ékezeteseket. A többjegyű (egyenként több karakterből álló: *cs*, *dzs*, *ddzs*) betűk azért visszatérnek. Közülük célszerű csak a kétjegyűekkel számolni, a három- és négyjegyűeket kivé-

* Elhangzott a Magyar Nyelvtudományi Társaság Heves megyei csoportjának felolvasóülésén 1991. április 16-án.

telként félretenni, külön listán tárolni. A kétjegyű elemeket valamiképpen kezelünk kell, éreztetve a géppel is egységüket: kétlépéses ciklusban dolgozzuk fel őket egy olyan egyszerűbb programozó nyelvben, mint a basic, s így, páronként hasonlítjuk össze a feldolgozandó szó karakterpárjaival; külön deklaráljuk őket a fejlettebb nyelvekben.

Az itt tárgyalandó feladathoz a betűrendbe állítás nem szükséges. Az ékezetes betűket és a kétjegyűeket inkább saját céljainknak megfelelően fogjuk sorba rakni. Így például az ékezeteseket (melyek nálunk szerencsés módon csak egyes magánhangzó fonémák jelölésére fordulnak elő -- vö. pl. a csehvel, ahonnan vették: *e, c, s*, stb.) úgy helyezzük el, hogy előbb álljanak a mélyeket jelölők: *a, á, o, ó* stb., majd a magasak, végül a labiális magasak -- könnyen belátható, mely okból. Így a „sima”, egylépéses összehasonlítgatás során mindjárt az fogja elárulni, milyen jellegű magánhangzó van az elemzett szóban, hogy a minta hanyadik magánhangzó elemével sikerült azonosítani: ha az 1--6. valamelyikével, akkor mély, ha magasabb sorszámúval -- akkor magas vagy semleges, és így tovább.

2. További információk, melyekre szükségünk van a szintéziskor:

a/ A fentiekben érintett csoportosítgatások során kiderül egy az ott érintett-hez képest sokkal egyszerűbb dolog: vajon magánhangzóra végződik-e egyáltalán a kérdéses szó? Erre az információra szükségünk lesz a superessivus megfelelő allomorfjának kiválasztásához. Itt kell elintéznünk a *brandy, guillotine*-féle kivételeket: az előző magánhangzóra végződik, az utóbbi nem, a látszat (a /szóvégi/ karakter) ellenére. (Az *y* betűt persze felsorolhattuk volna a magánhangzók között. Akkor a *gentamycin* illeszkedési osztályát tekintve nem lett volna kivétel: az *y* nem semleges karakter, mint édestestvére, az *i*. Ám ez az eljárás valószínűleg mégis több bajjal járt volna, mint haszonnal.) És így tovább, alább még látunk példát ennek az információnak a felhasználására.

b/ Mely illeszkedési osztályhoz tartozik? Míg a hangrendet mechanikusan, az illeszkedési osztályt csak egyedi elemzés után nyerjük általános esetben: mély, magas, ingadozó. (Mi a szintézis során csak az első kettőt vettük figyelembe: az ingadozók vagy az első, vagy a második osztályba sorolódtak.)

Mi több: az illeszkedés (*i* osztály) megállapításakor figyelembe kellett vennünk az ÉrtSz. (elektromechanikus) gépi feldolgozásának eredményeit is. A *hid*-tól a már említett *brandy*-n át az *empire*-ig épp e feldolgozás alapján tudtuk egyrészt az illeszkedés szabályát gazdaságosan megállapítani, másrészt egy jó kivétellistát összeállítani.

További kérdéseink a ragra vonatkoztak:

c/ Nyújtja-e a rag a nominatívusi alakot? A nem nyújtók voltak kevesebben: *-képpen, -kor, -ként*. De egyáltalán: mely nominatívusi végződést nyújt a ragok többsége? Az *o-t* is (*eszpresszo--eszpresszóban, allegro--allegróval --* vagy már az alapalak is hosszú *ó-val*?)?

d/ Hány alakú a rag?

Egyalakúak; *-ként, -kor, -ért, -ig*. Ahogy ezen esetek valamelyikének a képzését kapta feladatul a gép, azonnal, az illeszkedés vizsgálata nélkül alkotja ezen alakokat.

Kétalakúak: *-ban/-ben, -ból/-ből...*

Háromalakú: *-hoz/-hez/-höz*. Itt használjuk fel azt az ismeretünket, hogy míg a veláris-palatális illeszkedés a *tő* született s megfoszthatatlan sajátja (*indítékom*, mert az *indít* fiktív töve mély illeszkedésű), addig a palatálisok között a labiális-illabiális illeszkedés az alakgenerálás sajátja: *földön*, de *földemen*: a kétfajta illeszkedést tehát másutt s másutt végeztük el a programban.

Egész sajátos kérdéseket vetett fel a két *v-s* ragunk, az *instr-é* és a *factivu-é*: (i) magánhangzóra végződik-e a *tő*? (*almával, epével*), /ii/ két- vagy háromjegyű betűre végződik-e: *gennyé, lánnyá*, /iii/ *x-re* végződik-e? (*bóraxszal, főnixszé*).

3. A fentiek során többször utaltunk rá, hogy ezeket a nyelvi ismereteket kellett megfelelően felfűznünk egy algoritmusra, s akkor megkaptuk: *ember+dat* = EMBERNEK, *indíték+delativus* = INDÍTÉKRÓL, *gentamicin+causalis* = GENTAMICINÉRT, és így tovább.

The first part of the document discusses the importance of maintaining accurate records of all transactions and activities. It emphasizes that proper record-keeping is essential for ensuring transparency and accountability in financial operations. This section also outlines the various methods and tools used to collect and analyze data, highlighting the need for consistency and precision in data collection.

The second part of the document focuses on the analysis of the collected data. It describes the various statistical techniques and models used to interpret the data, including regression analysis, time series analysis, and hypothesis testing. This section also discusses the challenges associated with data analysis, such as missing data, outliers, and the need for appropriate statistical tests.

The third part of the document discusses the application of the analysis results. It describes how the findings are used to inform decision-making and to identify areas for improvement. This section also discusses the importance of communicating the results of the analysis to the relevant stakeholders and the need for ongoing monitoring and evaluation.

The fourth part of the document discusses the future of data analysis and the role of technology in this field. It describes the various emerging technologies, such as artificial intelligence, machine learning, and big data, and how they are being used to improve data analysis. This section also discusses the challenges associated with these technologies and the need for ongoing research and development.

The fifth part of the document discusses the ethical implications of data analysis. It describes the various ethical issues, such as privacy, security, and bias, and how they can be addressed. This section also discusses the need for ongoing education and training in data ethics and the role of regulatory bodies in ensuring ethical standards.

The sixth part of the document discusses the conclusion of the study. It summarizes the main findings and the implications of the study. This section also discusses the limitations of the study and the need for further research.

The seventh part of the document discusses the references. It lists the various sources used in the study, including books, articles, and online resources.

The eighth part of the document discusses the appendix. It contains the various tables, figures, and charts used in the study.

The ninth part of the document discusses the index. It provides a list of the various topics covered in the document and the page numbers where they can be found.

The tenth part of the document discusses the glossary. It provides definitions for the various terms used in the document.

The eleventh part of the document discusses the acknowledgments. It thanks the various individuals and organizations that have supported the study.

The twelfth part of the document discusses the disclaimer. It states that the study is for informational purposes only and does not constitute an offer of any financial product or service.

The thirteenth part of the document discusses the contact information. It provides the name, address, and phone number of the author.