Приближени функции. Двадесет години от кончината на професор Мирослав Янакиев

Александър Иванов

...който от вас е без грях, нека пръв хвърли камък върху нея.
Иоан 8:7

На 9 ноември 2018 година се навършват 20 години, откакто загубихме проф. Мирослав Янакиев.

И да – годишнина. И да – би трябвало да напиша нещо, което да ухае на хризантеми, да погребем за пореден път човека. За пореден път да се убедим, че не сме лоши хора – ето, помним нашите учители...

Ама това никак не подхожда на Мирослав Янакиев.

Той е очертал такава широка научна нива, че ние не само не сме я още разорали, ами не сме я огледали още, границите ѝ не проумяваме още. Та, струва ми се, малко преждевременно е да го погребваме още веднъж.

Затова аз реших да отбележа годишнината от кончината на Мирослав Янакиев с критика (и поправка) на негова грешка. Грешка лоша, защото е свързана с практически инструкции. И ако спазвате препоръките му, ще направите още по-големи грешки.

В [Котова, Янакиев 1979: 243] авторите пишат: „При това по данни, които дължим на студенти, работили под ръководството на К. Найденов и М. Котаров, имаме основание да твърдим, че в съвсем различни по стил текстове отношението между броя на буквите и броя на монофоните варира съвсем незначително: в три извадки всяка по 200 килофони – едната от белетристични, другата от публицистични, третата от научни текстове – се оказаха съответно средно по 1017, 1017 и 1009 букви в килофона или общо средно по 1014 букви (и с вероятност 99,7%, т. е. с практически пълна сигурност, не по-малко от 1001 и не повече от 1027 букви). Затова средно с грешка само от около 2% можем да смятаме текст, съдържащ 1000 букви, за килофона.“

И отдолу в бележка под линия продължават: „Но доколкото грешката е систематична, т. е. доколкото буквите в една килофона са сигурно повече от 1000 и средно 1014, ако не работим с машина, а работим с хора, вместо да ги затрудняваме да държат сметка за всяка дифона или нулофонна буква, разумно ще бъде да искаме от тях да отброяват по 1014 букви, т. е. да приемаме, че килофонът е текст, съдържащ 1014 букви. Нещо повече, когато работим с белетристични или публицистични текстове, грешката ще бъде съвсем пренебрежима, ако отброяваме по 1017 букви за килофона.“

Само че това няма как да бъде вярно. Практически невъзможно е в една килофона броят на буквите да прехвърли 1000. Защо?

При преброяване на килофона се спазват следните зависимости (впрочем, те се изучават още в училище под тема „съотношение между звук и буква“):

я → монофона (след буква за съгласна), иначе – дифона;
ю → монофона (след буква за съгласна), иначе – дифона;
щ → винаги дифона;
ь → винаги нулофона;
дж → монофона (когато не е на морфемна граница), иначе – дифона;
дз → монофона (когато не е на морфемна граница), иначе – дифона.

Вижда се, че буквите могат да станат повече от монофоните при изобилие на буквосъчетания (дилитери) *дж* и *дз*, когато въвеждат монофона, и при изобилие на буква (монолитера) *ь*. Обаче сумарната честота на дилитерите *дж*, *дз* (при това без разлика дали са монофонни или дифонни) и на монолитерата *ь* в български текст е с порядък по-ниска от честотата на *щ*, което винаги въвежда дифона.

Като прибавим към това дифонната интерпретация на *ю* и *я* (най-честа е тя при литерата *я*, която участва в писменото представяне на граматически морфеми, тоест на чести морфеми, сихноморфеми по Янакиев), става ясно, че килофоната, измерена в букви, няма как да „прехвърли“ 1000 букви.

Как се е достигнало до такова недоразумение? Аз се опитах да намеря повече информация за изследването на К. Найденов и М. Котаров, предположих, че то е провеждано по НИС и че там би трябвало да има заключителен научен доклад и направих запитване (от 19.X.2015 г.), но отзвук от тази университетска институция нямам.

Затова се наложи да направя допускане. Предположих, че данните, които цитират Котова и Янакиев, представляват брой на монофоните в проби от по 1000 букви (килолитери). Като се вземе предвид равнището на компютърната техника и развитието на софтуера от онова време, това предположение изглеждаше убедително.

Значи – на студентите се предоставя текст от 1000 букви (килолитера) и им се поставя задача да преброят монофоните в него.

Ако в работата на К. Найденов и М. Котаров задачата е била поставена по този начин, съвсем ясно става и объркването на Н. Котова и М.Янакиев, за които интересна беше килофоната като универсална единица за измерване на „езикова материя“, а не килолитерата (1000 букви), за която не може да се види никакво филологическо приложение.

Аз симулирах (повторих) според това допускане изследването на К. Найденов и М. Котаров и ето какви резултати се получиха[1]:

Таблица 1. Монофони в килолитера.

Тип текст	Размер на извадката в килолитери (N)	Средна аритметична за брой на монофони в килолитера ($\bar{x}$)	Средно квадратично отклонение (σ)	Стандартна грешка на средната аритметична ($\sigma_{\bar{x}}$)	Размах на данните	
Публицистика	200	1017.030	5.151	0.364	1003	1041
Белетристика	200	1014.485	4.936	0.349	1002	1027
Научен текст	200	1014.170	5.302	0.375	1004	1033

Тези данни се съгласуват много добре с данните, изнесени от Котова и Янакиев, та смятам предположението си (че данните се отнасят за бройката на монофоните в килолитера) за потвърдено.

Да видим сега как ще изглеждат нещата, ако оценим размера на килофоната в букви (литери).

Таблица 2. Монолитери в килофона.

Тип текст	Размер на извадката в килофони (N)	Средна аритметична за брой букви в килофона ($\bar{x}$)	Средно квадратично отклонение (σ)	Стандартна грешка на средната аритметична ($\sigma_{\bar{x}}$)	Размах на данните	
Публицистика	200	983.300	4.777	0.338	968	995
Белетристика	200	985.805	4.875	0.345	973	999
Научен текст	200	986.060	5.336	0.377	969	996

[1] Всички материали по това изследване са достъпни като текстови файлове в архива miryan.org/download/20k.zip. Там има описание (във файла opis.txt) как са формирани трите извадки и от какви източници. Всяка извадка е създадена от проби по 20 килолитери (20 000 букви), случайно подбрани от по 10 случайно подбрани текста в съответния „жанр“.

И сумарно за трите извадки $\bar{x}$ = 985.055 и σ = 5.449.

Потвърждава се изводът, че „имаме основание да твърдим, че в съвсем различни по стил текстове отношението между броя на буквите и броя на монофоните варира съвсем незначително“. Потвърждава се и изводът, че ако отброявате **985** (но не 1014!) букви и ги приемате за килофона, грешката ще бъде по-ниска от 2%. При това грешката ще бъде в двете посоки – положителна и отрицателна – и при формиране на извадка (например от десет килофони) грешките взаимно ще се „изяждат“.

Между средните аритметични на белетристичните и научните текстове t-критерият на Стюдънт не показва значимо различие. Но между публицистичните и белетристичните текстове (t = 5.19) и между публицистичните и научните (t = 5.45) различието е значимо. Затова при обработка само на публицистични текстове е разумно да използвате **983** букви като размер на килофоната.

Мисля, че Мирослав Янакиев би бил доволен, че е изчистено и коригирано това недоразумение. В крайна сметка той най-добре знаеше, че науката се развива, като преодоляваме грешките на учителите си. За да направим своите грешки.

Какво следва?

Приближени функции се използват отдавна във всички „пресмятащи“ дисциплини. Но доколкото ми е известно, във филологията, в глотометрията започва да ги използва за пръв път Мирослав Янакиев. Причината да се използват приближени функции е разяснена много добре в цитираната по-горе бележка под линия – пести се труд и време.

В [Котова, Янакиев 1978] се казва, че „за много езици (например за всички славянски с изключение на сърбохърватския, защото сърбите и хърватите не отбелязват задължително по различен начин дългите и кратките гласни) фонометризацията на правописно записаните съобщения може лесно да се автоматизира“.

Да, може! И за български аз съм направил такъв „питонски обект“, който отброява килофони в български текст. Той е в модула *gtools* (*glottometrical tools*) на miryan.org и е, разбира се, свободен за използване, както и всичко друго, достъпно на страниците, посветени на професор Мирослав Янакиев. С тези програмни средства аз „броя“ килофони и в данните за тази статия. Впрочем и в този „обект“ – *KphnBG()* – аз използвам приближена функция за обработването на двубуквените съчетания (дилитерите) *дж* и *дз*.

Но какво правим със сръбски и хърватски? Как да „отброяваме“ килофоните там?

Най-простото решение е да се създаде такава приближена функция – да делим текста на определен брой букви (литери), които да приемаме за килофона. Но при това е нужно да имаме достатъчно надеждна оценка на

грешката. И ако някой колега сърбохърватист се захване с тази работа, той може да разчита на моята пълна подкрепа.

Знам, че Мирослав Янакиев беше създал и приближена функция за измерване на вербалната температура. Той предупреждаваше, че за да работи тя, пробите трябва да бъдат по-големи от хектолекса. За съжаление, не съм намерил описание на тази функция в документите на Мирослав Янакиев. Но си струват усилията тя да бъде „възстановена“, както и да се помисли за създаване на приближени функции за други глотометрически характеристики.

Библиография

Котова, Янакиев 1978: Котова, Н. В., М. Янакиев. Глотометрията експлицира основите на съпоставителната лингвистика. – *Съпоставително езикознание*, III, 1978, № 3, 3–15.

Котова, Янакиев 1979: Котова, Н. В., М. Янакиев. Глотометрия на фонетичните стойности на буквата *а* в историята на българския език. – В: *Изследвания върху историята и диалектите на българския език. Сборник в памет на чл.-кор. Кирил Мирчев*. София, 1979, 244–249.

e-mail: sashodi@gmail.com
бул. „Черни връх“ № 20
София 1421