ЗАЛЕЖНОСТІ РАНГ–ЧАСТОТА ДЛЯ СИМВОЛЬНИХ N-ГРАМ У ПРИРОДНИХ ТЕКСТАХ

О. С. Кушнір, О. І. Мацюняк

кафедра оптоелектроніки та інформаційних технологій,
Львівський національний університет імені Івана Франка
вул. Тарнавського, 107, 79017 м. Львів, Україна
e-mail: o_kushnir@franko.lviv.ua

Одним із широко досліджуваних об'єктів у статистичній лінгвістиці є рангові залежності випадкових величин, якими можуть бути, наприклад, абсолютна або відносна частота f появи лінгвістичного елемента (букви, символу, складу, слова, N-грами тощо), від її рангу r, тобто порядкового номера даного елемента в спадному списку всіх елементів за частотою. Відомо [1–3], що залежність ранг–частота для літер наближено описується логарифмічною функцією:

$$f(r) \propto \log(r^{-1}). \tag{1}$$

З іншого боку, чи не найбільш досліджені на сьогодні рангові залежності для частоти слів із задовільною точністю визначаються законом Ціпфа, тобто відповідні функції $f(r)$ мають степеневий характер:

$$f(r) \propto f^{-\alpha}, \tag{2}$$

де α – це стала Ціпфа (див., наприклад, [2]). Нарешті, з літератури відомий ще один нетиповий приклад рангових залежностей для частоти "слів" із деяких східних мов (наприклад, китайської або корейської) [4]. Це експоненційна залежність, притаманна мовам з обмеженими розмірами "словника" ідеограм або фонограм:

$$f(r) \propto \exp(-ar), \tag{3}$$

де a – деяка константа.

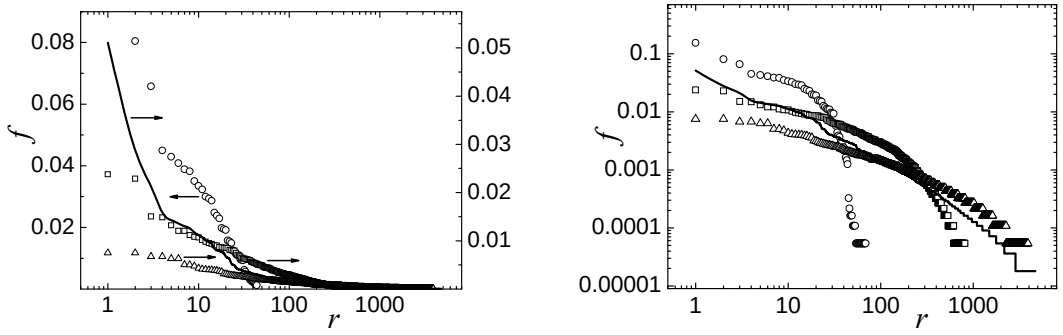
Статистичні або лінгвістичні причини таких якісно різних рангових залежностей досі майже не вивчено; незрозумілі і їхні взаємозв'язки, а також узагальнюючі функції, що допускають можливості граничних переходів поміж (1), (2) і (3). Принаймні якісно, ситуація нагадує відмінності поміж різними класами універсальності для критичної поведінки конденсованих систем. Скажімо, виразові (2) можна грубо поставити у відповідність степеневий характер поведінки деякого термодинамічного параметра при наближенні до критичної точки, а тому істотні флуктуації цього параметра, а виразові (1) – пригнічення флуктуацій і логарифмічну розбіжність зі зменшенням відносної температури. Попри всю умовність згаданих аналогій, поза сумнівом залишається одне – питання глибинних причин функціональних залежностей (1)–(3) потребує докладного і всебічного вивчення.

Метою цієї роботи було дослідження рангових залежностей для символьних N-грам – лінгвістичних структур, які складаються з N послідовних букв або символів (пропуск, кома, крапка, знак оклику тощо) – уніграм (N = 1), біграм (N = 1) і т. ін. Застосування N-грам має широко відомі практичні перспективи, зокрема у визначенні схожості чи відмінності тематики текстів [5] і розпізнаванні мов [6]. Наша основна гіпотеза полягала в тому, що в границі достатньо великих N ми мали би перейти від рангових залежностей типу (1) до якісно інших залежностей типу (2).

Для досліджень було використано текст новели Григорія Косинки "Політика" з орієнтовним обсягом 15,1 тис. знаків, у кодуванні UTF8. За загальноприйнятою методикою,

великих і малих літер у тексті не розрізняли. Для обробки природного тексту було створено програму в середовищі Qt, v. 5.3.2. З міркувань скорочення часу розрахунків ми обмежилися розглядом N-грам із N = 1÷3.

На рис. 1 показано рангові залежності окремо для уніграм, біграм, триграм і об'єднані рангові залежності для всіх вивчених N-грам. Усі результати представлено в напівлогарифмічному масштабі (див. рис. 1а), що відповідає гіпотезі, вираженій формулою (1), а також у подвійному логарифмічному масштабі (рис. 1б), що означає перевірку гіпотези (2). Спроба лінійної апроксимації за формулою (1) для рангових залежностей уніграм, біграм, триграм і об'єднаної залежності дає коефіцієнти детермінації R^2 відповідно 0,94, 0,93, 0,86 і 0,54. Це означає, що функція $f(r) \propto \log(r^{-1})$ найкраще, хоча й далеко не ідеально, описує дані для уніграм і дещо гірше – дані для біграм. Вона фактично не описує даних для триграм і об'єднану залежність $f(r)$.



(а) (б)

Рис. 1. Залежності від рангу відносної частоти уніграм (кола), біграм (квадрати) і триграм (трикутники), а також об'єднані рангові залежності $p(f)$ для N = 1÷3 (суцільні лінії), представлені в напівлогарифмічному (а) та подвійному логарифмічному (б) масштабах.

З іншого боку, залежність $\log f(\log r)$ для уніграм доволі далека від прямої, тобто функція $f(r)$ очевидно не степенева. Зі зростанням числа N ця функція для окремих N-грам, представлена в подвійному логарифмічному масштабі, поступово "лінеаризується", хоча й не повністю. Іншими словами, за цих умов рангові залежності наближаються до степеневої $f(r) \propto f^{-\alpha}$. Справді, лінійній апроксимації $\log f(\log r)$ для триграм відповідає помірна величина коефіцієнта $R^2 \approx 0{,}968$. Він стає ще більшим за умови об'єднання всіх N-грам ($R^2 \approx 0{,}984$). Така якість апроксимації загалом задовільна, оскільки з літератури відомо, що навіть рангові залежності для слів часто нелінійні в подвійному логарифмічному масштабі. Як правило, лінійність найкраще дотримана в центральній області незалежної змінної – в т. зв. "області ядра", яка виключає діапазони надто низьких рангів, де нахил кривої $\log f(\log r)$ помітно менший, і високих рангів, де спостерігаємо східчасту поведінку. У нас "область ядра" наближено відповідає діапазонові $r \sim 20\div500$, де визначений за нахилом показник степеня становить $\alpha \approx 0{,}93$. Це не надто відмінно від теоретично передбаченого Ціпфом одиничного значення.

Нарешті, можна очікувати, що описана вище еволюція кривих $f(r)$ стане ще виразнішою з подальшим зростанням N. З точки зору лінгвістики, тоді маємо перехід од логарифмічних до степеневих рангових залежностей, глибинною причиною якого, можливо, є перехід від лінгвістичних знаків суто символьного характеру (літер або розділових знаків)

до лінгвістичних знаків, навантажених семантикою (достатньо довгих послідовностей літер, яким можна приписати деякий зміст, або слів).

На думку автора піонерської роботи з опису рангових залежностей для лінгвістичних знаків С. М. Гусейн-Заде [1], наближено логарифмічна залежність $f(r)$ для літер зумовлена тим, що останні представляють собою систему невзаємодіючих об'єктів – на кшталт молекул ідеального газу. Відповідно, система слів, функція $f(r)$ для яких має альтернативний степеневий характер, мала би бути "взаємодіючою" в деякому сенсі. З іншого боку, емпіричні дослідження природних текстів засвідчили, що далекосяжні взаємодії та кореляційні явища, визначені за аналогією до відповідних явищ у статистичній фізиці, притаманні і літерам [7], і словам [8], а тому жодна зі згаданих систем лінгвістичних елементів, строго кажучи, не позбавлена взаємодій. Отже, з'ясування принципових відмінностей між ранговими розподілами літер і слів потребує подальших досліджень і залучення нових ідей.

Зазначимо, що вже після початку наших досліджень рангових залежностей N-грам (листопад–грудень 2014 року) ми ознайомилися зі статтею Л. К. Ха та ін. [9], у якій було наведено повніші (аж до N = 15) дані для суто буквенних N-грам із англомовних корпусів і зроблено висновок про те, що функція $f(r)$ для сукупності всіх N-грам справді є степеневою, причому показник степені $\underline{\alpha}$ виявився близьким до одиниці з доволі високою точністю. Проте автори праці [9] не зачіпали питання про еволюцію функціональної залежності $f(r)$ зі зростанням N, тобто основний предмет цієї роботи.

[1] S. M. Gusein-Zade. Probl. Peredachi Inform., 24, 102 (1988).
[2] I. Kanter, D. A. Kessler. Phys. Rev. Lett., 74, 4559 (1995).
[3] W. Li, P. Miramontes, G. Cocho. Entropy, 12, 1743 (2010).
[4] L. Lu, Z.-K. Zhang, T. Zhou. Sci. Rep., 3, 1082 (2013).
[5] M. Damashek. Science, 267, 843 (1995).
[6] G. Windisch, L. Csink. Proc. 2nd Romanian–Hungarian Joint Symposium on Applied Computational Intelligence, Timosoara, Romania, P. 243–255 (2005).
[7] W. Ebeling, A. Neiman. Physica A, 215, 233 (1995).
[8] M. A. Montemurro, P. A. Pury. Fractals, 10, 451 (2002).
[9] L. Q. Ha, P. Hanna, J. Ming, F. J. Smith. Artificial Intelligence Rev., 32, 101 (2009).