Grounding Coherence Properties of Discourse

Dan Cristea¹,² and Adrian Iftene¹

¹ Alexandru Ioan Cuza University of Iasi, Faculty of Computer Science
16, Berthelot St., 700486 Iasi, Romania
{dcristea@info.uaic.ro, adiftene@info.uaic.ro}

² Romanian Academy, Institute for Theoretical Computer Science, the Iasi branch
2, Gh. Asachi St., 700481, Iasi, Romania

Abstract. In this paper we investigate two fundamental issues related to the production of coherent discourse by intelligent agents: a cohesion property and a fluency property. The cohesion aspects of discourse production relate to the use of pronominal anaphora: whether and in what conditions intelligent agents could acquire pronouns as means to express recently mentioned entities? We show that the acquisition of pronouns in the vocabulary of an agent is conditioned by the existence of a memory channel recording the object previously in focus. The approach follows an evolutionary paradigm of language acquisition. Experiments show that pronouns spontaneously appear in the vocabulary of a community of 10 agents dialogging on a static scene and that, generally, the use of pronouns enhance the communication success. The processing load experiments address the fluency of discourse, measured in terms of Centering transitions. Contrary to previous findings, this side of discourse coherence seems to be grounded in an innate cognitive mechanism, which is driven by an economicity principle. We prove experimentally that the a model of immediate memory which resembles the stack data structure is optimal in terms of access costs and that, put at the base of the production of discourse, it leads to discourses which have similar fluency patterns as those produced by humans.

Keywords: Coherence, Pronominal Anaphora, Fluency of Discourse

1 Introduction

In this paper we study two fundamental properties of discourse coherence, cohesion and fluency, from a grounding perspective. The main question we want to answer is: what are the basic cognitive mechanisms that allow agents to develop coherence properties of discourse similar to those of human beings? In particular, we are interested in the acquisition of pronominal anaphora by intelligent agents in locally situated dialogues and in production of fluent discourses.

The research represents a step in the attempt to decipher the acquisition of language in communities of humans. Models of language acquisition ([2], [21], [24]) hypothesis that language users gradually build their language skills in order to optimise their communicative success and expressiveness, as triggered by the
need to raise the communication success and to reduce the cognitive effort needed for semantic interpretation.

The Talking Heads experiments ([20], [22], [23]) have already proven that a shared lexicon can be developed inside a community of agents which are motivated to communicate. The participants in the experiments are intelligent humanoid robots, able to move, see and interpret the reality around them (scenes of interrelated objects), as well as to point to specific objects. They are programmed to play language guessing games in which a speaker and a hearer, members of a community of agents, should acquire a common understanding on a situation which is visually shared by both participants in the dialogue. After tens of thousands of such games, played in pairs by members of the community, a vocabulary that give names to concepts which are needed to differentiate the properties of objects spontaneously arises. The vocabulary is shared by the majority of agents and is relatively stable at perturbing influences caused by population growth, decline, or mixing with other smaller groups. The next step deals with the acquisition of grammar. In [26] it is proved that rudiments of grammar can be developed as a result of interactions. Other studies developed in the ALEAR project showed that the capacity of agents of inventing grammatical markers for indicating event structures, the formation of semantic roles [19], the combination of markers into larger argument structure constructions through pattern formation [27], [28].

If perception, vocabulary, elements of grammar, semantic properties of language as space and time, and conceptualisations can be grounded in linguistic games, then what happens beyond the barrier of the sentence? At a certain level of language development, humans were able to interact in coherent dialogues and to produce long discourses. In mentally sane agents, dialogues and monologues had the goal to transmit information that are correctly received and deciphered by partners. Coherence of discourse is a major requirement for a successful communication.

To investigate the grounding of coherence in human produced dialogues and discourses, to see to what extend it is learnable, the first issue to find an answer is: how to measure coherence? This is not a simple question, because there are so many ways to look at coherence. To give just one example, remember the famous Chomsky’s utterance: Colourless green ideas sleep furiously which is given in schools as an example of a sentence that has a perfect syntax but which is incoherent. Still, there are people saying that this sentence can be considered perfectly coherent\(^1\). So, meaning and coherence is a matter of personal interpretation, has a lot to do with the context in which the utterance is placed and, in many cases, involves metaphoric uses of word senses. Seen as such, coherence is hardly measurable, is extremely exposed to subjective interpretation, and therefore impossible to be quantified. Leaving so many shadows behind, our rabbit is

---

\(^1\) Ideas, of course, have no colours. Unless they belong to a member of the ecologist party, in which case one can say that they are green. Seen in these context, ideas are simultaneously colourless and green. And if they boost during an agitated sleep, one can say, metaphorically, that they sleep furiously.
hard to chase. As recognised by many scholars, coherence implies both cohesion and fluency. Cohesion means links that tie together units of discourse, and one of the most important is the use of pronouns [11]. Fluency, on the other hand, means the easiness of deciphering utterances in their sequence and has a lot to do with the level of inference load put to work by the hearer.

In the first part of the research described in this paper we investigate whether the acquisition of pronominal anaphora inside a community of intelligent agents can be empirically proven by following an evolutionary approach and, if so, to point out which are the minimal cognitive requirements that allow the use of pronominal anaphors, how long would it take for a pronoun to appear in the vocabulary of a community, and what is the communication gain if pronouns are used. In the theory of discourse, the repeated occurrence of the same entity is one of the manifestations of cohesion, which, itself, is an aspect of discourse coherence. Among other things, a surface sign of the awareness the talking agent has that an entity is referred again is the use of a pronoun referring an entity named before by a noun, proper or common. We organized a number of game-based experiments, during which the agents were expected to achieve human-like cohesion performance with respect to the use of pronouns. A number of settings of increased complexity should show when the use of pronouns becomes a necessity and when it really enhances the understanding.

Influenced by the success of the evolutionary paradigm applied to different cognitive aspects of language, we believed that the same general framework should work also for grounding the fluency properties of discourse. The research on this aspect of coherence is reported in the second part of this paper. It describes a model of coherence in discourse which proves that the production of unconstrained discourse should not necessarily be associated with a learning process developed over a long series of interactions inside a community of intelligent agents. Instead, a pattern of fluency in discourse, similar to that characterising the discourses produced spontaneously by human subjects that are in need to communicate a conceptual representation, can be obtained by an intelligent agent possessing a very simple type of cognitive mechanism, a sort of internal short-term memory which emerges from an economy of effort principle.

In the initial stages of our research, we were looking forward to apply an investigation methodology that, in its basic aspects, would have had to replicate traits that have been used in the ALEAR project and before to prove the evolution of different aspects of language, as mentioned above. As we advanced in our modelling, we quickly understood that an important clue in the comprehension of a heard discourse or the production of a comprehensible discourse is the awareness the agent should have that an entity has already been mentioned. The necessity for a memory channel arose also while studying the acquisition of pronouns in language, so no surprise till now. We had to introduce a model of memory, which is responsible for recognising efficiently entities recently mentioned. In accordance also with many scholars (see, for instance [31]), this model of immediate memory should be as simple as a buffer in which recently mentioned entities are stored. The awareness that an entity has already been mentioned is
gained each time the mention of that entity is found in this buffer. We have, therefore, started by simplifying and parameterising a buffer model with the initial aim to prepare the ground on which to organise experiments that would prove that language fluency is a learnable process. Since recuperation of mentions from this memory would have had to be extremely quick, we have hypothesised that there should be one model which behaves better than others with respect to the cost of access. So, we proposed a model of cost that should be placed at the base of the experiments to prove this hypothesis. At that moment we hoped that, once the optimum memory model will be found, it could be incorporated in the cognitive mechanism of agents and then let them develop, by themselves, during repeated dialogues, the ability to produce coherent discourses.

At this point we had two targets ahead: to find the optimum short-term memory model, if there is one, and to find a methodology that would drive us towards a trainable ability to produce coherent discourses. The first target was reached soon, by experimenting with different types of knowledge graphs to be transmitted, different orders of transmitting them and different parameters of the memory model. We find out that, among a whole range of possibilities, a model that replicates (up to a certain point) the dynamics of a stack is the cheapest in terms of memory access costs. Not very surprisingly, since the stack model has been suggested before in relation with the processing of discourse [10].

In our model, the stack is used differently than in Grosz and Sidner’s attentional state model: it stores only mentions of entities, not complex states, this way resembling more Walker’s very simple cash model.

Next, we had to move towards finding a measure for fluency that should be used to control the experiments. A suggestion for this can be found in Centering [8], [1], [9]. Centering is known as a theory of local discourse coherence, not touching the issue of coherence of utterances. In Centering, only the entities (called centers) and their syntactic positions around the main verb (as subject, direct object or others) matter. In a simplified version, the type of transitions between consecutive utterances is dictated only by the order of centers mentions in the utterance. So, we did two things: because the meaning of sentences is ignored and we were interested only in the position of entities in an utterance, we have adopted a very schematic shape of an utterance, of the form Subject Verb Object. Then, in order to measure coherence based on Centering transitions, we adopted a range of 5 scores for transitions, from the easiest to process (most fluent, smoothest = 0) to most difficult to process (involving the most inference load = 4), a reversed scale than the one used in [5]. This gave us a measure of coherence.

But in order to find the methodology that would lead up towards developing in agents abilities to produce discourses as fluent as the ones produced by humans, we had to know how much fluent are human produced discourses. We formulated a hypothesis and looked for ways to prove it. We did three different tests in this direction, all three confirming its validity.

A discourse represents a sequencing of a cognitive representational space. It is clear that the ability to first represent in memory what to say and only af-
terwards to speak out this representation\(^2\) is a recognised sign of intelligence. As we are not concerned here with modelling gains in the level of intelligence of agents, this scenario had to be imposed. This yields the following general scheme of generating a discourse by an agent: a) the speaker agent builds in memory (another section than the short-term memory mentioned above) a conceptualisation of a situation; the representation of this conceptualisation can be seen as a graph (see, for instance, the dependency graphs of [17], or the semantic representations of [18]); b) the speaker then utters this graph by adopting a certain order in traversing it. The traversing order should give rise to more or less fluent discourses, and this order could be the result of an evolutionary process; c) the hearer agent uses his short-term memory to recognise entities already mentioned, this way building in his own memory space a conceptual representation which, ideally, should be identical with that of the speaker. As can be seen, at that moment we did not know how the short-term memory could trigger the order of utterances in the produced discourse. We thought that the order should be the result of either: a) a learning process or b) a kind of economy principle of the sort "choose to utter next whatever is easiest to grasp at each moment".

At this point we wrote algorithms for automatic generation of graphs and started to experiment on different shapes of graphs with different traversing strategies. We realised shortly that there are 4 strategies that are interesting to consider: breadth-first (BFS), depth-first (DFS), a kind of greedy approach (GREEDY), and a random selection (RANDOM). On the other hand, two predominant patterns of Centering loads were noticed on the histograms plotted after subjects generated short stories. After examination we realised that the two hums correspond to BFS and DFS orderings. In terms of costs of traversing the graph, they should have been more economical, compared to both GREEDY and RANDOM. So, at this point, we started to believe that subjects prefer to utter at each step the handiest to grasp, or the "laziest" selections, which would make superfluous the need for learning.

After measuring the Centering scores of the generated discourses following all 4 orders, it became clear that the Centering patterns of the DFS/BFS-generated discourses resemble very much the human-generated discourses. This discovery reinforced our belief that no evolution was necessary: coherence in the Centering sense yields naturally from the application of an economy principle at each step in selecting the next edge of the graph. This is so natural in human beings [33], [13], [30], that we did not believe anymore that the evolution along a series of interactions in a community of users of the language gave rise to the ability to talk coherently.

There remained only one question to be answered: in what way the optimum model of immediate memory that we have found correlates with the lazy approach and the human-like Centering? The link came from correlating observations on the histograms of Centering loads measured in controlled discourses produced by human subjects, and the notice that a stack model of memory supporting decisions of what to grasp next gives rise to a depth-first ordering in the speaker. In

\(^2\) as opposed to the other way round.
two different experiments the subjects investigated showed balanced preference for either DFS or BFS ordering in their generated discourses. Yet another experiment, developed on a corpus of short texts, revealed that human discourses are not Centering-optimal but quite close to it. On the other hand, the most economical memory model is the STACK, which yields a DFS ordering in traversing graphs. This correlation closes the circle: both DFS and BFS produce coherent discourses (as compared to RANDOM), but DFS is memory-cheaper, although it gives rise to a little bit less coherent discourses than BFS ordering. This is the pattern that applies also to most of the unconstrained human produced discourses.

Of course, languages have evolved to manage the complexity of the world around us but it would be unrealistic to believe that the human beings first gained a sophisticated cognitive apparatus and only after they started to invent the language. It is known that the evolution of the human cognition is closely correlated with the acquisition of language. For reasons of building the information model, however, we are forced to simplify this intertwined evolution. In our investigations we consider a given cognitive machinery which should be sufficient to enable the emergence of language. By our research we only delimit the minimum requirements for this cognitive platform and suggest some details of interactions that would lead to the present day performance.

The paper is organised as follows. In section 2 we describe the grounding of one of the main cohesion features the use of pronouns. The simple rules of a linguistic game are explained first. Then we ground the use of pronouns on a minimal cognitive property. Then we introduce some scene settings displaying an increasing difficulty of comprehension. The dialogue experiments conducted in these settings and the results are then presented and discussed. Section 3 is dedicated to the fluency aspect of the discourse. We start by giving a background of the research, then we define a set of hypothesis that lead step by step to the proposal of a grounding of one important aspect of the discourse coherence the fluency. The following subsections describe tests intended to support these hypothesis. A final argumentation linking both aspects of discourse is offered in section 4.

2 Grounding cohesion

2.1 The experiments framework

We organized a number of game-based experiments, during which the agents were expected to achieve discourse-level performance. For this we first developed a simple and parametric Java framework in which games can be easily defined and which allows for any number of experiments. This also offers support for the description of the closed worlds, for the description of agents’ lexical memories and of their dialogues.

In our experiments, a world has objects with properties (shape, colour, position, etc.), and a scene is a world with a specified configuration of objects.
There are two ways to generate a scene: by manually describing all the objects populating it, as well as all their properties, or by generating it randomly. The number of objects generated in a random scene can be set within specified lower and upper bounds.

Agents perceive the scene through a number of perception channels, which can be changed at will. Each channel targets a specific property of an object. Agents can perceive the values on each channel with their own granularities. This way some agents can have a greater acuity on a given perception channel than others.

A game is a specified protocol of interaction between two agents. An example of such a protocol is the "guessing game" [22], where one agent chooses an object in the scene, generates an utterance that describes it, and a second agent must guess the object described by the utterance (without knowing which was the chosen object). If the object is correctly indicated, the trust of both agents in the proper usage of the words describing the conceptualisation of the object increases. If the object is not guessed, a repairing strategy is applied: the speaker points to the object he chose and the trust it has in the words used to name it decreases, while the hearer either learns the words or associates a greater level of trust for this form to describe the conceptualisation of the object. After a large number of games of this kind, played in pairs by agents, the community shares a common vocabulary and associates words to concepts with a low level of ambiguity.

2.2 The inception of pronouns

Anaphora represents the relationship between a term (called "anaphor") and another one (called "antecedent"), when the interpretation of the anaphor is in a certain way determined by the interpretation of the antecedent (Lust, 1986). When the anaphor refers the same entity as the antecedent, we say that the anaphor and the antecedent are coreferential. When the surface realisation of the anaphor is that of a pronoun, the coreference relation also fulfills other functions:

- it enables conciseness by avoiding direct repetitions of a previous expression, thus contributing to the economy of expression a central principle in the communication between intelligent agents;
- it maintains the attention focused on a central entity, by referring it with extremely economical evoking means. Indeed only entities which already have a central position in the attention could be referred by pronouns and, once referred, their central position is further emphasised.

Anaphora, as a discourse phenomenon, presupposes non-trivial cognitive capacities. The one we are concerned about in this paper is the capacity of memorising the element in focus. This capacity is so central and elementary that we decided to consider it as being provided by a dedicated "perception" channel actually a memory channel. Indeed, both cognitive aspects of distinguishing between right
and left (to give a common example of perception) and of remembering that a
certain object was in focus recently involve primitive cognitive functions. The
lack of memory would make a dialogue impossible, the same way as the lack
of spatial perception abilities would make the recognition of spatial relations
impossible.

The type of games we are interested in when modelling anaphoric phenomena
are multi-turn, such that one entity, which has already been in focus previously,
could be referred again later. In this study, we are targeting only pronominal
anaphors. If we want an agent to develop the ability of using pronouns, the
dialogue should include a sequence of utterances in which an entity is mentioned
more than once.

The focusing memory is modelled through a channel called previous-focus,
with two values [true, false]. Excepting for the first utterance of the dialogue,
when there is no previously focused entity, on each subsequent utterance there is
one entity (object) which is remembered as being the focus of the previous game.
As such, each object in the scene has a value of false on the previous-focus
channel, except for the object which has been in focus previously, and whose
corresponding value on this channel is true.

2.3 The settings

The problem we are concerned with is when and why intelligent agents would
develop linguistic abilities for using anaphoric means in communication and how
anaphora could complete a conceptualisation.

It is clear that an agent has at least two reasons for choosing to name an
object by a pronoun:

- because it uses less words (for instance, it instead of the left circle);
- because this way the OLD (therefore, the entity previously in focus) is ex-
plicitly signaled, maintaining it there;

while it has also at least one reason why not using it:

- because it could introduce an ambiguity.

The use of pronouns should emerge naturally during the experiments. It
should not be enforced (given programmatically). To model the acquisition of
pronominal anaphora, four different settings have been used, which we believe
present an ascending degree of complexity. All are based on the paradigm of a
two-turn game. What makes the difference between these settings are changes
in the scene of the second turn as compared to the first, and the chosen focus.

In the first setting (Figure 1) both games are played in the same scene and
the co-speaker will focus in the second game the same object as the speaker
focussed in the first.

Turn 1: A names obj₁ by low left
Turn 2: B names obj₁ by that

In the second setting (Figure 2) new objects are introduced in the scene of
the second game, while the focus remains unchanged.
In the third setting (Figure 3), the objects in the second game’s scene are shuffled (their spatial properties like horizontal and vertical position are randomly changed). The co-speaker will keep the focus on the same object, although it might have changed its position.

Turn 1: A names obj₁ by *left*
Turn 2: B names obj₁ by *that*

In the fourth setting (see Figure 4), the scene of the second game is again a shuffled version of the scene in the first game and the focus can no longer be identified by any of the attributes used in the first game. In this particular scene, the agents do not distinguish colours or shapes, so the objects can be identified only through position and anaphoric means.

Turn 1: A names obj₁ by *low left*
Turn 2: B names obj₁ by *that*
All experiments have been run with the following parameters: \#agents = 10; \#multi-games = 5000; \#objects = 8 (between 8 and 10 in setting 2); channels: "hpos" (horizontal position), "vpos" (vertical position), "color", "shape", "previous-focus"; channels granularity: between 2 to 4.

As we see, in every multi-game the focus is maintained on the same object in both turns. Let us notice that it makes no difference who is the speaker in the second game. Only for the sake of displaying a dialogue we considered the second utterance as produced by the co-speaker.

2.4 The results

Figures 5-9 display success rates (averaged over the last 100 multi-games) along an experiment that lasted 5000 multi-games, in different configurations of objects and settings, as follows. Figures 5 and 6 show the success rate in setting 1, with scenes counting 5 and, respectively, 8 objects, while Figures 7-9 display the success rate in settings 2-4 when there are 8 objects in the scene. In all experiments, only multi-games which reported success after the first turn have been retained, as we were interested here only in the acquisition of pronouns (mentioned only in the second turn in each multi-game) and not in a stabilisation of a lexicon in general. So, if at the end of the first turn, agent B does not recognise the object indicated by agent A, the game is stopped. In all figures, the (darker colour) line above reports the percent of general success rate (after the second turn), while the (lighter colour) line below reports the success rate that is due to the use of pronouns.

The abruptly growing shapes of the lines above, in all four settings, show that, very quickly, the agents acquire a common understanding of the objects in the scene (to be more precise over the object in the focus). Only the last two settings show a more shaky shape, due to the increased complexity in the identification of the focus. In general, after 300-400 games, the success rate stabilises to 100%. However, as the (green) lower lines show, in fewer cases this common understanding is due to the use of pronouns. This should not be interpreted as an indication of the fact that the use of pronouns reduces the success rate, but that in some cases other referential expressions than pronouns are also used to identify an object which has been previously in focus (for example, in setting 3, they can use up instead of that).
However, if we compare Figures 5 and 6, we see that when the number of objects is larger, the need to use pronouns also goes up. This is clearly due to the fact that a greater agglomeration of objects in the scene makes their identification based on other features than being recently in focus more ambiguous. Indeed, the agents chose randomly among the shortest known categorisations which one to use for identifying the object in focus from those able to individualise it unambiguously. If all possible utterances have the same confidence (the confidence of a linguistic expression is calculated as the mean value of the confidence of the words used to utter the corresponding categorisation) one of them is chosen randomly. However, when more conceptualisations are at parity from the point of view of confidence, the shortest form is chosen. This is the only bonus that favours the economy of expression, therefore the use of pronouns. The graphs show that when there are more objects in the scene, being recently in focus remains the conceptual feature with the highest confidence.

An interesting thing is revealed by the graph in Figure 9: the two lines representing global success rate and pronoun-based success rate are practically identical. This means that when the situation is very complex, in almost all cases the agents prefer to use a pronoun to identify an already mentioned object.
Finally, we were interested to see what happens when we impose the use of pronouns. Figure 10 shows two lines, both drawn in setting 1: the lower (green) line represents the normal use of pronouns in the case of success in the second turn, while the upper (yellow) line represents the success rate in the second turn when we enforced the use of pronouns. The particular conditions of this experiment make superfluous the need for more than one channel (in this case "previous-focus") to identify the focus.

2.5 Discussion

In this section we have proven that the acquisition of pronouns in language can follow an evolutionist pattern, therefore pronouns can appear in language as a natural, spontaneous, process, driven by the necessity of the agents to acquire common understanding over a situation. The study does not show, however, that this is the only way in which pronouns could have appeared in natural languages. It simply shows a possibility.

We have used a paradigm in which a community of agents communicate. A common agreement over a focussed object in a scene is rewarded by an enhancement of the trust in both the conceptualisation used and the linguistics means...
to express it. After a number of experiments, a certain lexicon is acquired by the community.

The model we used has considered the existence of a memory channel remembering the object recently in focus. When such a channel is open, the identification of an object already mentioned, and which should be mentioned again, can be made quicker and with less ambiguity because it implies less categorisation. The linguistic expression of this economic categorisation is the pronoun. The experiments show a clear tendency of the agents to enhance their linguistic ability to use pronouns in more and more complex contexts.

When the number of objects in the scene increases, the chance that the "previous-focus" channel is the only channel that uniquely identifies an object is very high and therefore the use of pronoun becomes dominant.

3 Grounding fluency in discourse

3.1 Background

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 understand by the reader (as it occurs, 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 section 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 [3], as compared to those supporting an acquired view over features of language [15] (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 cognitive machinery which enables 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.

According to Centering, to each utterance in a discourse corresponds a list of forward-looking centers, which are semantic entities mentioned. This list, noted usually \( C_f \), is ordered according to syntactic criteria, not the same in all languages (see, for instance, [14] for Japanese, [7] for Italian, or [25] 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 \( C_p(U_n) \) is the most prominent center of \( C_f(U_n) \). The transitions between pairs of adjacent utterances define degrees of easiness of processing the sequence of utterances in monologues, or turn takings in dialogues, therefore degrees of fluency. We will refer in this section to the Centering second rule, which states that there are four types of 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 \( C_b \)'s and that between the \( C_b \) and the \( C_p \) of an utterance. When there is no intersection between the \( C_f \) list of the previous utterances and that of the current utterance, it means that the current utterance lacks a \( C_b \) (we will call this No \( C_b \), and note it as NOC).

In our experiments, we used the 4 Centering transitions and NOC (to whom we assigned values from 0 to 4) 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 \( (N - 1) \). 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 (describing a scene or situation).

Theories on discourse production and interpretation, as for instance [10], [31], [12], often bring into discussion a model of immediate memory (also called cash 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 is supposed to be slower. As mentioned already above in this paper, identifying elements in memory (either short or long term) is essential for a coherent communication.

3.2 Hypothesis

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

H1. On the "cheapest" immediate memory model

There should be a built-in model of immediate memory which minimises memory access costs.

Among all possible models of immediate memory, there should be one, which is the "cheapest" 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 that 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, and consequently a greater effort to bring information from long-term memory;
– on the other hand, a very long memory triggers a longer decision time to detect the first mentions (because an entity which is not more in the immediate memory should yield first a memory search fail and only after a trial is made for regaining it from the long-term memory).

H2. On the near-optimum coherence of the human discourse

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

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 above.
H3. On an economic (lazy) strategy in choosing the focus

In a space of yet-to-be-said elementary discourse units, an agent is tempted to select those entities to be uttered next which are 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 at an immediate minimisation strategy, a sort of greedy approach.

H4. On the global coherence effect of applying immediate laziness

A persistent application of lazy selections generates close to Centering-optimum discourses.

This hypothesis says that the human performance to produce coherent discourses is rooted in built-in cognitive mechanisms: the principle of least effort applied consistently at the moment of the selection leads naturally to a discourse which is not the most possibly coherent one, but not far from it.

H5. On the correlation between the memory model and the human-like performance in direct communication

The human-like performance is producing/interpreting discourse is a direct result of using a low-cost short-term memory model.

This hypothesis explains that the human under-optimum coherence behaviour is rooted in the immediate memory model. The memory model dictates what entity on which there are still pending (unuttered) relations to be considered next, and this materializes in a strategy of choosing the next entity to utter, which, applied consistently, is finally responsible for a human-like coherence performance. In all, the whole human performance in producing and interpreting discourses are manifestations of the same economy principle in language, which has often been recognised [32], [29], [33], [4], [13], [30].

3.3 Validating hypothesis H1

Since our intention is to find one optimal model in terms of memory costs, in [6] we considered the following basic memory operations: read (with constant cost $C_1$), delete (with constant cost $C_2$) and write (with constant cost $C_3$). The cost of the search should be equal to the number of reads until the element is found. In addition to that, we consider also a Penalty when the element is not found.

This makes the total cost of an utterance "x y"\textsuperscript{3} be expressed as:

\footnote{For reasons of simplification, we will consider that a discourse is made of a sequence of elementary utterances, of the type \texttt{<subject> <predicate> <object>}. Since we measure fluency following a Centering-inspired metric, 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, which simplify an utterance to a pair of two referential expressions.}
\[ c = \text{cost}(\text{search}(x, M)) \times C_1 + \text{if}(\text{exists}(x, M), C_2, \text{Penalty}) + C_3 + \\
+ \text{cost}(\text{search}(y, M)) \times C_1 + \text{if}(\text{exists}(y, M), C_2, \text{Penalty}) + C_3 \quad (1) \]

In the first model, Mem$_1$, the write position is fixed at a certain relative, $d_w$, distance between the left extreme and the right extreme of the list (Figure 11). 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.

![Fig. 11. The memory model Mem$_1$: buffer with fixed write position](image)

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. As can be seen, Mem$_1$ behaves like a STACK when $d_w = 0$ and like a QUEUE when $d_w = 1$.

In the second model, Mem$_2$, the zigzagged-list model, we will make the write pointer to be reset at the read position at the beginning of each utterance. After each insertion it moves right one position.

As we will show below, the results showed that the most efficient model is the one resembling a STACK (therefore when the write position coincides with the read position). This is in accordance with other researchers’ suppositions, as [31], for instance, reinforcing the importance of cheaply accessing recently mentioned entities. The ZIGZAGGED-LIST has a performance very similar to the STACK model.

We were interested in two aspects relative to H1: the influence of the memory type on the processing costs and the influence of the memory length on the processing costs, when we consider penalties for cases when an entity is not in a short memory.

We considered that the goal of the saying can be segmented in smaller parts and each part can be abstracted as a graph. We concentrated only on the transmission of such a graph (a component), leaving apart the assembling of more
such components 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).

To verify H1 (memory type) we have generated randomly 10,000 graphs and plotted the summed up costs, according to formula (1), for all 10,000 graphs, when the position of the writing point varies from 0 to 1 respective to the momentary size of the memory. In a first experiment the memory size was considered infinite (in fact, equal with the length of the graph) and in another experiment it was limited. Penalty was also varied on a scale from 0 to 40.

In all these plots the rising shape from STACK to QUEUE, as well as the near-STACK behaviour of the ZIGZAGGED model (that appear in Figure 12), are maintained. The figure displays the summed up memory for two graph traversing strategies: Depth-First Search (DFS) and Breadth-First Search (BFS).

![Fig. 12. Memory costs for BFS and DFS traversing strategies](image)

Interesting enough, Figure 12 also shows that the DFS strategy of traversing the graph gets better with the STACK model than with the QUEUE model, while the BFS strategy overrides the DFS strategy when the memory model is the QUEUE. Clearly, this happens because the DFS ordering is guarantied by a stack data structure, while the BFS ordering by a queue data structure.

### 3.4 Validating hypothesis H2

The scenarios described in this section are aimed at measuring the coherence of discourses produced by human agents in terms of Centering score, this way intending to see to what extend it is true that human subjects produce discourses that have a coherence score close to Centering-optimum.
If a discourse is seen as a sequence of utterances, each communicating a certain piece of elementary knowledge, Centering criteria of processing load depends to a large extend on the order these elementary propositions are uttered. We were therefore cautious to minimise the intrinsic constrains that would influence the order of utterances in the discourses under investigation. If this would be true, than the order the utterances produced would have been dictated solely by the processing load, hence fluency.

3.4.1 The processing load of human-produced texts

In our first experiments, described below, we tried to remove any "time sequence" constrains. The subjects\(^4\) received a picture that had to be described with any details (scenario 1) or a fixed number of details (scenario 2).

**The first scenario:** the subjects received the image in Figure 12 and they were asked to describe it using short sentences.

Fig. 13. Test image. From http://www.animationartgallery.com/WBL/WBLLB.html

The next plot (Figure 14) presents the Centering scores (the vertical axis) for the 104 students (considered on the horizontal axis) participating in the first scenario. A point of coordinates \((x, y)\) in the plot of Figure 14 should be read as follows: exactly \(x\) students have produced discourses having Centering scores below the value \(y\). The average score, computed over all discourses produced, was 1.31 (close to the RETAINING value, which is 1), and most students (65) have transmitted text with the Centering value under this value.

\(^4\) Our subjects are a class of students in Computer Science, at the Alexandru Ioan Cuza University of Iasi, in their second year. The total number of students involved was 112, out of which 104 participated in the first scenario and 108 in the second scenario.
In the next histogram (Figure 15), on the horizontal axis are the possible Centering values of texts produced by students and on the vertical axis the number of students who have passed a certain Centering value for a discourse. A point of coordinates \((x, y)\) in this histogram should be read as follows: exactly \(x\) students have produced discourses having Centering scores equal with the value \(y\). As can be seen, most of the subjects produced texts having Centering values close to 1 (RET) (there were 16 students which produced texts with Centering score 0.8, 15 students with score 1, 16 with 1.2, and 14 with score 1.4).

We interpret this result as the tendency of subjects to produce texts which are Centering-low, rather than Centering-high. It shows that humans express the knowledge they want to transmit in a form which is easy to process.

Two preferred strategies to scan the image were identified in the students texts. One was to start in an area of the image (a character, the top-left corner of the image, the bottom right corner, etc.), to verbalize all the information related to that area, then to go to a neighbouring zone and repeat the process. This mainly gives rise to a breadth-first search order (BFS). The other one was to start
in a character, to verbalize one predication that relates that character to another one, to move on this other character and so on. This strategy mainly reproduces a depth-first search order (DFS). Both orderings tend to produce highly linked statements about the scene, this way decreasing the Centering global score. This suggests that students split an image before transmitting it, being careful to understand and transmit the information related to one area/character and afterwards extending the observation in the neighbouring, in the detriment of jumping in disorder from one corner to another. They build sub-graphs focussed on areas or characters and, after exhausting the information there, they migrate beyond of these sub-graphs in the neighbouring zones. In particular, we noticed that the two humps on the histogram can be put in correlation with the BFS ordering (Centering values between 0.8 and 1.4, aprox. 61 students), and DFS (Centering values between 1.6 and 2, aprox. 25 students).

The second scenario: we have distributed to the same group of students a knowledge graph (Figure 16), and we asked them to utter the knowledge graph by producing short sentences.

Short sentences have the form <subject> <predicate> <object>, where <subject> and <object> are names of vertexes, and <predicate> is an edge. In the knowledge graph, any edge (oriented) is accompanied by a reverse edge. When uttering the link between two nodes, they had to use only one of the two edges linking them. For example, if in the above example, between nodes X and Y exists predicat_1 (from X to Y) and predicat_2 (from Y to X), the student describes the link between X and Y, either "X predicat_1 Y" or "Y predicat_2 X". As a result, because we have only 5 double bonds between vertexes, the resulted text must have exactly 5 sentences.

As in the first scenario, a point of coordinates (x, y) in the plot of Figure 17 should be read as follows: exactly x students have produced discourses having Centering scores below the value y. As can be seen in the above figure, the average value is 1.48 (about in the middle range from RETAINING to SMOOTH SHIFT), but more students (62) produced Centering values of transmitted texts above this value rather then below it. In this case the range of possible values is smaller than in the first scenario (the minimum Centering value was 0.6 and the maximum Centering value was 2.6, while in the first experiment the values covered the range from 0.3 to 3.23).

In Figure 18, an (x, y) point represents, respectively, a Centering value, and the number of students whose transmitted discourse had the Centering value x. As we can see, most of them transmitted texts with centering scores close to 1.6 (we had 37 students with centering score 1.6, and 17 students with score 1.8).

Again, analysing the discourses, we noticed that the two humps on the histogram can be put in correlation with the BFS ordering (Centering value 1.2, 20 students), and DFS (Centering value 1.6, 37 students). In the previous case study, when the subjects had no knowledge graph under eyes but had to scroll an image instead, they preferred BFS. Now, when they received the knowledge graph and no image, more subjects preferred the DFS strategy of traversing the graph.
3.4.2 Texts produced by humans are close to Centering-optimum

In this section we try to see how close are written texts from Centering-optimum. In our tests we have used texts belonging to the GNOME corpus [16]. The GNOME corpus includes 5 files ("Dermovate Cream", "Estracombi TTS", "Texts from the Getty Museum web site", "Roman Jewellery at The Potteries Museum’ Art Gallery”, "Jewellery moves, NMS Publishing Limited, National Museums of Scotland, Edinburgh, pages 10-15") annotated with everything we need (named entities NE, and co-references Ante) that allows the automatic calculation of the Centering transitions and, hence, scores. The Centering-optimum value of a text represents that permutation of the utterances that amounts to a minimum Centering score. When texts are large (hundreds of sentences in the GNOME corpus), generating all possible permutations is practically impossible (time is exponential with the input length). So, instead, we considered only the prefixes of length 12 (12 utterances extracted from the beginning of the text) of all texts, permuted them, and counted how many of them have Centering scores below the original prefix.

The first finding of this test is that for a chosen text, there are fewer possibilities to rearrange prefixes of discourses in a Centering-lower manner (black
area) than in a Centering-higher manner (grey area). This means that, although
the text is not Centering-optimum, it is, however, close to this minimum. Should
the text have been Centering-optimum would have meant that humans are able
to globally optimize the content. This is not the case, but the continuous lo-
cal optimization produces discourses close to a minimum. Moreover, as can be
noticed in the going-down shapes of black and white regions of Figure 19, the
longer the discourse is, the less possibilities are to express in a more coherent
form a certain pull of knowledge. This finding makes us believe that hypothesis
H2 is true.

3.5 Validating hypothesis H3 and H4

In this section we will concentrate on the different possible strategies of making
the immediate choice in producing a discourse and on measuring the global
coherence, in terms of Centering costs, of the discourses produced by applying
these strategies.

As suggested above, production of discourses are exercised on connected
graphs. The agents are supposed to transmit these graphs. In order to see the
implications the general shapes of these graphs could have on coherence, we
have automatically generated 10,000 connected graphs, making such that their shapes fall in three main classes: long-shaped graphs (nodes have a small number of adjacent edges), wide-shaped graphs (nodes have a large number of adjacent edges) and random-shaped graphs (no constrains in the number of edges emerging from nodes). We were considering the two strategies of searching the graphs mentioned already, BFS and DFS, but also a Greedy selection, presented in [6] and a Random selection. The Greedy selection is basically a BFS traversing in which the selected node is the one with the greatest degree among the descendant nodes of the last mentioned node, and the RANDOM selection means a random selection of an edge among those not yet consumed. The results are presented below. The shape of the graphs are as follows: long-shaped graphs have 12 nodes and 13 edges, wide graphs have 12 nodes and 30 edges, and unconstrained graphs have 12 nodes and 20 edges.

In Figure 20, the Centering averaged scores are compared for long, wide and unconstrained graphs. Looking in each group of four, BFS yields discourses more coherent than DFS. As expected, the best one is Greedy and the worst is Random.

The figure shows identical Centering patterns for Long, Wide and Unconstrained: Greedy < BFS < DFS < Random. This finding is important for H4. It says that, irrespective of the graph shape, a persistent application of a Greedy method produces discourses more fluent than those produced by a persistent application of the BFS method, this of the DFS method and so on. In general, BFS and DFS are situated between Greedy and Random on the Centering scale. But, while BFS and DFS are "lazy" methods, because they imply very few movements on the graph from the node in focus at each step, both Greedy and Random are more costly from the point of view of operations on the graph (Greedy because at each step the selection is based on counting the emerging
Fig. 20. Centering coherence for different graph shapes and different strategies

edges of neighbouring nodes and ranking the nodes; Random because jumping from one node to another involves actually a traversal of the graph).

3.6 Supporting the H5 hypothesis

Hypothesis H5 claims that there is a correlation between the memory model and the sub-optimum coherence of generated discourses.

Indeed, we have proven in section 4 that the use of a STACK model of immediate memory, as well as that of the ZIGZAGGED-LIST model, yields minimum accessing costs, for a whole range of knowledge graphs and for all strategies of crossing the graphs. So we hypothesized that human agents have a built-in mechanism that implements one of these efficient accessing models. Since the two models are very close in terms of costs, we will refer in the following only to the STACK model. On the other hand, in producing discourses from graph-like internal knowledge representations guided by a short term memory implementing the STACK model (last in first out) yields a DFS ordering of traversing the graph. In other words, if the agents internal memory implements a stack, the search for another edge to mention next will happen around the last mentioned node, and this yields a DFS order. But the DFS order of searching graphs was proven in section 3.5 to produce discourses of a slightly worse coherence than the other method BFS, while both being below the Centering scores induced by RANDOM. Finally, human subjects seemed to prefer either DFS or BFS when uttering graphs, as demoed in section 3.4.1 (Figures 14 and 17). On yet another statistics employing human subjects (described in section 3.4.2) it was showed that human produced discourses are close to Centering-optimum. These findings close the demonstration circle. If we take Centering-optimum to correspond to a Greedy search, then the DFS ordering, as encumbered by the STACK model is sub-optimum. The BFS is even better, but it is more expensive. So, at the origin of the human-like behaviour with respect to discourse coherence, it seems
to stay an economy principle deeply implemented in a cognitive mechanism: minimisation of memory costs. If this is consistently applied, then the produced discourses have a fluency which resemble those of humans.

4 Conclusions

Summarising, this paper presents the research methodology and the first results of the search for a model to explain the human-like coherence in discourse. It is interesting that the two closely related aspects of coherence, cohesion and fluency, are grounded in completely different mechanisms: the use of pronouns an aspect of cohesion in discourse is grounded in linguistic games, while ordering of utterances an aspect of fluency is grounded in a basic cognitive machinery, the model of immediate memory. Both models, however, put at the base an economy principle: for the realisation of the communication goals through language there should be consumed a minimum cognitive effort. We think that the conclusion regarding the emergence of fluent discourses is spectacular in itself, because it says that it is sufficient for the agents to implement a low cost immediate memory model and this yields a level of coherence in direct communication that eases the understanding of the messages they produce.

Our investigation had the following components:

– the implementation of a parameterised framework of organising linguistic games permitted the description of a series of settings of increasing complexity, on which the spontaneous inception of pronouns was tested. It was revealed that the acquisition of pronouns in language can follow an evolutionist pattern, i.e. pronouns can appear as a natural, spontaneous, process, driven by the necessity of agents to acquire common understanding on a situation. This process is conditioned by the use of a memory channel remembering the object recently in focus. When such a channel is open, the identification of an object already mentioned, and which should be mentioned again, can be made quicker and with less ambiguity because it implies less categorisation;

– a parameterised model of immediate memory was proposed, as well as a formula to describe memory accessing costs. Then the cost of accessing the memory was investigated over a range of parameters, including the position of the write pointer and the memory length. The results showed that a model resembling the STACK data structure and the ZIGZAGGED-LIST model are more efficient then a model resembling the QUEUE data structure, when tested on 10,000 differently shaped graphs traversed following different strategies;

– to investigate the properties of the human discourse, we used a class of students which were asked to look at two types of images and to formulate discourses expressing the contained knowledge. Their discourses were then measured in terms of Centering transitions and histograms were produced. The experiment revealed that the discourses produced following a BFS order have a lower Centering load (being, therefore, more coherent) than those
produced using a DFS order. However, students prefer either a BFS or a DFS order of uttering a visual scene and a knowledge graph. In another experiment, we used all texts in a well-known corpus (GNOME), fabricated all permutations of prefixes up to a certain length (imposed by the power of computation), and showed that fewer instances were easier to process than the original discourse of the same length. The comparisons were expressed in terms of Centering transitions, which are believed to express adequately not only the discourse coherence but also the cognitive load. We interpreted this finding as an indication that fluency of human discourses are close to Centering-optimum, although not being optimal;

– to study the different strategies of selecting the next utterance and their implications over discourse coherence, we have generated programmatically a large number of differently shaped graphs, then we have computed the resulting Centering costs by using 4 different orders. It resulted that the 4 strategies as ranked as follows: Greedy < BFS < DFS < Random;

– finally, by correlating different findings, the final conclusion emerged: that at the origin of a coherent discourse following a pattern similar to that performed by humans stays a memory model, therefore a cognitive mechanism, and not a learning mechanism.

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 incremental optimisation implemented by the principle “grasp whatever is handiest”. In artificial intelligence, it is common knowledge that local optimisation amounts to local optima, and only by chance, to global optima. A similar thing happens with the human discourse. Humans are capable to globally obtain a comprehensible discourse without a significant consumption of processing power. They do this by grasping what is cheaper to utter at each step. As a result, they get a discourse which is only close to Centering-optimum, although not optimum. To obtain more coherence, they have to consume more inference resources. But, even more surprisingly, to obtain less coherence involves also more processing power.

We think that the research communicated in the present paper could be enhanced on at least the different directions:

– to study what are the semantic features that attract the specialisation of pronouns. Can the categories of male/female, animate/inanimate and singular/plural, as they are used to differentiate pronominal forms in most languages, be generalised? Could a class of experiments intended to put in evidence the different semantic features of anaphoric expressions be imagined within the limited worlds of the linguistic games? Another thing that we dont know yet is what are the levers that should be triggered to restrain the proliferation of lexical forms of pronouns in the community of agents, as in most natural languages there are few synonyms to express one category of pronouns;

– to get more evidence regarding the near-optimum coherence of human discourses, by performing more experiments and using more subjects. To find
a better way to compute the Centering-optimum value of a text, because proceeding along the prefixes of a given text and computing permutations results in statistics influenced by the uttered discourse. Then, to consider other means of experimenting discourses produced by human subjects in order to stabilise the preference for one of the searching strategies (maybe the balance between DFS and BFS is not real);

– to experiment the direct correlation between the memory model and the Centering scores produced, in the eventuality that the ordering of utterances is entirely driven by the memory;

– to study if the placement of the human discourse at a certain point in-between the optimum and the worst corresponds to a similar placement of the simulated discourses on the same scale.

Acknowledgments

This research is supported by the FP7 grant ALEAR (Artificial Language Evolution of Autonomous Robots). The part of the research on cohesion was also contributed by Corina Dima and Emanuel Dima.

References