STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVIII**, Number 2, 2013

HREBS AND COHESION CHAINS AS SIMILAR TOOLS FOR SEMANTIC TEXT PROPERTIES RESEARCH

DOINA TATAR[(1)], MIHAIELA LUPEA[(1)], AND EPAMINONDAS KAPETANIOS[(2)]

Abstract. In this study it is proven that the Hrebs used in Denotation analysis of texts and Cohesion Chains (defined as a fusion between Lexical Chains and Coreference Chains) represent similar linguistic tools. This result gives us the possibility to extend to Cohesion Chains (CCs) some important indicators as, for example the Kernel of CCs, the topicality of a CC, text concentration, CC-diffuseness and mean diffuseness of the text. Let us mention that nowhere in the Lexical Chains or Coreference Chains literature these kinds of indicators are introduced and used since now. Similarly, some applications of CCs in the study of a text (as for example segmentation or summarization of a text) could be realized starting from hrebs. As an illustration of the similarity between Hrebs and CCs a detailed analysis of the poem "Lacul" by Mihai Eminescu is given.

1. Introduction

Denotation analysis is a complex discipline concerned with the mutual relationships of sentences. An important tool used in Denotation analysis is the concept of *hreb* defined in [10] as a discontinuous text unit that can be presented in a set form or a list form, when the order is important. A hreb contains all entities denoting the same real entity or referring to one another in the text. This basic concept is baptized in this way in honor of L. Hebek ([3]) who introduced measurement in the domain of Denotation analysis, as it is known in Quantitative Linguistics. As we will show, the concepts as Lexical Chain or Coreference Chain (as in Computational Linguistics) subsume the notion of hrebs in the variant of word-hrebs. In fact, we are interested in this paper only in the notion of word-hrebs (for other kinds of hrebs: morpheme-hrebs, phrase-hrebs and sentence-hrebs see [10], [15]).

Received by the editors: March 28, 2013.

2010 *Mathematics Subject Classification.* 68T50,03H65.

Key words and phrases. Lexical Chains, Coreference Chains, Hrebs, Text segmentation, Text summarization.

This paper has been presented at the International Conference KEPT2013: Knowledge Engineering Principles and Techniques, organized by Babeş-Bolyai University, Cluj-Napoca, July 5-7 2013.

We will operate with the concept of Cohesion Chain (CC), defined as a Lexical Chain **or** a Coreference Chain, and we will show the relationship between CCs and hrebs (more exactly a slightly modified kind of word-hrebs, quasi-hrebs). Due to this relation, some denotational properties of a text defined using hrebs could be translated to CCs, in the benefit of the last ones. Similarly, some applications of CCs in the study of a text (as for example segmentation or summarization of a text) could be realized starting from quasi-hrebs.

The structure of the paper is as follows: Section 2 presents the concept of hreb and some indicators of a text connected with it. In Section 3 the Lexical Chains, the Coreference Chains, and their use in segmentation and summarization are introduced. In Section 4 we analyze a poem by Eminescu from the point of view of word-hrebs (as in [10]) and CCs. The paper ends with some conclusions and further work proposal.

2. Hrebs

A word-hreb contains all the words which are synonyms or refer to one of the synonyms. The hrebs usually are constructed using some rules such that a word belongs to one or more hrebs [10]. For example a verb with personal ending (1st and 2nd person) belongs to both the given verb and the person (subject) it overtly refers to. We will slightly modify the definition of a hreb eliminating the above syntactical constraint and will denote the new concept by *quasi-hreb*. Namely, for us verbs with personal ending (1st and 2nd person) belong to the given verb and don't have any connection with the hreb representing the subject of these verbs. In this way, a word belongs to only one *quasi-hreb*, similarly with the property that a word belongs to only one Lexical Chain or Reference Chain (Coherence Chain). The rest of the properties of hrebs mentioned in [10] are unmodified for *quasi-hrebs*: references belong to the quasi-hreb of the word they refer to, e.g. pronouns and Named Entities belong to the basic word; synonyms constitute a common quasi-hreb; articles and prepositions are not considered; adverbs may coincide with adjectives, and may belong to the same quasi-hreb.

According to the information and ordering of entities, [15] defines five kinds of hrebs:

(1) Data-hreb containing the raw data, e.g. words, and the position of each unit in text.

(2) List-hreb containing the data but without the positions of the units in the text.

(3) Set-hreb being the set containing only the lemmas (for word-hrebs).

(4) Ordered set-hreb is identical with (3) but the units are ordered according to a certain principle, e.g. alphabetically, or according to length, frequency, etc.

(5) Ordered position-hreb containing only the positions of units in the given text.

In our example in Section 4 we will use only the cases 1, 2 and 3.

Complete word-hreb analyses of several texts can be found in [15].

2.1. **Denotational analysis with hrebs.** Creating hrebs means a reduction of the text to its fundamental semantic components. Having defined them one can make statements both about the text and the hrebs themselves and obtain new indicators. A short introduction in these indicators is given below (for a complete presentation see [10]):

1. By lemmatizing the words occurring in a List-hreb, and eliminating the duplicates, the corresponding Set-hreb is obtained. If in a Set-hreb there are at least two words (different lemmas), then the hreb belongs to the $Kernel$ (core) of the text, i.e. if $|hreb_i| \geq 2$ then $hreb_i \in Kernel$. The hrebs of a $Kernel$ will be called *kernel hrebs*.

2. An important indicator of a text is the size of the $Kernel$, denoted by $|Kernel|$.

3. Topicality of a set-*kernel hreb* H_i, is calculated as:

$$T(H_i) = \frac{|H_i|}{|Kernel|}$$

4. Kernel concentration is defined as the size of the kernel divided by the total number n of hrebs in the text:

$$KC = \frac{|Kernel|}{n}$$

5. Text concentration is calculated based on the List-hrebs. If H_i is a List-hreb (containing all word-forms, not only lemmas) and L is the number of tokens in the text, then $p_i = |H_i|/L$ is the relative frequency of the List-hreb H_i. Text concentration TC is given as:

$$TC = \sum_{i=1}^{n} p_i^2$$

Relative text concentration, TC_{rel} is defined as:

$$TC_{rel} = \frac{1 - \sqrt{TC}}{1 - 1/\sqrt{n}}$$

6. Hreb diffuseness

The diffuseness D_H of a given hreb H with n_H elements, where the positions of tokens are (in an ascending order) $P = \{pos_1, ..., pos_{n_H}\}$, is defined using the maximal and minimal position of tokens occurring in it:

$$D_H = \frac{pos_{n_H} - pos_1}{n_H}$$

i.e. the difference of the last and the first position divided by the cardinal number of the hreb.

7. Mean diffuseness of the text is:

$$D_{Text} = \frac{1}{K} \sum_{j=1}^{K} D_{H_j}$$

where K is the number of kernel-hrebs ($|Kernel|$) in $Text$.

8. Finally, text compactness is defined as:

$$C = \frac{1 - n/L}{1 - 1/L}$$

where n is the number of hrebs in the text and L is the number of (word-)tokens.

3. Cohesion Chains

3.1. **Lexical Chains.** Lexical Chains (LCs) are sequences of words which are in a lexical cohesion relation with each other and they tend to indicate portions of a text that form semantic units ([8], [11], [5]). The most frequent lexical cohesion relations are the synonymy and the repetition, but could be also hypernyms, hyponyms, etc.. Lexical cohesion relationships between the words of LCs are established using an auxiliary knowledge source such as a dictionary or a thesaurus.

A Lexical Chain could be formalized as:

$$LC_i : [LC_i^1\ (Token_j), \cdots, LC_i^m\ (Token_k)]$$

where the first element of the chain LC_i is the word LC_i^1, representing the token with the number j in the text, the last element of the chain LC_i is the word LC_i^m, representing the token with the number k in the text (where $j < k$), the length of the chain LC_i is m. Because the analysis is made at the level of sentences, usually the sentences where the words occur are indicated. The representation in this case is:

$$LC_i : [LC_i^1\ (S_j), \cdots, LC_i^m\ (S_k)]$$

The first element of the chain LC_i is the word LC_i^1, and occurs in the sentence S_j , the last element of the chain LC_i is the word LC_i^m, and occurs in the sentence S_k of the text (where $j < k$).

LCs could further serve as a basis for Text segmentation and Text summarization (see [4]). The first paper which used LCs (manually built) to indicate the structure of a text was that of Morris and Hirst ([7]), and it relies on the hierarchical structure of Roget's thesaurus to find semantic relations between words. Since the chains are used to structure the text according to the attentional/intentional theory of Grosz and Sidner theory, ([1]), their algorithm divides texts into segments which form hierarchical structures (each segment is represented by the span of a LC). Some algorithms for linear segmentation (as opposite to hierarchical segmentation) are given in [12], [13], [14]. In all these algorithms it is applied the following remark of Hearst 1997 [2]: *There are certain points at which there may be radical changes in space, time, character configuration, event structure(...). At points where all of these change in a maximal way, an episode boundary is strongly present.* The algorithms are based on different ways of scoring the sentences of a text and then observing the graph of the score function. In this paper we introduce two new scoring functions for sentences (in the next subsection).

Let us remark that linear segmentation and the (extractive) summarization are two interdependent goals: good segmentation of a text could improve the summarization ([4]). Moreover, the rule of extracting sentences from the segments is decisive for the quality of the summary. Some largely applied strategies (rules) are ([12]):

1. The first sentence of a segment is selected.

2. For each segment the sentence with a maximal score is considered the most important for this segment, and hence it is selected (for example, the minima in the graph of the below $Score^1$ and $Score^2$ functions represent the sentences candidates for boundaries between segments of a text).

3. From each segment the most informative sentence (the least similar) relative to the previously selected sentences is picked up.

Thus, one can say that determining a segmentation of a text and using a strategy (1, 2 or 3), a summary of the text can be obtained, as well.

3.2. **Coreference Chains.** Coreference Chains are *chains of antecedents-anaphors* of a text. A complete study of Coreference Chains is the textbook [6]. A Coreference Chain contains the occurrences of the entities identified as antecedents for a given anaphor and also the occurrences of this anaphor.

The formalization of a Coreference Chain is as follows:

$$CR_i : [CR_i^1\ (Token_j), \cdots , CR_i^m\ (Token_k)], (where\, j < k\,)$$

or

$$CR_i : [CR_i^1\ (S_j), \cdots, CR_i^m\ (S_k)], (where\, j \leq k\,)$$

depending on the marks (tokens or sentences) picked out.

In the same way as the Lexical Chains, Coreference Chains express the cohesion of a text. The algorithms of segmentation (and summarization) of a text based on Lexical Chains could be adapted for Coreference Chains. In this paper we refer to both Lexical Chains and Coreference Chains by the name of Cohesion Chains.

3.3. **Scoring the sentences by Cohesion Chains.** Cohesion Chains (CCs) defined as in the above sections could be used to score the sentences such that when this score is low, cohesion is low, and thus the sentence is a candidate for a boundary between segments; similarly for a high score (a high cohesion) and the non-boundary feature of a sentence. In this paper we propose the following two new functions of score for sentences:

$$Score^1(S_i) = \frac{the\, number\, of\, tokens\, in\, S_i\, contained\, in\, at\, least\, one\, CC}{the\, number\, of\, tokens\, in\, S_i.}$$

Let us remark that $0 \leq Score^1(S_i) \leq 1$. When $Score^1(S_i) = 0$ (or close to 0), S_i is a candidate for a boundary between segments because S_i has a low connection with other sentences. When $Score^1(S_i) = 1$ (or close to 1), S_i is "very" internal for a segment. So, observing the graph of the function $Score^1(S_i)$ we could determine the segments of a text.

The second proposed scoring function is:

$$Score^2(S_i) = \frac{the\, number\, of\, CCs\, which\, traverse\, S_i}{the\, total\, number\, of\, CCs\, in\, the\, text}$$

Again $0 \leq Score^2(S_i) \leq 1$ and the above remarks remain valid: when $Score^2(S_i)$ is 0 (or close to 0), S_i is a candidate for a boundary between segments because S_i has a low connection with the other sentences. When $Score^2(S_i)$ is 1 (or close to 1), S_i is "very" internal for a segment.

As a final remark, let us observe that the hrebs (quasi-hrebs) could be used exactly in the same way to score the sentences: it is enough to put *quasi-hrebs* instead of *CCs* in the definitions for $Score^1(S_i)$ and $Score^2(S_i)$. Thus, hrebs (quasi-hrebs) could serve to segment and/or summarize texts.

In the same way, the indicators 1-8 used in Denotational analysis with hrebs could be extended to CCs. Let us remark that quasi-hrebs (and thus CCs) are defined in the Data-hrebs format. This is accordingly with the definition of Lexical Chains where the most important (frequent) lexical relation which is present in a Lexical Chain is the repetition [11]. The more frequently a word is repeated in a Lexical Chain, the more important this Lexical Chain is. Obtaining CCs from Data-hrebs (duplicates are not eliminated), we will impose the condition to a kernel CC to have at least a given number of elements. In other words, a kernel CC must have a size bigger than a minimal one. Further, the topicality of a kernel CC, text concentration, CC-diffuseness and mean diffuseness of the text could be defined.

Let us mention that nowhere in the Lexical Chains or Coreference Chains literature these kinds of indicators are introduced up to now.

4. Example in Romanian

For the Eminescu's poem "Lacul" we will exemplify **hrebs, quasi-hrebs and CCs**, and the relationships between them. We will begin with the **Rules** for hrebs formation in Romanian language [10].

Rules of hrebs formation for the Romanian language

The Rules for hrebs are of the form: $a \in B$. Here a is an expression containing a special element called *pos* indicator which is written in italic (*pos* is for *part of speech*). Particularly, a could be formed only from the *pos* indicator. B is a (name for a) given hreb written with capital letters. More exactly, the Rule $a \in B$ means: a (or *pos* indicator of a) is an element of the hreb B . The connection between a and B will result from the word used for *pos* indicator. As a word-form could be contained in more than one hreb, in the application of rules it is possible to obtain a result as: $a \in B, C, \cdots$ meaning: a is an element of hreb B and hreb C and $\cdots$. The rules are valid only for the *pos* of a being *noun*, *verb*, *adjective*, *adverb*, *pronoun*.

RULES:

R1. $verb \in VERB$

R2. personal ending of a *verb*, which could be a *noun* or a *pronoun*, $\in NOUN\, or\, PRONOUN$

R3. synonym of a $verb \in VERB$

R4. *pronoun* referring to a $noun \in NOUN$

R5. *pronoun* referring to a Named Entity $\in NAMED\ ENTITY.$

R6. synonym of a Named Entity $\in NAMED\ ENTITY.$

R7. non-referring $pronoun \in PRONOUN$

R8. $noun \in NOUN$

R9. synonym of a $noun \in NOUN$

R10. $adjective \in ADJECTIVE$

R11. synonym of an $adjective \in ADJECTIVE$

R12. $adverb \in ADVERB$

R13.synonym of an $adverb \in ADVERB$

The Rules 1-13 could be summarized as follows: a noun, its synonyms, referring pronouns and personal endings in a verb belong all to the given noun; a Named Entity, its synonyms, referring pronouns and personal endings in a verb belong all to the given Named Entity; a verb in all its forms, its synonyms, belong to the given verb, however, the personal endings belong also to the respective noun; an adjective (adverb) and its synonyms belong all to the given adjective (adverb).

We illustrate the rules as applied to the poem "Lacul". Namely, we will make a denotation of tokens in the poem, then we will extract:

- **A. Hrebs** (Table 1),
- **B. Quasi-hrebs** (Table 2)
- **C. Cohesion Chains** (Table 3).

The tokens numbered are only nouns, verbs, adjectives, adverbs, and pronouns (in this poem do not exist Named Entities).

LACUL (denotation of tokens)

(S1) Lacul (1) codrilor (2) albastru (3)
Nuferi (4) galbeni (5) îl (6) încarcă (7).
(S2) Tresărind (8) în cercuri (9) albe (10)
El (11) cutremură (12) o barcă (13).
(S3) Şi eu (14) trec (15) de-a lung (16) de maluri (17),
Parc-ascult (18) şi parc-aştept (19)
Ea (20) din trestii (21) să răsară (22)
Şi să-mi (23) cadă (24) lin (25) pe piept (26).
(S4) Să sărim (27) în luntrea (28) mică (29) ,
Îngînaţi (30) de glas (31) de ape (32),
Şi să scap (33) din mână (34) cârma (35),
Şi lopeţile (36) să-mi (37) scape (38).
(S5) Să plutim (39) cuprinşi (40) de farmec (41)
Sub lumina (42) blîndei (43) lune (44).
(S6)Vîntu-n (45) trestii (46) lin (47) foşnească (48),
Unduioasa (49) apă (50) sune (51)!
(S7) Dar nu vine (52)... **(S8)**Singuratic (53)
În zadar (54) suspin (55) şi sufăr (56)
Lîngă lacul (57) cel albastru (58)
Încărcat (59) cu flori (60) de nufăr (61).

Hreb	Elements of Data-hreb	SDH	SSH
EU	(eu 14, **trec 15, -ascult 18, -aştept 19, scap 33**, -mi 23,		
(EU cont)	-mi 37 , **suspin 55, sufăr 56**)	9	8
LAC	(lacul 1, il 6, **tresărind 8**, el 11, **cutremură 12**, lacul 57)	6	5
EA	(ea 20, **răsară 22, cadă 24, vine 52**)	4	4
NUFĂR	(nuferi 4, **incarcă** 7, nufăr 61)	3	2
APĂ	(ape 32, apă 50, **sune** 51)	3	2
NOI	(**sărim 27, plutim 39**)	2	2
BARCĂ	(barca 13, luntrea 28)	2	2
TRESTIE	(trestii 21, trestii 46)	2	1
ALBASTRU	(albastru 3, albastru 58)	2	1
A PĂREA	(parc- 18, parc- 19)	2	1
LIN	(lin 25, lin 47)	2	1
A SCĂPA	(scap 33, scape 38)	2	1

Table 1. **A.** The hrebs with the size bigger than 1 (extracted from the poem **Lacul**)

4.1. **From Hrebs to Cohesion Chains.** By the application of the above mentioned rules a total number of 51 hrebs are obtained. From all these, only 12 hrebs presented in Table 1 contain more than one element. In Table 1 the hrebs are constituted as Data-hrebs, where SDH means "Size of Data-hreb" and SSH means "Size of Set-hreb".

The names of all 51 hrebs are as follows:

A ASCULTA, A AŞTEPTA, A ÎNCĂRCA, A CĂDEA, A CUTREMURA, A FOŞNI, A PĂREA, A PLUTI, A RĂSĂRI, A SĂRI, A SCĂPA, A SUFERI, A SUNA, A SUSPINA, A TRECE, A TRESĂRI, A VENI, ALB, ALBASTRU, APĂ, BARCĂ, BLÂNDĂ, CÂRMĂ, CERC, CODRU, CUPRINS, EU, EA, FARMEC, FLOARE, GALBEN, GLAS, ÎNCĂRCAT, ÎNGÂNAT, LAC, LIN, LOPATĂ, LUMINĂ, LUNĂ, LUNG, MAL, MÂNĂ, MIC, NOI, NUFĂR, PIEPT, SINGURATIC, TRESTIE, UNDUIOASĂ, VÂNT, ZADAR.

From the set of Rules R1-R13, the Rule R2 makes the difference when the quasi-hrebs are calculated. This rule is reproduced here:

R2. personal ending of a *verb*, which could be a *noun* or a *pronoun*, $\in NOUN\, or\, PRONOUN$

In Table 1 are bold marked all the verbs which are contained in a NOUN or PRONOUN hreb due to the Rule R2. All these verbs are not present in Table

Quasi-hreb	Elements of Data-hreb	SDH	SSH
EU	(eu 14, , -mi 23, -mi 37)	3	2
LAC	(lacul 1, il 6, , el 11, , lacul 57)	4	3
EA	(ea 20)	1	1
NUFĂR	(nuferi 4, nufăr 61)	2	1
APĂ	(ape 32, apă 50)	2	1
BARCĂ	(barca 13, luntrea 28)	2	2
TRESTIE	(trestii 21, trestii 46)	2	1
ALBASTRU	(albastru 3, albastru 58)	2	1
A PĂREA	(parc- 18, parc- 19)	2	1
LIN	(lin 25, lin 47)	2	1
A SCĂPA	(scap 33, scape 38)	2	1

Table 2. **B.** The quasi-hrebs extracted from the poem **Lacul**

2, the table of quasi-hrebs. As a remark, the hreb ”NOI” is not a quasi-hreb, because both its elements (sărim 27, plutim 39) are obtained by Rule R2.

Let us remember that Lexical Chains are sequences of words which are in a lexical cohesion relation (synonymy, repetition, hypernymy, hyponymy, etc) with each other. Coreference Chains are chains of antecedents-anaphors of a text. Examining Table 2 of quasi-hrebs, we observe that: the quasi-hreb EU corresponds to a Coreference Chain (eu 14, -mi 23, -mi 37), the quasi-hreb LAC to a Coreference Chain (lacul 1, il 6, el 11, lacul 57). The quasi-hreb EA is not a chain (it has only one element). The rest of quasi-hrebs represent Lexical Chains: (nuferi 4, nufăr 61), (ape 32, apă 50), (barca 13, luntrea 28), (trestii 21, trestii 46), (albastru 3, albastru 58), (parc- 18, parc- 19), (lin 25, lin 47), (scap 33, scape 38). Table 3 contains the Cohesion Chains denoted as we will use further. We obtained CCs from the Data-hrebs, and the length of a Cohesion Chain is given in the SDH column, because the duplicates are not eliminated (as in SSH column).

Calculating the scores $Score^1$ for each sentence are obtained the following results:

$Score^1(S_1) = 4/7 = 0.57$
$Score^1(S_2) = 2/6 = 0.33$
$Score^1(S_3) = 6/13 = 0.46$
$Score^1(S_4) = 5/12 = 0.42$
$Score^1(S_5) = 0/6 = 0.$
$Score^1(S_6) = 3/7 = 0.43$
$Score^1(S_7) = 0/1 = 0.$

Denotation of CC	Elements of CC	Length of CC
CC1	(eu 14, -mi 23, -mi 37)	3
CC2	(lacul 1, il 6, el 11, , lacul 57)	4
CC3	(nuferi 4, nufăr 61)	2
CC4	(ape 32, apă 50)	2
CC5	(barca 13, luntrea 28)	2
CC6	(trestii 21, trestii 46)	2
CC7	(albastru 3, albastru 58)	2
CC8	(parc- 18, parc- 19)	2
CC9	(lin 25, lin 47)	2
CC10	(scap 33, scape 38)	2

TABLE 3. **C.** Cohesion Chains extracted from the poem **Lacul**

$Score^1(S_8) = 3/9 = 0.33$

Taking as segment boundaries the sentences with minimal score, the text is divided in 4 segments: $Seg1 = [S_1, S_2]; Seg2 = [S_3, S_5]; Seg3 = [S_6, S_7]; Seg4 = [S_8]$ or 3 segments: $Seg1 = [S_1, S_2]; Seg2 = [S_3, S_5]; Seg3 = [S_6, S_8]$ if mono-sentence segments are not permitted.

Scoring with $Score^2$ formula, the results are as following:

$Score^2(S_1) = 3/10 = 0.30$
$Score^2(S_2) = 4/10 = 0.40$
$Score^2(S_3) = 8/10 = 0.80$
$Score^2(S_4) = 9/10 = 0.90$
$Score^2(S_5) = 6/10 = 0.60$
$Score^2(S_6) = 6/10 = 0.60$
$Score^2(S_7) = 3/10 = 0.30$
$Score^2(S_8) = 3/10 = 0.30$

The text has only one segment $[S_1, S_8]$, with the most "internal" sentence S_4. A summary of the poem using $Score^1$ is formed by the sentences: S_1, S_3, S_6 and using $Score^2$, by the sentence S_1. In both cases the rule one (Section 3.1) has been applied.

4.2. **Indicators of Cohesion Chains.** Let us suggest how the indicators in Section 2.1 could be defined for the Cohesion Chains CC1 to CC10.

- Kernel CCs : Considering the minimal size of a kernel CC being 2, all CCs are in *Kernel*. Considering the minimal size of a kernel CC being 3, only CC1 and CC2 are in $Kernel$. The last supposition is more realistic, since a CC has always at least 2 elements;

- The size of the $Kernel$ is 2, in the last above case;
- Topicality of the kernel CC denoted by CC1 is $3/2 = 1.5$ and topicality of CC2 is $4/2 = 2$;
- Kernel concentration is $KC = 2/10 = 0.2$;
- $p_1 = 3/61; p_2 = 4/61; p_i = 2/61,\ i = 3\, to\, 10$. Text concentration is $TC = 0.0151$ and Relative Text concentration is $TC_{Rel} = 1.2830$;
- Diffuseness for each CC is as follows:
  $D_{CC1} = (37 - 14)/3 = 7.66; D_{CC2} = (57 - 1)/4 = 14; D_{CC3} = (61 - 4)/2 = 28.5; D_{CC4} = (50 - 32)/2 = 9; D_{CC5} = (28 - 13)/2 = 7.5; D_{CC6} = (46 - 21)/2 = 12.5; D_{CC7} = (58 - 3)/2 = 27.5; D_{CC8} = (19-18)/2 = 0.5; D_{CC9} = (47-25)/2 = 11; D_{CC10} = (38-33)/2 = 2.5$
- Mean diffuseness of the text is $D_{Text} = 10.75$;
- Text compactness is $C = (1 - 10/61)/(1 - 1/61) = 0.8505$.

The above indicators could make differences between CCs, such that some of them are kernel CCs, or have a higher topicality and/or diffuseness.

5. Conclusions and Further work

Lexical Chains and Coreference Chains (CCs) are intensively studied, but few indicators are standard for them. The indicators inspired from the hrebs must be studied and adopted for CCs. These indicators, in the context of some applications using CCs, could become instruments for the evaluation of these applications and for improving them. For example, there is a large debate about how to select CCs to construct the summaries of a text: selecting long or short CCs is one of the questions. Using only kernel CCs, or kernel CCs with a high topicality and /or high diffuseness could be a solution.

As a general remark, Quantitative Linguistics and Computational Linguistics are considered two distinct fields with their own journals, techniques and specialists. It is important to identify those parts they have in common, and to try to extract the advantage from this commonality. This paper is a step toward this desirable aim.

References

[1] B. Grosz and C. Sidner. 1986. "*Attention, Intentions and the Structure of Discourse*". Computational Linguistics 12: 175 - 204.

[2] M. Hearst 1997. "*TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages*". Computational Linguistics 23: 33 - 76.

[3] L. Hrebicek 1997. "*Lectures on Text Theory*". Prague: Oriental Institute.

[4] E. Kapetanios, D. Tatar and C. Sacarea. 2013. "*Natural Language Processing: semantic aspects*", Science Publishers, (to appear).

[5] A. Labadie and V. Prince. 2008. "*Finding text boundaries and finding topic boundaries: two different tasks?*" Proceedings of GoTAL08.

[6] R. Mitkov 2002. "*Anaphora Resolution*", Pearson Education, Longman.

[7] J. Morris and G. Hirst. 1991. "*Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text*". Computational Linguistics 17: 21 - 48.

[8] M. Okumura and T. Honda. 1994. "*WSD and text segmentation based on lexical cohesion*", 755 - 761. Proceedings of COLING-94.

[9] I.I.Popescu, J. Macutek, E. Kelih, R. Cech, K. H. Best, and G. Altmann. 2010. "*Vectors and Codes of Text*". Studies in Quantitative Linguistics 8, RAM Verlag.

[10] I.I. Popescu, M. Lupea, D. Tatar, and G. Altmann. 2013. "*Quantitative analysis of poetry*", Ed. Mouton de Gruyter, to appear.

[11] N. Stokes, J. Carthy and A.F. Smeaton. 2004. "*Select: a lexical cohesion based news story segmentation system*". AI Communications, 17(1): 3 - 12.

[12] D. Tatar, A. Mihis and D. Lupsa. 2008. "*Text Entailment for Logical Segmentation and Summarization*", 233 – 244. In Kapetanios, E., Sugumaran, V., Spiliopoulou, M. [eds.] Proceedings of 13th International Conference on Applications of Natural Language to Information Systems, London, UK. (LNCS 5039).

[13] D. Tatar, E. Tamaianu-Morita and G. Serban-Czibula. 2009. "*Segmenting text by lexical chains distribution*", Proceedings of Knowledge Engineering Principles and Techniques (KEPT 2009), University Press, Cluj-Napoca, Romania.

[14] D. Tatar, M. Lupea and Z. Marian. 2011. "*Text summarization by Formal Concept Analysis approach*". Proceedings of Knowledge Engineering Principles and Techniques (KEPT 2011), Cluj-Napoca, Romania.

[15] A. Ziegler and G. Altmann. 2002. "*Denotative Textanalyse*", Wien, Praesens.

[(1)] Babeş-Bolyai University, Department of Computer Science, Cluj-Napoca, Romania
E-mail address: dtatar@cs.ubbcluj.ro

E-mail address: lupea@cs.ubbcluj.ro

[(2)] University of Westminster, UK
E-mail address: E.Kapetanios@westminster.ac.uk