# СЕКЦІЯ 3.
## *Методи та засоби ущільнення інформації*

# Подільні коди та їх застосування

Анатолій Анісімов, Ігор Завадський
кафедра математичної інформатики,
Київський національний університет ім. Тараса Шевченка
Київ, Україна
ava@unicyb.kiev.ua, ihorza@gmail.com

# Splittable codes and their applications

Anatoliy Anisimov, Igor Zavadskyi
Department of Mathematical Informatics
National Taras Shevchenko University of Kyiv
Kyiv, Ukraine
ava@unicyb.kiev.ua, ihorza@gmail.com

***Анотація*—Введено поняття подільних стискальних кодів і досліджено їхні загальні властивості. Коди Фібоначчі та мультироздільникові коди класифіковано як частинні випадки подільних кодів. Пояснено переваги мультироздільникових кодів над кодами Фібоначчі через їхню належність до різних класів подільних кодів.**

***Abstract*—Splittable codes are introduced, and their general properties are investigated. Fibonacci codes and multi-delimiter codes are classified as partial cases of splittable codes. Advantages of multi-delimiter codes over Fibonacci codes are explained by means of splittable code classification.**

***Ключові слова*—подільні коди; стискальні коди; коди Фібоначчі; мультироздільникові коди.**

***Keywords*—splittable codes; data compression codes; Fibonacci codes; multi-delimiter codes.**

## I. Вступ

Структура багатьох префіксних кодів передбачає поділ кодового слова на дві частини, що зумовлюється насамперед необхідністю відокремлення кодових слів у потоці закодованих даних. Так, у найпростішому кодуванні довжин серій використовуються пари (*кількість символі у серії*, *символ*). Відомі коди Еліаса [1], Левенштейна [2] та ін. [3] конструюються з пар (*бітова довжина*, *двійкове подання елемента даних*). У кодах Голомба [4] і Голомба-Райса [5] використовуються пари (*частка*, *залишок*). Згідно з наведеним в [6] означенням, код Фібоначчі Fib$m$ визначається як множина таких двійкових кодових слів довжини не менше $m$, що кожне слово містить тільки одне входження рядка з $m$ послідовних одиниць, причому це входження є суфіксом слова. Кодові слова Fib$m$, очевидно, також можна розглядати як конструкції з двох елементів: інформативної частини та суфіксу. Характерною особливістю цих кодів є фіксованість як довжини, так і вмісту суфіксу. Це справедливо і для більш широкого класу кодів, так званих патерн-кодів, або кодів із фіксованим суфіксом, властивості яких досліджено в [7].

В основу ще одного способу префіксного кодування послідовності натуральних чисел покладено їх подання у двобазисній системі числення [8]. Наприклад, у системі числення з основами 2 і 3 число подається як послідовність пар $(k_i,\Delta_i)$, де $k_i$ – довільне додатне ціле число, а $\Delta_i$ набуває значень 0, 1 або 2. Для кодування значень $k_i$ та $\Delta_i$ можна застосовувати різноманітні префіксні кодування, одне з яких детально досліджено в [9]; це так звані (2,3)-коди. Зауважимо, що в цьому випадку кодове слово являє собою не пару, а послідовність пар двійкових кодів з тією властивістю, що довжина коду одного елемента кожної пари обмежена, а іншого – ні. Подібну структуру мають і мультироздільникові коди, детально досліджені в [10], що за деякими характеристиками, зокрема за стискальною ефективністю, перевершують коди Фібоначчі.

Отже, широкий клас стискальних кодів можна розглядати як реалізації одного загального методу кодування:

- Згідно з певним математичним принципом елементу вхідного алфавіту зіставляється послідовність пар $\{(k_1,\Delta_1),\ldots, (k_n,\Delta_n)\}$, де $k_i$ і $\Delta_i$ – невід'ємні числа, $k_i > 0$, причому множина допустимих значень $\Delta_i$ обмежена, а множина допустимих значень $k_i$ – ні.
- Для подання значень $k_i$ та $\Delta_i$ використовуються окремі коди, незалежні від індексу $i$, кодованої послідовності та одне від одного.

- Коди $k_i$ та $\Delta_i$ конкатенуються; якщо спочатку записують код $k_i$, а потім – $\Delta_i$, отримуємо *(k,Δ)-блок*, якщо навпаки – *(Δ,k)-блок*.
- $(k,\Delta)$- чи $(\Delta,k)$-блоки конкатенуються в прямому чи зворотному порядку.

Коди описаної структури ми називатимемо *подільними*. У цій статті ми дамо означення подільних кодів, класифікуємо їх та покажемо, що як коди Фібоначчі, так і мультироздільникові коди є подільними кодами, але належать до різних їх класів. Також ми спробуємо пояснити доведені на практиці переваги мультироздільникових кодів над кодами Фібоначчі з теоретичного погляду, тобто з точки зору їх належності до різних класів подільних кодів.

## II. Подільні коди

Насамперед дамо означення $(\Delta,k)$-коду, призначеного для кодування впорядкованої послідовності пар цілих чисел $(\Delta_1, k_1),\ldots, (\Delta_n, k_n)$.

**Означення 1.** $(\Delta,k)$-код є множиною кодових слів, що задовольняють таким умовам:

- слово являє собою конкатенацію $(\Delta,k)$-блоків.
- кожен $(\Delta,k)$-блок – це конкатенація кодів значення $\Delta$ і значення $k$, де $\Delta$ – невід'ємне ціле число, що не перевищує деякої константи $d$, а $k$ може бути довільним натуральним числом.
- значення $\Delta$ і $k$ кодуються незалежними префіксними кодами;
- $(\Delta,k)$-код є префіксним.

$(k,\Delta)$-код означується так само, як і $(\Delta,k)$-код, за тим винятком, що в кожному блоці коди значень $\Delta$ і $k$ поміняні місцями. Разом $(k,\Delta)$- і $(\Delta,k)$-коди утворюють множину *подільних кодів*.

Коди, у яких параметр $\Delta$ може набувати не більше $d$ різних значень, називатимемо *d-подільними*. Одним із найпростіших $d$-подільних кодів є код Голомба, у якому кожне кодове слово містить тільки одну $(k,\Delta)$-групу, що подає число $km + \Delta$, $0 \le \Delta < d$. Цікавішим прикладом подільних кодів є коди Фібоначчі. Покажемо, що за певного кодування $k$ і $\Delta$ слова коду Fib$m$ можна інтерпретувати як послідовності $(k,\Delta)$-блоків, однак не вдається в загальному випадку подавати як набори $(\Delta,k)$-блоків.

**Теорема 1.** Будь-який код Фібоначчі Fib$m$ є $m$-подільним $(k,\Delta)$-кодом, але не є $(\Delta,k)$-кодом.

**Доведення.** Покажемо, що будь-яке слово будь-якого коду Fib$m$ складається з цілої кількості $(k,\Delta)$-груп бітів, де значення $k\geq 1$ необмежене зверху і кодується послідовністю 0…01, що містить $k$–1 нулів, а величина $\Delta$ може набувати $m$ різних значень, які кодуються рядками 0, 10,…,1…10, що містять до $m$–2 одиниць, а також рядком 1…1 із $m$–1 одиницями.

Справді, якщо слово складається з $m$ послідовних одиниць, то перша з них є кодом $k$, а наступні $m$–1 – кодом $\Delta$. Розглядаючи інші кодові слова, зауважимо, що початок будь-якої двійкової послідовності може бути інтерпретовано і як код деякого значення $\Delta$, і як код деякого значення $k$. Вилучимо з початку кодового слова код $k$, потім – код $\Delta$ і отримаємо коротше кодове слово, до якого застосуємо ту саму процедуру, аж поки не поділимо все слово на $(k,\Delta)$-групи бітів або не отримаємо залишок, у який не вкладається ціла $(k,\Delta)$-група. Подивимось, як згідно з цією процедурою оброблятиметься закінчення кодового слова, що має вигляд 01…1 і містить $m$ одиниць. Перший біт "0" цього закінчення може бути або передостаннім бітом деякого коду $k$, або останнім бітом коду $\Delta$. У першому випадку перша одиниця суфіксу 1…1 належатиме коду $k$, а решта $m$–1 одиниць складатимуть код $\Delta$. У другому випадку на останній ітерації оброблятиметься рядок 1…1 із $m$–1 одиницями, а він, як було зазначено вище, є $(k,\Delta)$-групою.

Тепер покажемо, що коди Фібоначчі не є $(\Delta,k)$-кодами. Тобто не існує таких префіксних кодувань параметрів $k$ і $\Delta$ (значення першого необмежені зверху, а другий може набувати лише скінченну множину значень), за яких усі слова коду Fib$m$ можна подати у вигляді цілої кількості $(\Delta,k)$-груп.

Оскільки кодові слова Fib$m$ можуть починатися з довільної кількості нулів, то для їх подання у вигляді послідовностей $(\Delta,k)$-груп серед можливих кодів $\Delta$ має бути рядок вигляду 0…0. Припустимо, що він містить $s$ нулів, і розглянемо кодове слово вигляду 0…01…1, що містить $s$ нулів і $m$ одиниць. Його нулі відповідають коду $\Delta$, а отже, в частині 1…1 спочатку міститься код $k$, а потім – ціла кількість $(\Delta,k)$-груп. Зауважимо, що послідовність, яка містить лише $m$ одиниць, також є кодовим словом Fib$m$, а значить, жоден код $k$ не може містити $m$ одиниць (інакше це слово не становило б повної $(\Delta,k)$-групи). Припустимо, що згаданий вище код $k$ містить $t$ одиниць, $t<m$, а подальші $m$–$t$ одиниць складаються з цілої кількості $(\Delta,k)$-груп. Останнє означає, що будь-які $m$–$t$ одиниць поспіль мають декодуватися як ціла кількість $(\Delta,k)$-груп. Зокрема, це справедливо і для перших $m$–$t$ одиниць слова, що складається лише з $m$ одиниць, а тому й останні $t$ одиниць цього слова являють собою конкатенацію цілої кількості $(\Delta,k)$-груп. Останній факт означає існування деякого коду $k$, який містить менше $t$ одиниць. Однак $t$ одиниць поспіль також є кодом $k$, звідки робимо висновок про непрефіксність кодування $k$. Ця суперечність завершує доведення теореми. ■

Зауважимо, що якщо подільні коди застосовувати для стискання інформації, то вони будуть тим ефективнішими, чим коротшою буде середня довжина коду послідовностей $(k_1,\Delta_1),\ldots, (k_n,\Delta_n)$. З цієї точки зору кодування параметрів $\Delta$ і $k$

в одинковій системі числення, як, наприклад, у кодах Фібоначчі, не економне. Більш заощадливим є, наприклад, кодування значень $\Delta$ та/або $k$ у скороченій двійковій системі. Проте для параметра $k$ таке кодування неможливе, оскільки множина його значень необмежена, а отже і довжина коду, що відповідатиме кожному значенню $k_i$, невідома. Однак у скороченій двійковій системі можна кодувати значення параметра $\Delta$. Наприклад, у розглянутому в [9] коді $D_1$ три можливих значення $\Delta$ кодуються двійковими рядками 0, 10 та 11; якби $\Delta$ могло набувати лише 2 значення, то їх варто було б кодувати рядками «0» і «1», 4 значення можна закодувати всіма можливими двохбітовими рядками тощо.

Що стосується параметра $k$, то для нього існує всього два префіксних одинкових кодування: вигляду $0\ldots0\pi_1$, де кількість нулів дорівнює $k$–1, а $\pi_1$ – деякий незмінний рядок, що містить одиницю, і вигляду $1\ldots1\pi_0$, де кількість одиниць дорівнює $k$–1, а $\pi_0$ – деякий незмінний рядок, що містить нуль. Якщо ми конструюємо економні коди для $k$, то доцільно вибирати найкоротші рядки $\pi_1$=1 і $\pi_0$=0. Звичайно, теоретично для $k$ можна застосовувати й інші префіксні кодування, наприклад, коди Еліаса. Однак у відомих нам застосуваннях подільних кодів розподіл значень $k$ є геометричним, а для такого розподілу найефективнішим префіксним кодуванням є саме одинкові коди вигляду 0…01 або 1…10.

Описаним принципам цілком відповідають коди Голомба. Як уже зазначалося, це одні з найпростіших ($k$,$\Delta$)-кодів, у яких кожне кодове слово складається з однієї ($k$,$\Delta$)-групи. Якщо ж розглядати складніші коди, слова яких можуть містити кілька ($\Delta$,$k$)- чи ($k$,$\Delta$)-груп, то останні групи в кодовому слові мають виконувати функцію роздільника, тобто відокремлювати одне кодове слово від іншого. Зауважимо, що внаслідок особливостей одинкового кодування параметра $k$ останній біт ($\Delta$,$k$)-блоку завжди має одне те саме значення, на відміну від останнього біта ($k$,$\Delta$)-блоку. Тому подільні коди доцільно конструювати саме з ($\Delta$,$k$)-, а не ($k$,$\Delta$)-блоків, задаючи роздільник у вигляді $0\alpha0$, де $\alpha0$ — розділовий блок, а «0» перед ним — останній символ попереднього блоку чи попереднього кодового слова (якщо для $k$ обрано кодування 1…10). Це дає змогу утворювати короткі кодові слова, які не містять цілого роздільника. Наприклад, вони можуть складатися лише з розділового блоку вигляду $\alpha0$, у той час як роздільник має вигляд $0\alpha0$. Довший роздільник забезпечує кращу асимптотичну щільність коду, у той час як наявність коротких кодових слів дає змогу ефективніше організовувати стикання за невеликих обсягів вхідного алфавіту.

Крім того, будь-яка множина роздільників вигляду 01…10 сама по собі утворює префіксний код, на відміну від роздільників вигляду 1…1 із кодів Фібоначчі. Тому в одному ($\Delta$,$k$)-коді можна використовувати кілька роздільників. Такі коди ми назвемо *мультироздільниковими* і їх дослідженню буде присвячено наступний розділ.

## III. МУЛЬТИРОЗДІЛЬНИКОВІ КОДИ

Припустимо, що $m_1$ або… або $m_t$ — деякі цілі числа, такі що $0 < m_1 <\ldots< m_t$. Дамо означення мультироздільникового коду $D_{m_1,\ldots,m_t}$.

**Означення 2.** Мультироздільниковий код $D_{m_1,\ldots,m_t}$ містить усі слова вигляду 1…10 із $m_1$ або… або $m_t$ одиницями і всі слова, що задовольняють таким умовам:

- слово не починається з послідовності 1…10, що містить $m_1$ або $m_2$ або… або $m_t$ одиниць;
- слово закінчується послідовністю 01…10, що містить $m_1$ або $m_2$ або… або $m_t$ одиниць;
- слово не містить послідовності 01…10 із $m_1$ або $m_2$ або… або $m_t$ одиницями ніде, крім кінця.

Роздільниками у такому коді є послідовності вигляду 01…10, що містять $m_1$ або $m_2$ або… або $m_t$ одиниць, однак код містить також слова вигляду 1…10 із $m_1$ або $m_2$ або… або $m_t$ одиницями, які утворюють роздільник разом із останнім нулем попереднього кодового слова в тексті.

Покажемо, що мультироздільникові коди є ($\Delta$,$k$)-кодами.

**Теорема 2.** Будь-який код $D_{m_1,\ldots,m_t}$ є ($\Delta$,$k$)-кодом.

**Доведення.** Нам потрібно вказати деяке натуральне число, якого не може перевищувати значення параметра $\Delta$, і побудувати такі префіксні коди для $\Delta$ і $k$, що будь-яке слово з $D_{m_1,\ldots,m_t}$ складатиметься з цілого числа ($\Delta$,$k$)-груп.

Нехай $d$ – деяке ціле число, $0 \le d < m_1$, а величина $\Delta$ може набувати $2^d$+1 значень, які кодуватимемо рядком “0” та всіма двійковими рядками довжини $d$+1 із першим символом “1”. Значення параметра $k$, що може бути довільним додатним цілим числом, подаються у вигляді послідовностей 1…10, які містять $k$–1 одиниць.

Якщо слово має вигляд 1…10, то, згідно з означенням 2, його довжина є не меншою за $m_1$+1. Але довжина коду $\Delta$ не перевищує $m_1$ і тому код $\Delta$ не включає в себе всього слова вигляду 1…10, а отже, решта цього слова може бути інтерпретована як код $k$.

Розглядаючи інші кодові слова, зауважимо, що початок будь-якої двійкової послідовності може бути інтерпретовано і як код деякого значення $\Delta$, і як код деякого значення $k$. Будемо застосовувати той самий принцип, що і в доведенні теореми 1: вилучимо з початку кодового слова код $\Delta$, потім – код $k$ і отримаємо коротше кодове слово, до якого застосуємо ту саму процедуру, аж поки не поділимо все слово на ($\Delta$,$k$)-групи бітів або не отримаємо залишок, у який не вкладається ціла ($\Delta$,$k$)-група.

Подивимось, як згідно з цією процедурою оброблятиметься закінчення кодового слова, що має вигляд 01…10 і містить не менше $m_1$ одиниць. Перший біт “0” цього закінчення може бути або кінцем деякого коду $k$, або належати коду Δ. У першому випадку на останній ітерації отримуємо залишок 1…10 із не менш ніж $m_1$ одиницями, що, як показано вище, є повною (Δ,$k$)-групою. У другому випадку зазначимо, що код Δ містить не більш ніж $m_1$ бітів і після його вилучення отримаємо послідовність вигляду 1…10, яка представляє певне значення $k$. Таким чином, ситуація, коли на останній ітерації отримується залишок, у який не вкладається ціла (Δ,$k$)-група, неможлива. ■

Усі мультироздільникові коди є однозначно декодовними, повними та універсальними в розумінні Еліаса, що доведено в [10]. Якщо порівнювати мультироздільникові коди з кодами Фібоначчі, то першим властиві розглянуті в попередньому розділі переваги (Δ,$k$)-кодів над ($k$,Δ)-кодами. Так, у табл. 1 показано кількість коротких кодових слів у мультироздільникових кодах і кодах Фібоначчі, а також наведено їхню асимптотичну щільність. Як уже зазначалося, мультироздільникові (Δ,$k$)-коди можуть містити кодові слова, що коротші за роздільник, на відміну від ($k$,Δ)-кодів Фібоначчі. А отже, і загальна кількість коротких кодових слів у мультироздільникових кодах може буде вищою. Так, наприклад, код $D_{2,4,5}$ містить 34 кодових слова довжини не більше 8, порівняно з 28 словами у коді Fib3, який серед кодів Фібоначчі вважається найкращим для стискання природномовних текстів.

| Код | Асимптотична щільність | Кількість кодових слів довжини ≤ $n$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $n$=2 | $n$=3 | $n$=4 | $n$=5 | $n$=6 | $n$=7 | $n$=8 | $n$=15 |
| Коди з найкоротшим словом довжини 2 | | | | | | | | | |
| Fib2 | $1{,}618^n$ | 1 | 2 | 4 | 7 | 12 | 20 | 33 | 986 |
| $D_1$ | $1{,}755^n$ | 1 | 2 | 3 | 5 | 9 | 16 | 28 | 1432 |
| $D_{1,2}$ | $1{,}618^n$ | 1 | 3 | 5 | 7 | 10 | 16 | 27 | 799 |
| $D_{1,3}$ | $1{,}674^n$ | 1 | 2 | 4 | 7 | 11 | 18 | 30 | 1106 |
| Коди з найкоротшим словом довжини 3 | | | | | | | | | |
| Fib3 | $1{,}839^n$ | 0 | 1 | 2 | 4 | 8 | 15 | 28 | 2031 |
| $D_2$ | $1{,}867^n$ | 0 | 1 | 2 | 4 | 7 | 13 | 24 | 1906 |
| $D_{2,3}$ | $1{,}785^n$ | 0 | 1 | 3 | 6 | 11 | 19 | 33 | 1874 |
| $D_{2,4}$ | $1{,}823^n$ | 0 | 1 | 2 | 5 | 9 | 17 | 30 | 1998 |
| $D_{2,3,4}$ | $1{,}731^n$ | 0 | 1 | 3 | 7 | 13 | 23 | 39 | 1721 |
| $D_{2,4,5}$ | $1{,}796^n$ | 0 | 1 | 2 | 5 | 10 | 19 | 34 | 2019 |
| $D_{2,4,6}$ | $1{,}809^n$ | 0 | 1 | 2 | 5 | 9 | 18 | 32 | 2032 |

Табл. 1. Кількість коротких кодових слів та асимптотична щільність кодів

| Текст | Слів у тексті | Розмір словника | Ентропійні біти | SCDC | Fib3 | $D_{2,3,5}$ | $D_{2,3}$ | $D_{2,4,5}$ |
|---|---|---|---|---|---|---|---|---|
| Гамлет, Шекспір | 30 694 | 4 501 | 9,2112 | 10,47 +13,7% | 10,01 +8,7% | 9,76 +6,0% | 9,83 +6,7% | 9,85 +6,9% |
| Робінзон Крузо, Д. Дефо | 121 325 | 5 994 | 8,73519 | 10,13 +16,0% | 9,41 +7,7% | 9,13 +4,5% | 9,13 +4,5% | 9,21 +5,4% |
| Біблія | 779 079 | 12 452 | 8,6279 | 10,138 +17,5% | 9,219 +6,9% | 8,954 +3,8% | 9,044 +4,8% | 9,071 +5,1% |
| Статті з Вікіпедії | 19 507 783 | 288 179 | 11,0783 | 12,869 +16,2% | 11,564 +4,4% | 11,492 +3,7% | 11,488 +3,7% | 11,471 +3,5% |

Табл. 2. Стискальна ефективність кількох кодів (середня довжина слова в бітах і % перевищення над ентропією)

Загалом можна відзначити дві закономірності:

- коди із більшою кількістю роздільників мають гіршу асимптотичну щільність, однак містять більше коротких слів;
- чим коротшою є довжина роздільника, тим більше коротких слів містить код, але тим гіршу асимптотичну щільність він має.

Варіюючи довжини роздільників та їхню кількість, можна гнучко підбирати мультироздільникові коди відповідно до розподілу частот елементів вхідного алфавіту, щоб максимізувати коефіцієнт стискання. У застосуванні до стискання природномовних текстів найбільш ефективними видаються коди $D_{2,3,5}$ та $D_{2,4,5}$. Як свідчать результати експериментів, наведені в табл. 2, розмір стиснутих за допомогою цих кодів текстів лише на 3,5–7% перевищує ентропійну межу, що 1,2–1,6 разів краще за аналогічний показник кодів Фібоначчі і більш ніж удвічі краще за SCDC-коди [11].

## IV. Висновки

Подільні префіксні коди – це широке сімейство кодів, застосовних насамперед до стискання інформації. Серед них можна виділити 2 класи кодів: ($k$,Δ)-коди (найвідоміший представник – коди Фібоначчі) та (Δ,$k$)-коди (наприклад, мультироздільникові коди). У силу особливостей будови (Δ,$k$)-коди видаються більш ефективними для стискання певних типів даних, зокрема природномовних текстів, що підтверджується результатами експериментів.

## Література References

[1] P. Elias, "Universal codeword sets and representation of the integers", IEEE Trans. Inf. Theory, vol. 21, no. 2, pp. 194–203, 1975.

[2] V.I. Levenshtein, "Redundancy and delay recovery of coding of natural numbers," (in Russian) Probl. Cybern., vol. 20, pp. 173-179, 1968.

[3] D. Salomon, Variable-Length Codes for Data Compression London, U.K.: Springer-Verlag, 2007.

[4] S.W. Golomb, "Run-length encodings," IEEE Trans. Inform Theory, IT-12(3), pp. 399–401, 1966.

[5] R.F. Rice and R. Plaunt, "Adaptive variable-length coding for efficient compression of spacecraft television data," IEEE Trans. Communications, vol. 16(9), pp. 889–897, Dec. 1971.

[6] S.T. Klein and M.K. Ben-Nissan, "On the usefulness of Fibonacci compression codes," Computer J., vol. 53, no. 6, pp. 701–716, Jul. 2010.

[7] K.B. Lakshmanan, "On universal codeword sets," IEEE Trans. Inform. Theory, vol. 27, no. 22, pp. 194–203, 1975.

[8] A. Anisimov, "Prefix encoding by means of the 2,3-representation of numbers," IEEE Trans. Inf. Theory, vol. 59, no. 4, pp. 2359–2374, 2013.

[9] A. Anisimov, I. Zavadskyi, "Robust prefix encoding using lower (2; 3) number representation," (in Russian) Cybernetics and System Analysis, no. 2, pp. 3–14, 2014.

[10] A.V. Anisimov, I.O. Zavadskyi. Variable-Length Prefix Codes With Multiple Delimiters // IEEE Transactions on Information Theory, vol. 63, issue 5, pp. 2885-2895. –2017.

[11] N. Brisaboa, A. Farina, G. Navarro and J. Parama. "Lightweight natural language text compression," Information Retrieval, vol. 10, no. 1,pp. 1–33, 2007.