DISCOURSE COHERENCE – A BUILT-IN COGNITIVE MECHANISM?

Dan CRISTEA* ** and Adrian IFTENE*

* Alexandru Ioan Cuza University, Faculty of Computer Science
** Romanian Academy, Institute for Theoretical Computer Science
E-mail: {dcristea, aiftene}@info.uaic.ro

Abstract: The paper presents a proposal for correlating human’s performance in discourse coherence with a linear model of immediate memory. We begin by estimating experimentally the discourse coherence as produced by humans, using for that a measure based on Centering transitions. Then we introduce a parametrised model of immediate memory, and we propose a simple access cost model, which mimics cognitive effort during discourse processing. We show that an agent, equipped with the most economical model of immediate memory and manifesting a greedy behaviour in choosing the focus at each step, produces discourses having similar qualities as those produced by humans.

Key words: Discourse coherence, cognitive load/effort, immediate memory, Centering, economy principle applied to language.

1. INTRODUCTION

Human discourse is, most often than not, a coherent one. When humans communicate to others a situation or an argument, the common result is a message made of a sequence of utterances which are easy to understand. Producing easy to understand discourses is almost a reflex behaviour. Unless the discourse is the result of a damaged brain, which has difficulty to assemble utterances and is prone to a random sequencing, and unless the discourse is on purpose encrypted in order to make it difficult to decipher by the reader (as, sometimes, in the literary works of writers like William Falkner, Marcel Proust, Gabriel Garcia Marques or Herta Müller), the common human behaviour is one which produces simple to understand discourses. In this paper we argue that, provided the content to transmit is clear, it is cognitively cheaper to produce coherent discourses than incoherent ones. We show that producing and understanding discourse at a quality similar to that characteristic to humans can be modelled by very simple mechanisms. This, however, should not be taken as a proof that the human mind is indeed built this way. It only shows a possible way. Altogether, demonstration raises the credibility of theories which advocate an innate (Chomsky, 1965), as compared to those supporting an acquired view over features of language (Piaget, 1985) (to name only the representative names of classical linguistic schools), at least for those features influencing the human performance in discourse. It is possible that humans possess a cognitive machinery which enable them to produce and to process discourses at low costs and this machinery is also responsible for the default coherence of their discourses, provided the agents have a clear image of what has to be uttered.

Interest in determining the cognitive capacity of humans to produce and process discourses is extremely vivid. See (Pan & Snow, 1999; Cicourel, 2006; Pho, 2008), for instance. On the other
hand, a range of linguistic theories analyse empirically the properties of human discourse. Of importance for our construction is Centering (Grosz et al., 1986, 1995; Brennan et al, 1987), a theory of local coherence, which reveals two important interrelationships between features of human discourse: one is explaining restrictions for the use of pronouns in sequences of utterances, the other is defining a typology of transitions between utterances, which suggests a coherence measure. According to Centering, to each utterance corresponds a list of forward-looking centers, which are semantic entities mentioned in the discourse, or discourse entities (Marcu, 2000). This list, noted usually \( C_U \) is ordered according to syntactic criteria, not the same in all languages (see, for instance, (Kameyama, 1998) for Japanese, (di Eugenio, 1998) for Italian, or (Strube & Hahn, 1999) for German). Then, each utterance has a unique backward-looking center, \( C_b \), and a principal center, \( C_p \). \( C_b \) of an utterance \( U_n \) is defined as the first center of the previous utterance \( U_{n-1} \) which is realized also in the current utterance, while \( C_p \) is the most prominent center of \( C_f \). The transitions between pairs of adjacent utterances define degrees of easiness of pursuing the sequence of utterances, or turn takings, in monologues and dialogues. We will refer in this section to the Centering second rule, which states that there are four transitions, from easiest to most difficult: continuing (CON), retaining (RET), smooth shift (SSH) and abrupt shift (ASH), in this order, all evaluating the relationship between consecutive \( C_f \)'s and that between the \( C_b \) and the \( C_p \) of an utterance. When there is no intersection between the \( C_f \) lists of consecutive utterances, a \( C_b \) is lacking (we will call this No \( C_b \), and note it as NOC).

Poesio et al. (2004) have tried a parametrisation of Centering, by questioning each of the theory’s statements and evaluating them on a corpus.

Theories on discourse production and interpretation, as for instance (Grosz & Sidner, 1986; Walker, 1989; Jensen & Lisman, 1998), often bring into discussion a model of immediate memory as being responsible for the operations which allow recuperation of mentions. A short term memory model is the minimum cognitive device which makes possible accessing and managing discourse entities, over short interval of time, such that an entity already introduced is recuperated once it is mentioned again. The short term memory should be quick. If an entity is not found in this working memory, it will be searched for in a long term memory, but the access there could be slower. As mentioned already above in this section, identifying elements in memory (either short or long term) is essential for a coherent communication.

We are not aware of researches aimed to measure the coherence of human discourses in different settings. Let’s note that such an investigation is not easy to do. An event-focussed situation, therefore one implying involvement of time, is not proper to act as a discourse theme in this type of investigation, because it yields a discourse constrained with respect to the sequence of elementary utterances. In such a discourse, coherence mainly follows as a by-product of the narration and cannot be the result of a cognitive capacity. To study the unconstrained proficiency to produce coherent discourse one needs a static reality as a theme, one made up of objects found in diverse relations ones to the others. Such a scene, with relations between objects uttered in any order, produces the same final result in the mind of the hearer, although re-assembling by hearer could be cheaper or more expensive. The difference, therefore, stays in coherence.

This paper is structured as follows. The next section presents a set of simplifications and conventions for the constructions to follow. Section 3 puts forward a set of 4 hypotheses, which we will try to defend in the rest of the paper. We describe in section 4 two models of immediate memory and we propose a measure for evaluating the access effort to memory. In section 5, we present experiments that are aimed to evaluate the hypotheses, excepting for one which is considered to hold by default. Finally, in section 6, we discuss the results and formulate conclusions.

This work is triggered by the intention to build an artificial agent displaying discourse capabilities similar to those of humans.
In this paper we will use a number of simplifications. One of the most important is that, for abstraction reasons, we will consider that a discourse is made of a sequence of elementary utterances, of the type \(<\text{subject}> <\text{predicate}> <\text{object}>\). Since we measure coherence following a Centering-inspired metric, when discussing coherence we are interested only to recuperate the meaning of referential expressions in the context of similar mentions. As such, predicates are of little relevance and we will ignore them. An utterance becomes therefore a pair of two referential expressions, and a discourse can be seen as a sequence of such pairs. In terms of costs, this means that for the production, as well as for the interpretation, predicates are cost-free. Referential expressions represent textual mentions of discourse entities, which are concepts, characters, etc., if considered situated on a semantic level. Seen on this level, a reality or a situation which the speaker intends to transmit can be simplified to a graph, where entities (nodes) are connected by relations or predicates (edges). This graph will become a discourse when it will be “uttered”, meaning – serialized, i.e. expressed as a sequence of node-node pairs (if names of relations/predicates are ignored). We will also make no difference between the situation when the reality (graph) to be communicated should be retrieved from a long term memory or is visualized directly (“under the eyes”) by the agent.

As said already, and in connection with this simplified perspective of discourse, we will consider that a discourse is coherent if its entities are easily recoverable from the previous discourse. Coherence, therefore, has a lot to do with a memory machinery, because recuperation of previously uttered entities is an activity which involves a memory. The memory is a place where entities are stored during discourse unfolding and where they are found when mentioned again. If, during a memory search, such an entity is found, then a connection is made and the hearer “understands” that reference. If, on the contrary, the entity is not found, than it is supposed to be new. For instance, it is clear that if in the sequence

1. John loves Mary.
2. He offers her candies.

he would not be identified as John and her as Mary, the understanding would be exposed to a failure, because a person different than John will be seen as offering candies to a person different than Mary.

As close to anaphora resolution this could seem, we are actually not concerned here with this aspect. In other words, we do not discuss idiosyncrasies of antecedent recuperation of an anaphor in discourse. As such, we can make another simplification, by considering that each discourse entity has a unique referential string on the textual level.

As said, to measure coherence we will stick to Centering (Grosz et al., 1986, 1995; Bresnan et al., 1987), as a theory of local coherence. The Centering transitions between consecutive utterance of a discourse segment are, from the most coherent (and easiest to make) to the least coherent (hardest to make): CON, RET, SSH, ASH, and NOC. The last one expresses a difficult transition between two consecutive utterances and, conforming to some scholars (Walker et al., 1998) could be an indication of a segment boundary. We will use a more permissive view on discourse segmentation, by actually totally ignoring segment boundaries and counting NOC as a typical transition, together with all the others. When there is no \(C_b\) between two adjacent utterances it simply means a (temporal) change of topic, not necessarily a discourse segment boundary.

---

1 In view with our simplified notation of S V O utterances, the relation (predicate) here would be offers_candies.
To express numerically the cognitive load, we will use the values: CON=0, RET=1, SSH=2, ABH=3, and NOC=4. As such, the relative higher cognitive load at a point where there is no common center between adjacent utterances is mirrored by the corresponding high Centering value (4). By using this range of values we state that these five transitions, taken in this order, correctly model descending degrees of discourse coherence and, accordingly, ascending degrees of cognitive load.

We will use these values to compute a **global coherence score** of a discourse of length $N$ (utterances) by summing up the $N-1$ transition values and dividing the sum to the number of transitions $(N-1)$. A discourse will be called **Centering-optimum** if, among all possible permutations of utterances (in which centers are stable, therefore the realization relations between surface references and semantic representations are frozen to those in the original variant), the global coherence score is minimum. Following this definition, a Centering-optimum discourse is the smoothest possible verbalisation of a reality.

3. **HYPOTHESIS**

Our construction is based on a number of hypotheses, which we introduce below.

**H1. On the near-optimum coherence of the human discourse**

*On average, human discourses have a high degree of coherence, but rarely the highest.*

As said already, to measure human discourse coherence, we will make use of the five Centering transitions. The hypothesis says that, on average, and for descriptions which are free of semantic/time constraints, discourses generated by humans are close to Centering-optimum, as defined in section 2. In other words, unless the discourse is the result of a damaged brain or the agent has a special interest to encrypt the message, the agents have a tendency to produce coherent discourses.

**H2. On the “best” immediate memory model**

*The built-in model of immediate memory does influence coherence.*

Among all possible models of immediate memory, there should be one, let’s call it **OSTMem** (from Optimal Short-Term Memory), which is the best for discourse management, in the sense that it optimizes the total cost of transmitting a graph between a speaker and a hearer. This means that distinct ways in which the internal short term memory is organized induces different processing loads in reading and writing.

Empirical evidences for this hypothesis are:

- it should make a difference whether the access mechanics of the memory resembles those of a stack or of a queue. For instance in stack-type buffer an entity recently uttered is found at once, while in a queue-type buffer – only after searching a certain part of the buffer;
- the length of the buffer is important, because a very short buffer means a low capacity of recording, therefore accommodating only a small number of entities. This means a high forgetting rate;
- on the other hand, a very long memory triggers a long decision time to detect the first mentions (because the first mentioned are recognised by memory search fails).

**H3. On the local strategy in choosing the focus**

*An agent is tempted to utter next that entity which is handiest to grasp.*

This hypothesis says that, at any point during the production of a discourse, agents tend to be lazy with respect to the choice for next utterance. If an agent has several possible choices to make at a
certain moment, then he will choose one among those which are most straightforward to choose, i.e. are nearby, are at hand. It says that the agents are inclined to spend minimum effort in choosing what to utter at every moment. This resides to an **immediate minimisation strategy**, a sort of **greedy approach**.

Forming in memory long, un-uttered yet, discourses, for the sake of improving them mentally until a coherence optimum is found, with the ultimate goal to ease the comprehension by the hearer, is not a common behaviour. Instead, the common speaker tries to optimize her/his effort of producing discourse at each step.

**H4. On the correlation between the memory model and the local strategy in acquiring a human like performance in direct communication**

*If the memory model is OSTMem and the agent uses a greedy approach in selecting the next utterance than any transmitted graph will be close to Centering-optimum.*

This hypothesis is central for this paper. It says that the human performance to produce discourse is rooted in built-in cognitive mechanisms: a certain adaptation of the agents’ mind with respect to the memory model that optimises the costs of recuperating mentioned entities in a discourse and an innate preference for low efforts in choosing what to say next. In fact, both these triggers are manifestations of the same economy principle in language, which has often been recognised (Whitney, 1877; Vendryes, 1939; Chomsky, 1995; Kager, 1999). We hypothesise therefore that the blind application of the same economy principles twice abides to the production of a comprehensive discourse.

### 4. THE MEMORY MODEL

We will consider that the goal of the saying can be segmented in smaller parts and each part can be abstracted as a graph. In the following we will concentrate only on the transmission of such a graph, leaving apart the assembling of these parts into a long discourse displaying a rhetorical structure. We place ourselves, therefore, at a local level. The agent communicates the graph as a sequence of utterances, each expressing the knowledge that an edge (relation, predicate) links two nodes (discourse entities).

In the process of uttering the graph, the agent should be aware whether a node has already been uttered or not. The reason for this is that a mentioned entity (node of the graph) will most probably be mentioned differently, for instance, by using pronouns instead of using more powerful qualifying expressions (see Cristea et al., 2009 for a model of anaphora in the evolutionary approach). Similarly, the hearer should recognize an entity which has been mentioned in the preceding discourse, or otherwise it will not be able to accumulate the related facts as applying to the same entity.

These considerations make clear the need for a short term memory for production and interpretation of the discourse. In this section we will propose two models of short term memory, which are built on the skeleton of a general type of data structure. In accordance with other approaches in the formalization of human immediate memory related to language (Atkinson & Shiffrin, 1968; Anderson & Bower, 1980; Jensen & Lisman, 1998) the memory is seen as a list structure (buffer) of a certain length where read and write operations are performed at specific positions. As will be seen, the list will not be aprioristically ascribed to any known type (among
which – the stack model and the queue model).\(^2\) Each position in this list will be occupied with a mention of a discourse structure.

Since our intention is to find one optimal model in terms of memory costs, we will consider the following costs for the basic memory operations:

- read: the element at the position of the read head is read; the cost is constant \((C_1)\);
- search: go linearly from the first position to the end of the list; the cost of the search should be equal to the number of reads until the element is found: it is \(C_1\) if the looked-for element is on the first position; in average it is \(C_1 \times L/2\) \((L = \text{length of the list})\), and \(C_1 \times (L+1)\) if the looked-for element is not in the list;
- delete: a found element is deleted; the cost of delete is constant \((C_2)\);
- write: an element is inserted on the position of the write pointer; the cost of the write is constant \((C_3)\).

When uttering an edge that connects nodes \(x\) and \(y\), as well as when hearing the utterance "\(x y\)"\(^3\), the operations performed with the memory \(M\) are as follows:

1. search node \(x\) in \(M\);
2. if found, delete it from the position it was found;
3. write \(x\) in \(M\);
4. search node \(y\) in \(M\);
5. if found, delete it from the position it was found;
6. write \(y\) in \(M\);

One can notice that the writing of the mention of a node in \(M\) is performed irrespective whether it was found in \(M\) or not. By deleting the mention in case it exists, we make sure that only one occurrence of each entity name will ever be there. And, in any case, the memory is updated with respect to the node under operation (in focus). The assumption is that writing an entity in memory is equivalent to bringing it in focus.

This makes the total cost of an utterance "\(x y\)" be:

\[
c = \text{cost(search}(x, M)) \times C_1 + \text{if(exists}(x, M), C_2, 0) + C_3 + \text{cost(search}(y, M)) \times C_1 + \text{if(exists}(y, M), C_2, 0) + C_3
\]

\[
(1)
\]

4.1. The linear model: Mem1

In this model, the write position is fixed at a certain relative, \(d_w\), distance between the left extreme and the right extreme of the list (see Figure 1). The read head moves right in searching the element. When the memory is not full, the size of the memory, therefore also the right extreme, is not fixed. When the memory is full, writing an element in the write position means shifting the rest of the elements to right or to left and forgetting one element.

\(^2\) We will make no speculations at this moment whether this structure is the result of an evolutionary process, acquired during the life time of the human agent, during a long series of interactions, or is innate, cabled in the cognitive mechanism of humans.

\(^3\) As said, in the present model the processing of a relation is ignored.
Discourse Coherence – a Build-in Cognitive Mechanism?

The behaviour of Mem₁ when an entity e is processed is as follows:
- R-pointer at 0; W-pointer at position $d_w$ (a parameter);
- search for e starting at the R-pointer and advance the R-pointer one position right after each element;
- if e is found then delete e;
- write e at the W-pointer;
- advance the W-pointer one position to the right;

The concept of focus is dissolved in the memory model, actually transforming the binary status (in focus/not in focus) into a continuum (anywhere in between the first position of the memory and the last position). The element in focus should be that element which necessitates the minimum access time to be retrieved. This is the element at the initial read position.

4.2. The zigzagged-list memory model: Mem₂

Centering makes a clear conceptual separation between utterances. It does not consider the last element of the previous utterance and the first element of the current utterance as following in sequence. In other words, the discourse is not seen as a continuum. Centers of each utterance are ordered in a list, and the order in these lists is given by syntax. Why is that?

As nature in general, living entities are cyclical agents. We live in a world which is motored by cycles: years, nights and days, breath, digestion, sanguine circulation. Seen at a certain granularity, nothing in this world goes on linearly, non-repetitive. Human agents fragment their sayings because they are unable to utter long stories breathless. We need to make pauses while we tell stories. If we have a reality to communicate, which for the sake of abstractness, in this paper we see as a graph of relations between some entities, there should be a natural segmenting factor which forces us to utter it in portions. These discourse chunks are the sentences (or clauses) – called utterances in Centering. But then, the production and the interpretation of the discourse should also be sensible to this segmenting necessity. In what way? What are the natural mechanisms acquired by humans to maximize their communication success, as encumbered by the natural segmentation of sayings into sentences?
In this model we will make the write pointer to be reset at the read position at the beginning of each utterance. As in the first model, it moves to the right after each insertion (see Figure 2). The movement to the right of the R-pointer in searching is identical as in the first model.

So, when processing an utterance, the behaviour of Mem₂ is as follows:

- reset the W-pointer at 0;
- while there is still an entity e to process {
  - search for e starting at the R-pointer and advancing the R-pointer one position right;
  - if e is found then delete e;
  - write e at the W-pointer;
  - advance W-pointer one position to the right;
}

5. EVALUATION

In this section we are preoccupied to validate the hypotheses formulated in section 3.

5.1. H₁: when rephrasing, how coherent are the discourses produced compared to the original?

We were interested to see if human subjects really transmit discourses with a cohesion score close to Centering-optimum. In order to verify this hypothesis, in a first experiment we have distributed to a class of students⁴ a discourse composed of 12 Romanian utterances, and we asked them to read the text, to memorize the content and then to write it down again, freely. We wanted to see how close to the original the coherence scores of these discourses are. The global Centering score of the original text was 1.6 (therefore between RET and SSH).

We received 7 discourses shorter than 12 utterances, 2 longer than 12 utterances and the rest having 12 sentences, as the original. Figure 8 presents a histogram of different Centering scores of the students’ compositions.

![Fig. 3 Histogram of discourses displaying different Centering scores](image)

⁴ The class had 32 master students in Computational Linguistics.
As can be seen on Figure 3, comparing the areas to the left and to the right of the vertical line, most of the students retransmitted the initial content of the discourse with a global Centering score lower or equal than/to the original (there were 22 students which got Centering values lower than 1.6, 3 got Centering values equal to 1.6, and 7 – higher than 1.6). We interpret this result as the tendency of human agents to produce texts which are Centering-low, rather than Centering-high. It means that humans reinterpret the knowledge they want to transmit and put it in a form which is more often easier to process than harder.

5.2. H1: how close to Centering-optimum are the texts produced by humans?

In a different set of experiments addressing H1 we wanted to see how close to Centering-optimum are the texts produced by humans. For that we took a discourse belonging to the GNOME corpus (Poesio, 2004), and generated all possible permutations of its k-prefixes (with k in the range 2 to N, N being the length of the text; a k-prefix is the sub-discourse keeping only the first k utterances of the original discourse). We counted how many of the prefixes have a global Centering score under the Centering score of the original prefix. Table 1 shows the results.

<table>
<thead>
<tr>
<th>No of uttrs (k)</th>
<th>Min Centering Value</th>
<th>Initial Centering Value</th>
<th>Max Centering Value</th>
<th>No of Permutations With Lower Than Initial Values</th>
<th>No of Permutations with Higher Than Initial Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>0.50</td>
<td>1.50</td>
<td>2.00</td>
<td>4 (66.67%)</td>
<td>2 (33.33%)</td>
</tr>
<tr>
<td>4</td>
<td>1.33</td>
<td>2.33</td>
<td>3.33</td>
<td>8 (33.33%)</td>
<td>16 (66.67%)</td>
</tr>
<tr>
<td>5</td>
<td>1.25</td>
<td>2.00</td>
<td>4.00</td>
<td>16 (13.33%)</td>
<td>104 (86.67%)</td>
</tr>
<tr>
<td>6</td>
<td>1.40</td>
<td>2.40</td>
<td>4.00</td>
<td>136 (18.89%)</td>
<td>584 (81.11%)</td>
</tr>
<tr>
<td>7</td>
<td>1.17</td>
<td>2.66</td>
<td>4.00</td>
<td>1.766 (39.79%)</td>
<td>2.672 (60.21%)</td>
</tr>
<tr>
<td>8</td>
<td>1.57</td>
<td>2.86</td>
<td>4.00</td>
<td>10.938 (28.72%)</td>
<td>27.151 (71.28%)</td>
</tr>
<tr>
<td>9</td>
<td>1.37</td>
<td>2.50</td>
<td>4.00</td>
<td>30.891 (7.85%)</td>
<td>362.396 (92.15%)</td>
</tr>
<tr>
<td>10</td>
<td>1.22</td>
<td>2.22</td>
<td>4.00</td>
<td>113.320 (3.12%)</td>
<td>3.511.640 (96.80%)</td>
</tr>
<tr>
<td>11</td>
<td>1.10</td>
<td>2.00</td>
<td>4.00</td>
<td>547.631 (1.37%)</td>
<td>39.368.693 (98.63%)</td>
</tr>
<tr>
<td>12</td>
<td>1.18</td>
<td>2.18</td>
<td>4.00</td>
<td>6.437.184 (1.34%)</td>
<td>472.564.416 (98.66%)</td>
</tr>
</tbody>
</table>

In Figure 4 we display a chart based on columns (1, 5, and 6) of Table 1. We plotted the percentage of permuted prefix-discourses having Centering scores greater than (grey area) and lower than (black area) the Centering scores of the initial prefix, for different values of k. The plots show that for the chosen text there are fewer possibilities to rearrange prefixes of discourse in a Centering-lower manner than in a Centering-higher manner. This means that, although the text and none of its prefixes (excepting for the trivial 2-prefix) is not Centering-optimum, they are, however, close to this minimum. Should these prefixes have been Centering-optimum would mean that humans were able to globally optimize the discourse.

Figure 5 shows the placement of the prefixes’ Centering scores between the possible minimum and maximum.

---

5 We used prefixes of a single text instead of many texts of the original corpus because permuting long discourses is computationally very expensive. We are aware that this experiment should be reconsidered by involving more computational power.
We were interested in two aspects relative to H2: the influence of the memory type on the memory processing costs and the influence of the memory length on the memory processing costs. We present here only findings relative to the memory type, not also to the memory length.

To verify H2, and implicitly to find the “best” memory type, we have generated randomly 10,000 graphs and plotted the summed up memory costs, as given by formula (1), for all graphs. For the first model, Mem₁, we made the writing point to vary from 0% to 100% of the momentary size of the memory. The memory size was considered infinite (in fact, equal with the length of the graph). Accordingly, the momentary size of the buffer is given by the number of elements in memory at a certain moment.

We have taken the precaution that the generated graphs be trees as well as non-trees, with the following algorithms (to attach more than one parent to a node means to leave, for a while, a node in the Open class, even following the moment when an edge towards it has been generated).
Also, because, clearly, the order of the nodes in memory influences the cost of the search, we have transmitted randomly the edges to the memory (actually in the rhythm they were generated).

In Figure 6, on the horizontal axis the position of the write pointer is displayed. The experiments with $\text{Mem}_1$ put in evidence a clear qualitative distinction between the stack (position 0) and the queue (position 100%). It shows that the stack model gives a cheaper solution on average.

On the other hand, the model $\text{Mem}_2$, acts very close to the optimum of $\text{Mem}_1$, this showing that a zigzagged list is the next to optimum solution.

5.4. H4: is it true or false?

We were interested to see whether the STACK memory model (built from the $\text{Mem}_1$ family), which was proven in the previous section to behave as $\text{OSTMem}$, really produces discourses which are close to Centering-optimum, when the agent uses a greedy strategy in choosing the next utterance. Greedy means as follows: the agent has uttered a part of the graph; at this moment the edge it grasps next to utter is one among those edges whose summed-up memory cost is minimal among those remained to be transmitted.

Proving this hypothesis is the central point of this research. In order to verify this hypothesis we have generated randomly a graph (with 10 edges), we have chosen randomly one edge as the start utterance, then at each step we have chosen the next edge applying the greedy strategy, we have verbalised that edge as “x y” if the edge linked the nodes x and y, and so on, until the whole
graph is spelled-out. Finally, we computed the global Centering score of the generated discourse. The evaluation shows a very small distance from the Centering-optimum: 95.75% of possible variants have the Centering score above the generated discourse and only 4.25% are below this score.

6. CONCLUSIONS

In this paper we have discussed experiments intended to give cautious answers to important questions in the modelling of human performance in producing and interpreting discourse, as the following: What is the memory model humans use when producing and processing discourses? Is there a best and a worst discourse, in terms of cohesion, that can be produced with a number of elementary discourse units? How are the discourses produced by humans placed in this range? Is there a metric we can use to measure the cognitive load in terms of operations performed by the memory? What triggers the selection of the next utterance? If a best memory model is identified and its processing costs are measured using the identified metric, is it that the discourses produced this way are close to human-like performances?

If the answer to the last question is true than we believe that an important step in the search for a cognitive mechanism that explains the production and interpretation of discourse has got close to be deciphered. We are caution in showing exaggerate optimism, because it only shows a possible way. Our construction represents only the if part of a demonstration: if the conditions enumerated hold then the discourse generated has that property. But, in principle, other models could give rise to the same results.

Summarising, the research methodology used in our investigation is as follows. We started by stating a set of 4 claims, called hypotheses: the first addresses the human discourse performance, while the following three – cognitive features of our discourse processing machinery. To investigate the properties of the human discourse, we used a class of master students which were asked to reproduce from memory the facts that were presented in a short story, by forming their own short compositions. Their discourses were then measured in terms of Centering transitions and compared against the global Centering score of the initial witness discourse. The experiment revealed the tendency to produce discourses easier to interpret than the original. In another experiment, we used one of the texts in the GNOME corpus (Poesio, 2004). From all discourses cut out of the original one by taking longer and longer prefixes (number of initial utterances), we automatically generated all possible permutations of utterances. Then we showed that, for all these partial discourses, fewer were easier to process than the original discourse and much more were harder to process than the original. The comparisons to express “less” and the “harder” were done in terms of Centering transitions, which we believe express adequately not only the discourse coherence but also the cognitive load. We interpreted this finding as an indication that human discourses are close to Centering-optimum, although not optimal.

Following this result, we have tried to find a model of cognition which yields a similar behaviour on an artificial agent. For that we had first to work on a memory model able to manage communication most advantageously, in terms of memory operations, since we hypothesised that the immediate memory has an important influence over discourse performance. It showed up that the STACK model is, on average, the least costly, therefore the most productive, in terms of memory operations. Close to it comes then the ZIGZAGGED-LIST model. Then, we had to prove that when the best memory model combines with a greedy-like selection strategy, the resulted discourse is characterised by a degree of coherence similar to that produced by humans. It is interesting to comment on the near-optimum Centering character of the human discourse cognition. It is clear that this is the result of the greedy, incremental, optimisation. In artificial
intelligence it is common knowledge that local optimisation amounts to local optima, and only by chance, to global optima. This is what happens with the human discourse. Humans are not capable to globally optimise the discourse comprehension. Instead, they have the ability to do this at each step. As a result, they get a discourse which is only close to Centering-optimum, although not optimum.

Once again, these considerations take into account only reality graphs unconstrained to sequentiality, therefore where there are no narration parts necessitating a certain time ordering. Such constraints could give rise to transitions which follow no more the near-Centering-optimum pattern.

Still, there are a number of weaknesses in our approach, which we will have to fix in a further investigation. First, the length of the memory model has to receive attention. Secondly, when implementing the “greedy” approach of selecting the next edge to be uttered, an important aspect has been neglected: the computational effort to find the cheapest edge among those remained to be communicated. In fact, the greediness behaviour was supposed to express an economist principle: the agents’ brain prefers to work less than more. So it grasps the more economical edge (from the point of view of memory operations) at every step. But there is no label indicating “the most economical” edge at a certain moment during generation. Finding it involves a computation by itself. And this computation is neglected in our experiments till now. Thirdly, it is evident that the experiments aiming to measure human discourse coherence are still not very comprehensive, because they are limited to only a class of students and one discourse from the GNOME corpus. More data is necessary here and this is an important objective for the near future.

If the findings claimed in this paper are true, then this investigation puts in evidence a missing link between immediate memory models intended to offer an interpretation of the human capacity to produce/interpret discourses (Sidner, 1983; Walker, 1989), empirical models of discourse attention and structure (Grosz & Sidner, 1986), and models investigating the coherence of discourse (Grosz et al., 1983; 1995). We now begin to understand the cognitive mechanisms of humans which trigger their ability to compose and process discourses the way they do. It is as simple as that: an immediate memory apparatus adapted to process knowledge efficiently (which cannot appear otherwise but as a result of evolution) and a minimalist behaviour (perhaps, also innate). These two ingredients are enough to boil comprehensive discourses (provided the reality to express is comprehensively represented itself). So, the most important conclusion is that the ability to produce discourse by agents is not the result of a learning process acquired during the agents’ life. It is wired deep inside their minds.

ACKNOWLEDGEMENTS

This research is supported by the EC FP7 project „ALEAR – Artificial Language Evolution on Autonomous Robots”.

REFERENCES