Applied Mathematical Sciences, Vol. 3, 2009, no. 17, 813 - 824

The Cardinality of a Sphere Relative to an Edit Distance

Othman Echi

Department of Mathematics, University Tunis-El Manar
Faculty of Sciences of Tunis
"Campus Universitaire" 2092 Tunis, Tunisia
othechi@yahoo.com

Abstract. Let Σ be an *alphabet* (a finite set). We denote by Σ^* the set consisting of all finite words (or strings) that can be made from the letters (or phonemes). The set of all n-letter words over Σ will be denoted by Σ_n. Let w be an n-letter word in Σ_n. This paper deals with the cardinality of the sphere $S_{\mathrm{H}}(w,p) := \{u \in \Sigma_n \mid \mathrm{H}(w,u) = p\}$ of center w and radius p ($p \in \mathbb{N}^*$) relatively to the Hamming distance H on Σ_n. A new distance T is defined on the language Σ^* and the cardinality of the corresponding sphere $S_{\mathrm{T}}(w,p) := \{u \in \Sigma_n \mid \mathrm{T}(w,u) = p\}$ is also computed.

These cardinalities are showed to satisfy some curious recurrence relations. These recurrence relations incite us to introduce new types of binomial coefficients and binomial formula.

Mathematics Subject Classification: 05A05; 05A10; 68R15

Keywords: Alphabet; Binomial coefficient; Binomial theorem; Cardinality; Edit distance; Hamming distance; String

Introduction

Let Σ be a finite set (usually called an *alphabet*). We denote by Σ^* the set consisting of all finite words (or strings) that can be made from the letters (or phonemes, following the linguistic terminology). A word does not have to be meaningful in any language.

A *word* over an alphabet Σ is an element $w = (a_1, \ldots, a_n)$ of the cartesian product Σ^n, the number n is called the length of w (we denote by $l(w) := n$); a_i is called the i^{th} letter of the word. We write simply $w = a_1 a_2 \ldots a_n$.

It is also convenient to include in Σ^* an empty word denoted by ε, and defined to be the word with no letters (we say that ε is of length 0, $l(\varepsilon) = 0$).

Two words $w, w^{'}$ are said to be equal if they have the same length, and for each i, the i^{th} letter of w is the i^{th} letter of $w^{'}$.

The set of all n-letter words over Σ will be denoted by Σ_n. When we use the alphabet $\Sigma = \{0, 1\}$, then Σ^* is the set of all possible binary words. Binary words are extremely important in storing data into computers.

Let Σ be an alphabet; then any subset of Σ^* is called a *language* over Σ. For example the set of all English (resp. French) words over the alphabet $\Sigma = \{a, b, c, \dots, x, y, z\}$ is a language over Σ called the English (resp. French) language.

The language Σ^* will be called the *improper language over* Σ.

0.1. **Hamming distance.** Richard Hamming is one of the founders of modern coding theory [5]. His research has included work in the areas of coding theory, numerical methods, statistics and digital filtering. Hamming defined the distance between the bit strings $x = x_1x_2 \dots x_n$ and $y = y_1y_2 \dots y_n$ as the number of positions in which these strings differ; we denote by $\mathrm{H}(x, y)$ this distance (thus the hamming distance between two strings equals the number of changes in individual letters needed to change one of the strings into the other).

Many codes (for instance, ASCII code) have the property that each character is given a code of the same length. Such a character code is called a *fixed length character code*. For codes with fixed length character, there is an important type of decoding, namely the *nearest neighbor decoding*. This type of decoding is based on Hamming distance. It is also known that nearest neighbor decoding gives us the most likely codeword sent, so that it is also maximum likelihood decoding.

The Hamming distance is popular in the Knowledge Representation community; however, the assumption that it measures only the distance between two strings with the same length is extremely compelling and gives the Hamming distance very little flexibility.

0.2. **Edit Distance.** The *Edit distance*(or the *Levenshtein distance*) between two strings $x = x_1 \ \dots \ x_n$ and $y = y_1 \ \dots \ y_m$ is the minimum number of "errors" (edit operations) needed to transform x into y, where possible operations are:

- Insert a character:
  $\mathrm{insert}(x, i, a) = x_1x_2 \ \dots \ , \ x_iax_{i+1} \ \dots \ x_n$.
- Delete a character:
  $\mathrm{delete}(x, i) = x_1x_2 \ \dots \ x_{i-1}x_{i+1} \ \dots \ x_n$.
- Modify a character:
  $\mathrm{modify}(x, i, a) = x_1x_2 \ \dots \ x_{i-1}ax_{i+1} \ \dots \ x_n$.

The Edit distance between the strings x and y will be denoted by $\mathrm{Ed}(x, y)$.

When you run a spell checker on a text, and it finds a word not in the dictionary, it normally proposes a choice of possible corrections.

If it finds "stell" it might suggest steal, steel, stele, sell, shell, spell, stall, still, stull and swell. As part of the heuristic used to propose alternatives, words that are "close" to the misspelled word are proposed. The suggestions or propositions are based on the Edit distance.

Edit distance is extensively used in computational biology to explore which DNA sequences are likely mutations of one another (see Waterman 1989 [12], Waterman 1995 [13], Crochemore and Gusfield 1994 [2], Farach- Colton 1998 [3]).

Dialectology, is the study of how language varies geographically. It is one of the oldest branches of linguistics; the early works date from the 19th century (see Petyt (1980) [7]).

Levenshtein distance is also employed in Linguistics and especially in Dialectology (see for instance the application of Levenshtein distance to speech recognition Veldhuijzen van Zanten et al. 1999 [10]). Other researchers have used variants of Levenshtein distance to diagnose potentially pathological pronunciation deviation (see Connolly 1997 [1]).

0.3. **A new distance.** We introduce, here, a new distance defined on the language Σ^* over an alphabet Σ.

Let us make some notations:

- If $w = w_1 \dots w_n \in \Sigma_n$, then for $i \leq n$, we denote by $[w]_i$ the word $w_1 \dots w_i$.
- If $u, v \in \Sigma^*$, then we denote by $\mathrm{H}(\underline{u}, \underline{v})$ the Hamming distance between $[u]_i$ and $[v]_i$, where $i = \min\{l(u), l(v)\}$

Constructing a distance is always interesting; as said "When you can measure, [. . .] you know something" -Lord Kelvin-

Definition 0.1. Let Σ be an alphabet. We define the mapping

$$\mathrm{T} : \Sigma^* \times \Sigma^* \longrightarrow \mathbb{N}$$

which takes the pair (u, v) to $\mathrm{T}(u, v)$, where

$$\mathrm{T}(u, v) = \mathrm{H}(\underline{u}, \underline{v}) + |l(u) - l(v)|.$$

with $\mathbb{N} = \{0, 1, 2, 3 \dots\}$.

We will prove in the next section that T is a distance (a metric) on Σ^*.

This distance may also be useful in dialectology and pathological pronunciation deviation.

In this paper, we compute the cardinalities of the spheres $S_H(w, p)$ and $S_{\mathrm{T}}(w, p)$ of center w and radius p ($p \in \mathbb{N}^*$) relatively to the metrics H and T respectively. These cardinalities are showed to satisfy some curious recurrence relations. These recurrence relations incite us to introduce new types of binomial coefficients and binomial formula.

1. The metric T

The following notion is interesting since it links integer valued metrics on Σ^* and the Hamming distance.

Definition 1.1. Let Σ be an alphabet. A metric $\delta : \Sigma^* \times \Sigma^* \longrightarrow \mathbb{N}$ is said to be *compatible with the Hamming distance*, if for any two strings with the same length u, v in Σ^*, we have $\delta(u, v) = \mathrm{H}(u, v)$.

Example 1.2. *The Edit distance is not compatible with the Hamming distance. Indeed, if we let* $\Sigma := \{0, 1, 2\}$ *and* $u = 012$, $v = 120$, *then* $\mathrm{H}(u, v) = 3$. *Let us make the following edit operations:*
– $\mathrm{insert}(u, 3, 0) = 0120 := u_1$,
– $\mathrm{delete}(u_1, 1) = 120 = v$.
Thus, $\mathrm{Ed}(u, v) = 2$.

Firstly, it is worth noting that "T is a metric" do not follows immediately from the fact that "H is a metric". Indeed, the expression $\mathrm{H}(\underline{u}, \underline{v})$ does not satisfy the triangle inequality nor the axiom of separation. For the triangle inequality, take for example $\Sigma = \{0, 1\}$ and $u = 001$, $v = 0001$ and $w = 00$; then we have

$$1 = \mathrm{H}(\underline{u}, \underline{v}) \leq \mathrm{H}(\underline{u}, \underline{w}) + \mathrm{H}(\underline{w}, \underline{v}) = 0 + 0.$$

Proposition 1.3. *Let* Σ *be an alphabet. Then the following properties hold:*
(i) $\mathrm{T}(u, v) \geq 0$, *for all* $u, v \in \Sigma^*$.
(ii) $\mathrm{T}(u, v) = 0$ *if and only if* $u = v$.
(iii) $\mathrm{T}(u, v) = \mathrm{T}(v, u)$, *for all* $u, v \in \Sigma^*$.
(iv) $\mathrm{T}(u, v) \leq \mathrm{T}(u, w) + \mathrm{T}(w, v)$, *for all* $u, v, w \in \Sigma^*$.
That is to say T *is a metric on the improper language* Σ^*.

Proof. Only (iv) needs to be shown.

It will be continent to denote $m_{u,v}$ and $M_{u,v}$ respectively the numbers $\min\{l(u), l(v)\}$ and $\max\{l(u), l(v)\}$, for $u, v \in \Sigma^*$. Since T is symmetric (statement (iii)), we may suppose without loss of generality, that $l(u) \leq l(v)$.

First, remark that $\mathrm{T}(u, v) = |A|$, where

$$A = \{i \in \mathbb{N}^* \mid 1 \leq i \leq m_{u,v} \text{ and } u_i \neq v_i\} \cup \{i \in \mathbb{N}^* \mid m_{u,v} < i \leq M_{u,v}\}.$$

Consider the following sets

$$B = \{i \in \mathbb{N}^* \mid 1 \leq i \leq m_{u,w} \text{ and } u_i \neq w_i\} \cup \{i \in \mathbb{N}^* \mid m_{u,w} < i \leq M_{u,v}\}$$

and

$$C = \{i \in \mathbb{N}^* \mid 1 \leq i \leq m_{v,w} \text{ and } v_i \neq w_i\} \cup \{i \in \mathbb{N}^* \mid m_{v,w} < i \leq M_{v,w}\}.$$

By discussing three cases, namely $l(w) \leq l(u)$, $l(u) < l(w) \leq l(v)$ and $l(v) < l(w)$, one may prove easily that $A \subseteq B \cup C$; and thus $|A| \leq |B| + |C|$, which gives immediately $\mathrm{T}(u, v) \leq \mathrm{T}(u, w) + \mathrm{T}(w, v)$, proving that T is a metric on the improper language Σ^*. □

Remark 1.4. It is easily seen that T is compatible with the Hamming distance and that for each $u, v \in \Sigma^*$, we have $\mathrm{Ed}(u,v) \leq \mathrm{T}(u,v)$ (we may have $\mathrm{Ed}(u,v) < \mathrm{T}(u,v)$; take for instance $u = ab$ and $v = b$; then $\mathrm{Ed}(u,v) = 1$ and $\mathrm{T}(u,v) = 2$).

Question 1.5. Let Σ be an alphabet and δ an integer valued metric on Σ^* which is compatible with the Hamming distance. Is $\mathrm{T} \leq \delta$?

2. Spheres in the Improper Language

Notations 2.1. Let Σ be an alphabet of size k. Let w be an n-letter word of Σ^*. Then we denote by;

- $S_H(\Sigma^*, w, p) = \{v \in \Sigma_n \mid \mathrm{H}(w,v) = p\}$, where $p \in \mathbb{N}$;
- $S_{\mathrm{T}}(\Sigma^*, w, p) = \{v \in \Sigma^* \mid \mathrm{T}(w,v) = p\}$;
- $S_{\mathrm{Ed}}(\Sigma^*, w, p) = \{v \in \Sigma^* \mid \mathrm{Ed}(w,v) = p\}$;
- The cardinality of $S_H(\Sigma^*, w, p)$ (resp. $S_{\mathrm{T}}(\Sigma^*, w, p)$, resp. $S_{\mathrm{Ed}}(\Sigma^*, w, p)$) will be denoted by $\mathrm{H}(n,k,p)$ (resp. $\mathrm{T}(n,k,p)$, resp. $\mathrm{Ed}(n,k,p)$).

The following result compute the cardinality $\mathrm{H}(n,k,p)$ and $\mathrm{T}(n,k,p)$. Let us first fix some notations:

For $n, p \in \mathbb{N}$, we set $\binom{n}{p} := \dfrac{n!}{p!(n-p)!}$ if $p \leq n$ and $\binom{n}{p} = 0$, if $p > n$.

Theorem 2.2. *For each $n, k, p \in \mathbb{N}$, we have:*

(1) $\mathrm{H}(n,k,p) = (k-1)^p \binom{n}{p}$.

(2)

$$\mathrm{T}(n,k,p) = \sum_{i=0}^{n} \mathrm{H}(n,k,i) k^{p-i} = k^{p-n}(2k-1)^n;$$

if $p > n$; and

$$\mathrm{T}(n,k,p) = \sum_{i=0}^{p} \mathrm{H}(n,k,i) k^{p-i} + \sum_{i=n-p}^{n} \mathrm{H}(i,k,i-(n-p))$$

$$= \sum_{i=0}^{p} (k-1)^i k^{p-i} \binom{n}{i} + \sum_{i=0}^{p} (k-1)^i \binom{i+n-p}{i}$$

if $p \leq n$.

Proof. (1) Let I be the set of subsets E of $\{1, \ldots, n\}$ with cardinality p; for such a subset, we may write in a unique way $E = \{e_1, e_2, \ldots, e_p\}$, with $e_1 < e_2 < \ldots < e_p$.

Let $w = w_1 \ldots w_n$ be an n-letter word of Σ^*, we denote by $S_{\mathrm{H}}(w,k,p)$ the sphere of center w and radius p of the improper language Σ^* relatively to the Hamming distance. If $p > n$, then $S_{\mathrm{H}}(w,k,p)$ is empty.

Now, suppose that $p \leq n$. Consider the map

$$\Phi : \cup[\{E\} \times (\Sigma \setminus \{w_{e_1}\}) \times \ldots \times (\Sigma \setminus \{w_{e_p}\}) : E \in I] \longrightarrow S_{\mathrm{H}}(w,k,p)$$

defined by $\Phi(\varphi, \lambda_1, \dots, \lambda_p)$ is the word $l_1 l_2 \dots l_n$, where $l_i = w_i$ whenever $i \in \{1, \dots, n\} \backslash E$ and $l_i = \lambda_i$, for $i \in E$.

Clearly, $\mathrm{H}(w, \Phi(E, \lambda_1, \dots, \lambda_p)) = p$; this is to say $\Phi(E, \lambda_1, \dots, \lambda_p) \in S_H(w, k, p)$. On the other hand, it is easily seen that Φ is a bijective map.

Thus,

$$\mathrm{H}(n, k, p) = \sum_{E \in I} (k-1)^p = |I|(k-1)^p = (k-1)^p \binom{n}{p}.$$

(2) Suppose that $p > n$. Then any word of $S_{\mathrm{T}}(w, k, p)$ is divided into two words $u \in S_H(w, k, i)$ and $v \in \Sigma_{p-i}$. More precisely the following mapping is clearly a bijection.

$$\begin{array}{rccc} \gamma : & \bigcup_{i=0}^{n} (S_H(w, k, i) \times \Sigma_{p-i}) & \longrightarrow & S_{\mathrm{T}}(w, k, p) \\ & (u, v) & \longmapsto & uv \end{array}$$

where uv is the concatenation of the two words u, v.

Thus,

$$\mathrm{T}(n, k, p) = \sum_{i=0}^{n} k^{p-i} \mathrm{H}(n, k, i) = \sum_{i=0}^{n} (k-1)^i k^{p-i} \binom{n}{i}$$

$$= k^{p-n} (\sum_{i=0}^{n} (k-1)^i k^{n-i} \binom{n}{i}) = k^{p-n} (2k-1)^n.$$

Now, suppose that $p \leq n$

Let us write $S_{\mathrm{T}}(w, k, p) = (S_{\mathrm{T}}(w, k, p) \cap \Gamma_n) \cup (S_{\mathrm{T}}(w, k, p) \cap \Sigma_{>n})$, where $\Gamma_n = \Sigma_0 \cup \dots \cup \Sigma_n$ and $\Sigma_{>n} = \cup[\Sigma_i : i \geq n+1]$

Thus $\mathrm{T}(n, k, p) = |S_{\mathrm{T}}(w, k, p) \cap \Gamma_n| + |S_{\mathrm{T}}(w, k, p) \cap \Gamma_{>n}|$

By using a similar argument as in (1), we see easily that

$$| \, S_{\mathrm{T}}(w, k, p) \cap \Gamma_{>n} \, | = \sum_{i=0}^{p} \mathrm{H}(n, k, i) k^{p-i}.$$

For the other cardinality, it suffices to remark that $S_{\mathrm{T}}(w, k, p) \cap \Gamma_n$ is equal to the disjoint union

$$\bigcup_{i=n-p}^{n} S_H([w]_i, k, i - (n-p))$$

where $[w]_i = w_1 \dots w_i$. Hence

$$|S_{\mathrm{T}}(w, k, p) \cap \Gamma_n| = \sum_{i=n-p}^{n} \mathrm{H}(i, k, i - (n-p))$$

$$= \sum_{i=n-p}^{n} (k-1)^{i-(n-p)} \binom{i}{n-p}$$
$$= \sum_{i=0}^{p} (k-1)^{i} \binom{i+(n-p)}{i}$$

Therefore,

$$\mathrm{T}(n,k,p) = \sum_{i=0}^{p} (k-1)^{i} k^{p-i} \binom{n}{i} + \sum_{i=0}^{p} (k-1)^{i} \binom{i+n-p}{i}.$$

□

A well known recurrence relation(Pascal's triangle) of the double sequence $(\binom{n}{p},\ n,p \in \mathbb{N})$ is the following:

$$\binom{n}{p} = \binom{n-1}{p} + \binom{n-1}{p-1}$$

for $n,p \geq 1$. This yields immediately the following recurrence relation for the cardinality $\mathrm{H}(n,k,p)$:

$$\mathrm{H}(n,k,p) = \mathrm{H}(n-1,k,p) + (k-1)\,\mathrm{H}(n-1,k,p-1)$$

for $n,p \geq 1$ and $k \in \mathbb{N}$.

We begin by solving the above recurrence relation under some reasonable initial conditions.

Proposition 2.3. *Let $(R,+,\times)$ be a unitary commutative ring and*

$$U : \mathbb{N} \times (\mathbb{N} \setminus \{0\}) \times \mathbb{N} \longrightarrow R$$

be a triple sequence of elements of R satisfying the following properties:

(1) *If $p > n$, then $U(n,k,p) = 0$, for each $k \in \mathbb{N} \setminus \{0\}$.*
(2) *There exists $\lambda \in R$ such that $U(n,k,0) = \lambda$, for each $(n,k) \in \mathbb{N} \times (\mathbb{N} \setminus \{0\})$.*
(3) *U satisfies the following recurrence relation:*

$$U(n,k,p) = U(n-1,k,p) + (k-1)U(n-1,k,p-1)$$

for $n,p \geq 1$ and $k \in \mathbb{N} \setminus \{0\}$.

Then

$$U(n,k,p) = \lambda\,\mathrm{H}(n,k,p) = \lambda(k-1)^{p}\binom{n}{p},$$

for $k \neq 1$ or $p \neq 0$.

Proof. We use induction(on the integer p). Let $\mathcal{P}(p)$ be the following proposition that involves p;

$\mathcal{P}(p)$: " for each integer $0 \leq i \leq p$ and each integers n,k such that $k > 1$, we have $U(n,k,i) = \lambda(k-1)^{i}\binom{n}{i}$ ".

Since $U(n,k,0) = \lambda$, $\mathcal{P}(0)$ is true. Suppose $\mathcal{P}(p)$ is true and compute $U(n,k,p+1)$.

By the recurrence relation (3), we have

$$U(n,k,p+1) = U(n-1,k,p+1) + (k-1)U(n-1,k,p).$$

But, induction hypothesis, gives

$$U(n-1,k,p) = \lambda(k-1)^p \binom{n-1}{p}.$$

Thus, by a trivial iteration, we get

$$U(n,k,p+1) = \lambda(k-1)^{p+1} \sum_{i=p}^{n-1} \binom{i}{p}.$$

On the other hand, we have $\sum_{i=p}^{n-1} \binom{i}{p} = \binom{n}{p+1}$. Therefore,

$$U(n,k,p+1) = \lambda(k-1)^{p+1} \binom{n}{p+1},$$

finishing the induction. □

Corollary 2.4. *The cardinality* $\mathrm{H}(n,k,p)$ *is characterized by the following properties:*

(*i*) $\mathrm{H}(n,k,0) = 1$, *for* $k > 1$.
(*ii*) *If* $p > n$, *then* $\mathrm{H}(n,k,p) = 0$ *for each* $k > 1$.
(*iii*) $\mathrm{H}(n,k,p) = \mathrm{H}(n-1,k,p) + (k-1)\,\mathrm{H}(n-1,k,p-1)$, *for* $n,p \geq 1$ *and* $k > 1$.

Question 2.5. Compute the cardinality $\mathrm{Ed}(n,k,p)$.

3. Binomial Coefficients and Binomial Formula

The classical binomial formula (or Newton formula) in an arbitrary ring with 1 where the elements x, y satisfy the commutation relation $xy = yx$ is

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} y^k x^{n-k},$$

where $\binom{n}{p} := \dfrac{n!}{p!(n-p)!}$, for $n \geq p$ and $\binom{n}{p} = 0$, otherwise.

An important generalization first given by Schützenberger [9] is:

$$(x+y)^n = \sum_{k=0}^{n} \frac{(q;q)_n}{(q;q)_k (q;q)_{n-k}} y^k x^{n-k},$$

where x, y satisfy the commutation relation $xy = qyx$, q a scalar, and $(a;q)_k := (1-a)\ldots(1-aq^{k-1})$ if $k = 1, 2, \ldots$, and $(a;q)_0 = 1$.

Motivated by some applications in quantum group theory and non-commutative geometry, several authors have considered more general commutation relations. For instance, Rosengren considered the relation $xy = ax^2 + qyx + by^2$ in [8] and found a nice binomial formula.

Over the years, several generalization of classical binomial coefficients and binomial formula have been done.

Our goal in this paper is the generalization of Newton formula in a more general setting (see Definition 3.2).

Motivated by the recurrence relation in Corollary 2.4(*iii*), we introduce a new type of binomial coefficients.

Definition 3.1. Let $(R, +, \times)$ be a commutative unitary ring and $f : \mathbb{N} \times \mathbb{N} \longrightarrow R$ be a double sequence of elements of R. We say that $(f(n,p))$ are $\mathcal{T}$-binomial coefficients if the following properties hold:

(*i*) $f(n,0) = 1$, for each $n \in \mathbb{N}$;
(*ii*) $f(n,p) = 0$, for each $p > n$;
(*iii*) there exist $\alpha, \beta \in R$ such that

$$f(n,p) = f(n-1,p) + (\alpha + p\beta)f(n-1,p-1),$$

for each $n, p \geq 1$.

We denote by $\mathcal{T}^p_{(\alpha,\beta)n} := f(n,p)$ and if there is no ambiguity, this will be denoted by $\mathcal{T}^p_n$.

Let $n, p \in \mathbb{N}$.

Set $A^p_n := \dfrac{n}{(n-p)}$, for $n \geq p$ and $A^p_n = 0$, otherwise. Then we have the well known recurrence relations:

$$A^p_n = A^p_{n-1} + pA^{p-1}_{n-1}$$

for $n, p \geq 1$; and "Pascal's formula":

$$\binom{n}{p} = \binom{n-1}{p} + \binom{n-1}{p-1}$$

for $n, p \geq 1$.

It is easily seen that the double sequences (A^p_n) and $(\binom{n}{p})$ are $\mathcal{T}$-binomial coefficients.

The following definition is a generalization of the well known binomial formula in a more general setting.

Definition 3.2. Let R be a ring and $(E, +, ., \times)$ be an R-algebra with unit element 1. Let $(F_n, n \in \mathbb{N})$ be a sequence of elements of E. We say that $(F_n, n \in \mathbb{N})$ is a *Newton formula* if there exists a binary operation $\star : E \times E \longrightarrow E$ such that the following properties hold:

(1) $(F_n, n \in \mathbb{N})$ is *iterated under* $\star$ (i.e.; $F_{n+1} = F_1 \star F_n$ for each $n \in \mathbb{N}$).
(2) $f \star 1 = f$, for each $f \in E$.

(3) $f \star (g+h) = f \star g + f \star h$, for each $f, g, h \in S$.
(4) $f \star (\lambda g) = \lambda (f \star g)$, for each $f, g \in S$ and each $\lambda \in R$.

[Such operation $\star$ will be called *a compatible binary operation with the structure of R-algebra* on E".]

We will, simply, write $F_n = (F_1)^{\star n}$.

In connection with $\mathcal{T}$-binomial coefficients and Newton binomial formula, we introduce the following

Definition 3.3. Let R be a commutative ring and x, y be two indeterminates over R. We call *$\mathcal{T}$-binomial formula* each sequence $(F_n(x,y), n \in \mathbb{N})$ in $R[x,y]$, such that

$$F_n(x,y) = \sum_{i=0}^{n} \mathcal{T}_n^i x^i y^{n-i},$$

where $(\mathcal{T}_n^i)$ are $\mathcal{T}$-binomial coefficients.

The main goal of this section is to compute $\mathcal{T}$-binomial coefficients and show that each $\mathcal{T}$-binomial formula is a Newton formula in the sense of Definition 3.2.

Proposition 3.4. *Let $(R, +, \times)$ be a commutative unitary ring and $(\mathcal{T}_n^p, n, p \in \mathbb{N})$ be a double sequence of elements of R such that $\mathcal{T}_n^0 = 1$, for each $n \in \mathbb{N}$ and $\mathcal{T}_n^p = 0$, for each $p > n$. Then the following statements are equivalent:*

(i) $(\mathcal{T}_n^p)$ are $\mathcal{T}$-binomial coefficients;

(ii) there exist α, β in R such that $\mathcal{T}_n^p = (\prod_{i=1}^{p}(\alpha + i\beta))\binom{n}{p}$, for each $n \in \mathbb{N}$ and $p \in \mathbb{N} \setminus \{0\}$.

Proof. $(ii) \Longrightarrow (i)$. Straightforward.

$(i) \Longrightarrow (ii)$. We use induction on $k \in \mathbb{N} \setminus \{0\}$.

– If $k = 1$, then $\mathcal{T}_n^1$ satisfies the recurrence relation:

$$\mathcal{T}_n^1 = \mathcal{T}_{n-1}^1 + (\alpha+\beta)\mathcal{T}_{n-1}^0 = \mathcal{T}_{n-1}^1 + (\alpha+\beta).$$

Thus, clearly, $\mathcal{T}_n^1 = n(\alpha+\beta) = (\alpha+\beta)\binom{n}{1}$.

– Suppose that $\mathcal{T}_n^l = (\prod_{i=1}^{l}(\alpha+i\beta))\binom{n}{l}$, for $1 \leq l \leq k$; and let us compute $\mathcal{T}_n^{k+1}$.

By induction hypothesis, we have

$$\mathcal{T}_n^{k+1} = \mathcal{T}_{n-1}^{k+1} + (\prod_{i=1}^{k+1}(\alpha+i\beta))\binom{n-1}{k},$$

which gives immediately

$$\mathcal{T}_n^{k+1} = \sum_{j=k}^{n-1}(\prod_{i=1}^{k+1}(\alpha+i\beta))\binom{j}{k} = (\prod_{i=1}^{k+1}(\alpha+i\beta))\binom{n}{k+1},$$

finishing the induction. □

It will be interesting to give analogous properties of classical binomial coefficients for the new type introduced here. This is not our purpose in the present paper.

Theorem 3.5. *Let R be a commutative ring and x, y be two indeterminates over R. Let $\mathcal{T}_n^p$ be $\mathcal{T}$-binomial coefficients and $(F_n(x,y), n \in \mathbb{N}\setminus\{0\})$ be a $\mathcal{T}$-binomial formula; with*

$$F_n(x,y) = \sum_{i=0}^{n} \mathcal{T}_n^i x^i y^{n-i}.$$

We define the operation $\star : R[x,y]\times R[x,y] \longrightarrow R[x,y]$, by $f\star g = fg+\beta x^2\dfrac{\partial g}{\partial x}$.

Then, for each $n \in \mathbb{N}\setminus\{0\}$, we have $F_n = (F_1)^{\star n}$; accordingly each $\mathcal{T}$-binomial formula is a Newton formula in the R-algebra $(R[x,y],+,.,\times)$.

Proof. Let us write $F_n(x,y) = \sum\limits_{i=0}^{n} \mathcal{T}_n^i x^i y^{n-i}$, where $(\mathcal{T}_n^i)$ are $\mathcal{T}$-binomial coefficients. There exist $\alpha, \beta \in R$ such that

$$\mathcal{T}_n^k = \mathcal{T}_{n-1}^k + (\alpha + k\beta)\mathcal{T}_{n-1}^{k-1},$$

for each $n, k \geq 1$. Then, for $n \in \mathbb{N}^*$, we have

$$F_n(x,y) = \sum_{k=0}^{n} \mathcal{T}_n^k x^k y^{n-k} = y^n + \sum_{k=1}^{n} \mathcal{T}_{n-1}^k x^k y^{n-k} + \sum_{k=1}^{n} (\alpha + k\beta)\mathcal{T}_{n-1}^{k-1} x^k y^{n-k}$$

$$= y^n + y(\sum_{k=1}^{n-1} \mathcal{T}_{n-1}^k x^k y^{(n-1)-k}) + \alpha \sum_{k=1}^{n} \mathcal{T}_{n-1}^{k-1} x^k y^{n-k} + \beta \sum_{k=1}^{n} k\mathcal{T}_{n-1}^{k-1} x^k y^{n-k}$$

$$= y^n + y(F_{n-1}(x,y) - y^{n-1}) + \alpha \sum_{k=0}^{n-1} \mathcal{T}_{n-1}^k x^{k+1} y^{(n-1)-k} + \beta \sum_{k=0}^{n-1} (k+1)\mathcal{T}_{n-1}^k x^{k+1} y^{(n-1)-k}$$

$$= yF_{n-1}(x,y) + \alpha x F_{n-1}(x,y) + \beta x F_{n-1}(x,y) + \beta x^2 \frac{\partial F_{n-1}}{\partial x}(x,y)$$

$$= (y + (\alpha+\beta)x)F_{n-1}(x,y) + \beta x^2 \frac{\partial F_{n-1}}{\partial x}(x,y)$$

$$= F_1(x,y).F_{n-1}(x,y) + \beta x^2 \frac{\partial F_{n-1}}{\partial x}(x,y).$$

If we define $\star : R[x,y] \times R[x,y] \longrightarrow R[x,y]$, by $f \star g = fg + \beta x^2 \dfrac{\partial g}{\partial x}$, then clearly, $\star$ is a binary operation which is compatible with the structure of R-algebra on $(R[x,y],+,.,\times)$; and we have $F_n = (F_1)^{\star n}$, for each $n \in \mathbb{N}^*$. Therefore, $(F_n, n \in \mathbb{N})$ is a binomial formula. □

References

[1] J. Connolly, Quantifying Target-Realization Differences, *Clinical Linguistics and Phoentics,* **11**(1997), 267298.

[2] M. Crochemore, D.Gusfield, Combinatorial Pattern Matching, Proceedings of the Fifth Annual Symposium (CPM 94) held in Asilomar, California, June 5–8, 1994, *Lecture Notes in Computer Science*, 807, Springer-Verlag, Berlin, 1994.

[3] M. Farach-Colton, Combinatorial Pattern Matching, Proceedings of the 9th Annual Symposium (CPM 98) held in Piscataway, NJ, July 20–22, 1998, Edited by Martin Farach-Colton, *Lecture Notes in Computer Science*, 1448. Springer-Verlag, Berlin, 1998.

[4] P. Hall, A. Meulen, R. Wall, *Mathematical Methods in Linguistics*, Dordrecht: Kluwer Academic (1990).

[5] R. Hamming, *Coding and Information Theory*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1980.

[6] J. Nerbonne J, Wilbert, *Computationele Vergelijking en Classificatie van Dialecten*, Taal en Tongval 51 (1999).

[7] K. Petyt, *The Study of Dialect : An Introduction to Dialectology*, London: André Deutsch (1980).

[8] H.Rosengren, A non-commutative binomial formula, *J. Geom. Phys.* **32**(2000), 349 - 363.

[9] M. Schützenberger, Une interprétation de certaines solutions de l'équation fonctionnelle: $F(x+y) = F(x)F(y)$, *C. R. Acad. Sci. Paris* **236** (1953), 352 - 353.

[10] Gert Veldhuijzen van Zanten, Gosse Bouma, Khalil Sima'an, Gertjan van Noord, Remko Bonnema , *Evaluation of the NLP Components of the OVIS2 Spoken Dialogue System, OVIS Technical Report* 84 (1999), arXiv:cs/9906014v1.

[11] T. Warnow, Mathematical approaches to comparative linguistics, *Proc. Nat. Acad. Sci. U.S.A.*, , **94**(1997), 6585 - 6590.

[12] M. Waterman Sequence Alignments, *In Mathematical Methods for DNA sequences*, 53–92, CRC, Boca Raton, FL, 1989

[13] M. Waterman M, *Introduction to Computational Biology: Maps, Sequences and Genomes*, London: Chapman and Hall(1995).

Received: April, 2008