tripleC 4(2): 265-276, 2006
ISSN 1726-670X
http://tripleC.uti.at

Information Structure Representation And Extraction From A Corpus Of Patient Data, Using An Ontology

Christian Cote

ERSICOM - Université Jean Moulin Lyon 3
6, cours Albert Thomas - BP 8242 - 69 355 Lyon cedex 08
cote@univ-lyon3.fr

Abstract: We propose a methodology to model the information structure for its extraction from any medical text. We experiment this extraction in a corpus that represents the information system of a specific professional activity in the hospital pharmacy. the information structure represent how the meaning of a sentence is specified by the constraints of the information flow.

But the information structure is systematically recognized and interpreted in the context of a text: it's also the last object of the information system. Then we consider the text as a contextual frame to model the recognition and the extraction of the information structure.

A text can't be considered only as a linguistic object in a professional and information context: it's an implemented (or externalised following situated and distributed cognition) ontology. The updated text articulates the ontology of the patient body and the referential dimension of the information.

The model of the information structure presupposes we know what are the constraints of the information system on the symbolic entities (in a way to distinguish the information structure to any sentence description). In a way to determine these constraints, we propose to model the information process by the information flow: we represent in this way how any fact in the body of the patient is symbolised, conveyed and represented into a text.

The information flow characterizes only the constraints of the information on the linguistic entities and structures. But the information is linguistically a referential semantic object: it's the representation at distance of a new fact in the world in the frame of a text that accepts this information. Then the model of
At last, The information flow allows the articulation of an ontology and a semantic precisely on the question of the information structure. We unify the model of the information structure by the definition of five primitives. A sign representation allows both the characterisation of the structure and of each of its components.

Keywords: information structure, information flow, semantic.

Acknowledgement: Text

1 INTRODUCTION

We propose a methodology to extract any information structure (considered as the conveyance of a new fact) from text (considered as a location of representation and interpretation). An information structure is an utterance that conveys a new fact in the world. It overlaps the frame of the sentence (Steedman and Kruijff-Korbayovà, 2003). In a way to propose a model, we will deduce its representation from the information flow (Barwise and J. Seligman, 1997).

The question of the information (and especially the information structure) is both a linguistic, cultural and cognitive problem and is ordinarily considered essentially under one or two parameters: linguistic (Harris, 1988), linguistic and cognitive (Jackendoff, 1990), or cultural (Hutchins, 1995). I propose a more pluri-disciplinary framework: in a cognitive background, (following the propositions of the cultural cognition (D'Andrade, 1989)) an information system (considered as a cultural tool) uses linguistic entities (and their specific syntactic and semantic properties) to convey any fact form the world to its location of

interpretation (or use). Our project is founded on the articulation of three different objects: information system, linguistic entities and knowledge.

Our essential claim is that the context of the activity and the information system constraint the linguistic operations. Then the question of the Information Structure is not only a linguistic problem, but the articulation between linguistic tools, cultural process and knowledge. (The information is systematically recognized in the frame of a common knowledge, (Dretske, 1981)). We will represent this articulation by primitives associated to an ontology.

Less technically, an information has meaning considering the operations of information building: " M. Smith, kinetic, concentrations, 5,56 µg/l, T0, 12h32 " means under the recognition of the fact that an entity marked on time ("T0, 12h32 ") contains a property ("µg/l") that is measured by an operation and its issue is the numeral "5,56" (chosen in a scale). The represented fact is a part of the individual (" M. Smith") process of "kinetic".

1.1 Corpus: definition and representation

In the frame of an hospital, the adaptation of posology is an activity that mobilises an information system from the patient to the pharmacist that controls the therapy. This task is performed only with the aid of the information that stay in the regularly updated sheets dedicated to this patient.

The corpus is defined as an ideal-type (Weber, 1910/1995): it's a limited and autonomous frame to observe the common process of information production, conveyance and interpretation. The corpus is constructed, that means there is different ways to characterize it: the updating of the sheets, the production and conveyance of the information structure and its context. We limit its description to a whole included material information system (rules of encoding, selection, conveyance and decoding) containing the linguistic processes: the whole available lexicons (a limited sub-language) and the situation of interpretation (a frame for reasoning at distance and the agent ability to memorize and learn some operations and features of the information system).

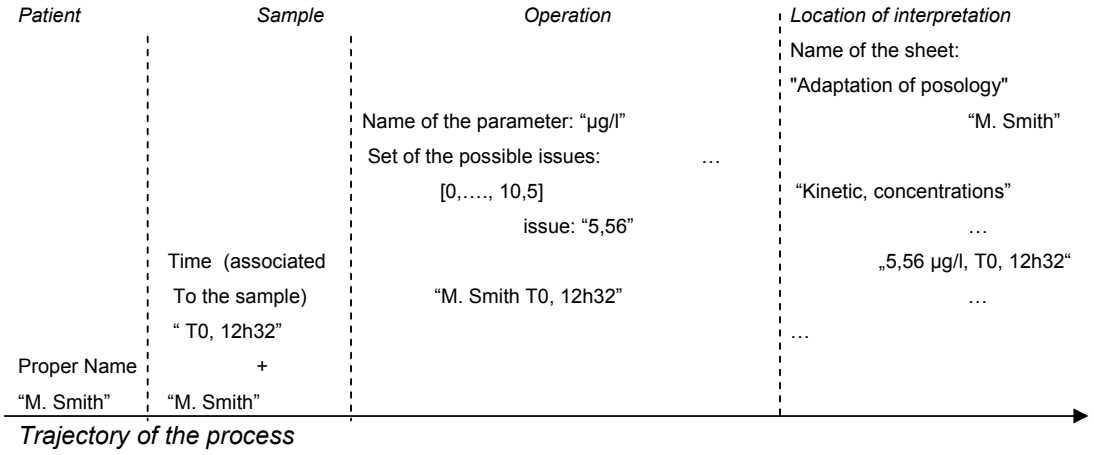


Figure 1: Schematisation of the corpus

(Cf. Appendix A).

In the frame of the information production, the agents interventions are limited to actions from individuals to restricted objects in time (samples). The operations are functions from an object to a symbol under a parameter (that is a previously selected entity or a lexicon). The text is the channel: it classifies any issue of the symbolization process accepted by the model of the patient. We identify the information structure by the classification of a conveyed suit of entities.

The location of interpretation designates the spatial frame of the sheet: the text is composed of a lexicon where each entity is associated to a spatial frame in a way to define a diagrammatic representation. The first operation of conveyance associates a proper name to the whole spatial dimension, and then to each constant of the sheet. In the second operation, in any attributed diagram, a

time marker (eventually associated to an order marker) classifies the conveyed succession of entities into a diagram to update an event.

1.2 Hypothesis

The theoretical background of the distributed and situated cognition (Agree, 1997) allows the definition of the whole information system as a cultural tool –a regular process in a specific activity with its proper sub-language- and the interpretation as inferences on the state of the world by reasoning at distance. In this frame, the information system is a process: the operations are implemented into material tools in a way to symbolize some objects or parts of objects. The operations are implemented reasoning (Hutchins, 1995). Following this theory, the text is a tool of interpretation. The diagrammatic dimension of the sheet in relation to the specific distribution of the printed lexicon allows associating the textual frame to the knowledge of all the possible events in the world. The text is characterized by a model (Halliday and Hasan, 1985) that accepts some specific structure: information is recognized and interpreted as the representation of a new fact in the world under the parameter and constraints of the information flow. If we consider the linguistic processes in this frame, they are constrained operations (syntactic dependencies and semantic interpretation). For example, the argument selection is captured by the models of the distributed cognition but only as operation; the translation of this selection into constrained linguistic operations (words dependencies and distribution) is described by a relation between a larger and a limited processes. The information structure will be precisely explained by two levels of rules: the high level of the information flow process and low level of the deep syntax: the syntactic representation is the optimal description of the realized succession of entities.

Our segmentation is both lexical and cognitive, considering that the syntactic and semantic rules are constrained by the lexicons. The lexicons are some under-specified meanings and the constraints of the flow allow the inference of a precise reference. The representation of these regular constraints on the lexicons reveals common sense ontology. The ontology proposes a precision because each entity of the composition refers into a specific domain. The precision is obtained both by the arguments of the relation, and by the parameter. The parameter is the issue of a previously treated situation.

1.3 Methodology

The coherency of the whole corpus is defined by the information system.An information system is a regular process of information production and conveyance from a part of the world (using a channel represented by one constant of the text) to the required textual frame. The operations of symbolization have an order that can be described by MARKOV chains (Van Der Lubbe, 1997).The chains are relative to only one distributed “channel” of the activity. Because every channel converges to the textual frame, then the sheet is the global channel of the distributed system. The corpus can be regarded as an “implemented” information system: a concrete process that satisfies the requirements of the distributed information theory.

If the selection is a principle of the information system, how the selection articulates an operator and an argument to symbolize is the process of predication (defined for example by HARRIS). The succession of the lexicons and their non-substitutivity, the order of the operations in the information process, compose the conveyed structure; they are both questions of flow and language. (The probabilities of the realization of one specific occurrence are context dependent: they have no relation to an information problem).

The Information theory considers only how some quantities of units are selected from lexicons and the order of these operations. Each type of operation is different: the process of selection, the memory and the order can't be explained only by quantities and formal properties of the entities: the operations are explained by the contents.

The articulation between ground and focus are characterized by functions among classifications: the order of the operations represents exactly this dichotomy.

The regularity of these operations characterizes the primitives of the Information Structure. In a frame for the interpretation founded on the knowledge of the flow operations, these primitives are characterized as ontology.

We obtain a semantic considering that the discourse domain is constructed by the information flow (natural world at distance, information system, patient representation). We can propose a predicative representation of each type of entities using tools of Situation Theory (Devlin, 1991). The relational and contextual constraints of each component allow their characterization by an information ontology (that associates both lexical foundation and operations of the flow).

2 INFORMATION FLOW MODEL TO REPRESENT THE PROCESS

The symbolization is founded on the classification of an occurrence (that can be linguistically indexed). The type that classifies is recognized into a knowledge domain. The occurrence and the type are verified into a situation in the world. A classification is not information: the information requires a more precise and limited classification or inversely, a precise classification is translated into a more common and general classification.

2.1 Classification

The initial presentation of the information flow uses a set theoretic presentation. Accordingly to the previous remarks we prefer a type theoretic representation. We characterize three sorts of entities:

[a, b, c ,…] : set of occurrences (indexed or denominated by a marker of time) where [A, B, X, E] characterize the type of each occurrence.

[α, β, δ, κ, o, υ] : sets of types where [Γ, Δ, Φ, Ψ] characterize the sets of types.

Relation $\vDash$: operations between entities of different levels (occurrence-type, type-occurrence-channel). Each relation is then a triple: $<\vDash, a, \alpha>$. It associates two entities and a particular situation in the world (characterized by the symbols $[s_1, \ldots, s_n]$).

The operations produce some expressions (or constrained successions of selected entities). These conveyed expressions are represented in isolation and in a collection. This relation is non-directional. Every occurrence and every type are of lexicons characterized by different ways to refer. The situation that supports the classification is of a type that accepts the two entities (occurrence and type).

Ex. “M. Smith, kg,”

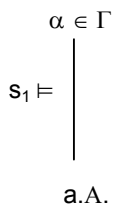


Figure 2: classification

2.2 Functions

The functions $f^{\wedge}$ and$f^{\vee}$, $g^{\wedge}$ and $g^{\vee}$,$h^{\wedge}$ and $h^{\vee}$ characterize the operations among entities of the same level (occurrence, types, channels) or “translation”. The functions are characterized by contra-variant pairs among two classifications: each function connects two entities of the same level. The initial form b $\leftarrow$a and $\alpha \rightarrow \chi$ represents an infomorphism: a common symbol specified on time by an entity in a scale, and the inclusion of sample into the patient body.

b ←a represents the type of property as condition to the numeral and $\alpha \rightarrow \chi$ the sample as condition to the analysis of the individual.

The triples are: < s_1□, a, α>, < s_2□, b, β>,

$f^{\wedge}$: Γ→Δ

$f^{\vee}$: A←B

Ex. “M. Smith, kg,”, “ 2,58 mg/ml ”

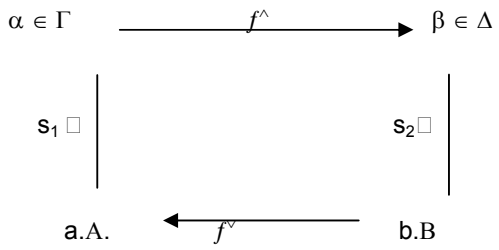


Figure 3: informorphism

The functions allow the characterization of constrained structures as “ 2,58 mg/ml ” or “ M. Smith 14:54 ”. If the functions represent how a type is translated into a more precise, the obtained structure is an extended situation (with the adding of the two types and the inclusion of the smaller occurrence). We represent now the fact that the functions represent more than two classifications:

α→ α ⊕ β ← βcharacterizes the fact that the channel conveys both the condition and the issue at the level of the types.

a → <a ⊇ b> ←b represents each entity and their inclusion at the level of the occurrences.

The distinction of operators is due to the specific structure of the domains.

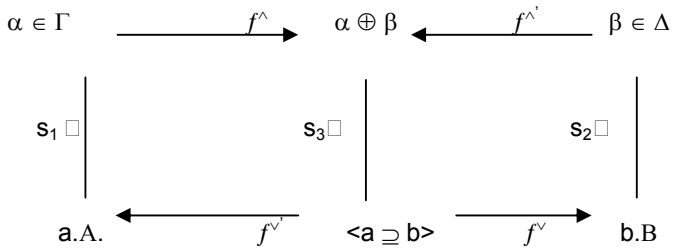


Figure 4: information

Every new information is a set of functions from one classification to another where one type is translated into a more specific and inversely at the level of the occurrences. Each classification and its adding under the channel characterize the whole content of the information.

2.3 Channel

The infomorphism represents only a focus and a ground in a situation. This fact can be conveyed only if the infomorphism is classified into another type. This type or channel is common to the information production frame and the location of interpretation. Every channel is a symbolic entity that classifies and conveys an infomorphism to its location of interpretation.

Each conveyed and time indexed structure represents a fact: the channel is an event structure (the representation of a process in duration) when a collection of information is conveyed. It's the first classification of information.

The duration allows understanding the second classification as the classification of information collections into an information state (Guinzburg and Cooper, 2004). An information is more precisely

interpreted in its whole context than in isolation: the information state of a channel bounds the possible interpretations of any of its classified information.

A channel in the frame of a distributed information system is classified into a global channel.

Every channel is a classification of the two initial classifications and the information classification. Then it represents another situation.

The functional rule that translates a type into a more precise is verified in this case. In fact, as in a conditional frame, the choice of the lexicon of precision (Δ) is associated to both Γ, Φ and Ψ. (Φ represents the lexicon of channels and Ψ the lexicon of the global channel).

In the representation, $\kappa \in \Gamma$, $\sigma \in \Phi$ and $\upsilon \in \Psi$.

The functions $g^{\wedge}$ and $h^{\wedge}$ represent respectively the type to channel translation and the distributed to global channel translation. The inverse of each function ($g^{\vee}$ and $h^{\vee}$) is the translation into a local lexicon (Δ) for the channel and any accepted expression for the global lexicon (Φ). The global channel (the text name) selects the whole the sub-language.

The distinction between the two schematizations can be summarized by the distinction between the process of symbolization and the textual operations of symbol manipulation. (These last represent more extended objects in the world)

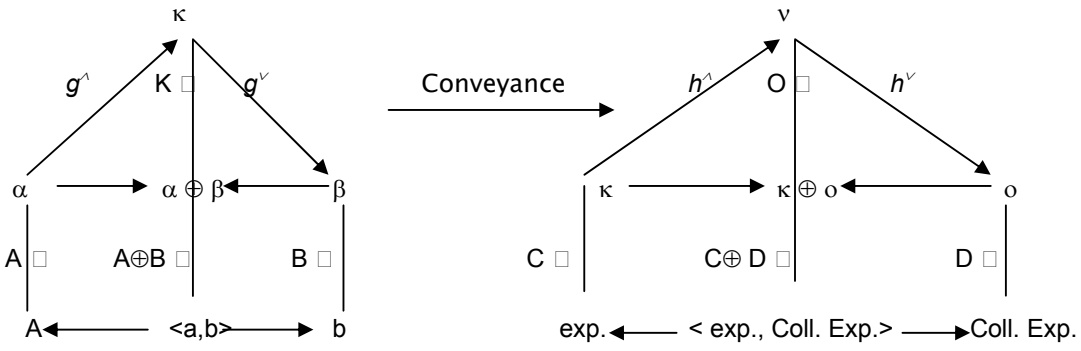
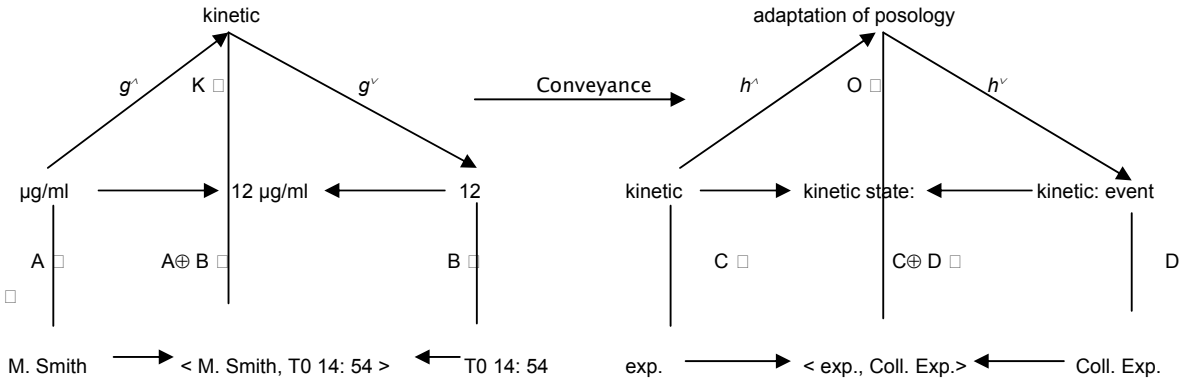


Figure 5: information flow



"Exp.": expression or the whole conveyed suit of symbols.
" Coll. Exp" : collection of expressions into a diagram.

Figure 6: Information flow: example

The information flow characterizes (1) the symbolization at one time and (2) the identification of the newest symbolized fact into a larger representation by a pair of contra-variant functions from a fact to its precision (obtained by the causality or the context of the previous facts).

We can precise the text model. The global channel conveys (or duplicates) any obtained suit of symbols from the location of information production to a text. The condition is that this text accepts the same individual a (“M. Smith”).

If we consider the global channel (“adaptation of posology”), it classifies any event contained in text and the patient name (a). The model represents only the process that allows the interpretation: the text is characterized in the information process, where one individual occurrence is functionally relied to its more extended representation.

2.4 Questions of states spaces

Our previous description associates systematically the information process to the symbolization. But this description is unsatisfying when we consider the knowledge or more simply, the empiric succession of the operations. The state space is then the representation of the trajectory in the real world. Each pair of function is co-variant.

3 DOMAINS, PRIMITIVES AND ONTOLOGY

The flow allows the characterization of conventional constraints to select lexicons and determine the behavior of the linguistic entities.

We introduce some consequences of the flow on some conceptual tools used into the semantic analysis. The interpretation of the information is not only linguistic: it implies the memory of the process of information production. Every information is interpreted under the capacities of the distributed channels. This memory is not heterogeneous but associated to the interpretation at distance because the flow is the condition for any information about a patient.

3.1 Domains of discourse

The first consequence of the information flow on the semantic analysis is the definition of distributed domains of discourse. The domains of discourse [14, p. 446] are constituted by the limited sorts of things allowed by the discourse. (This can be distinguished to the reference of the expression because it includes the counter-factual and the other inferred entities). But the distributed domains are not only more or less extended, they represent more and less precise segmentations into the objects in the world. Entities and expressions are interpreted both into the domain of the information production and into the domain of the patient with the aid of the specific domain of the text (the pre-printed sheets) that overlaps the previous.

The different steps of the information structure entail the definition of three ordered domains:

D: $d_0 \leq d_1 \leq d_2$

d_0 for the natural world: it corresponds to the entities designated by the occurrences.

This is a domain of individuals:

$d_0 : [(a_1, \ldots., a_n) \supseteq (b_1, \ldots., b_n.)]$ The inclusion characterizes the minimal structure of the universe. This universe is only composed of material objects. This domain is not finite and the symbols of these entities are only indexicals. If we consider every possible occurrence, we write the occurrence level by a type: λa.T, λb.T.

d_1 is the domain of the symbolization: it's composed by the different functional operations and the specific organization of each lexicon: every type is integrated into this universe. (We have in this universe exactly the trajectory of the information, then the state spaces previously mentioned). This domain is organized by a semi-lattice representation (Link, 2002). $d_1 : <<\{\alpha_1 \vee \ldots \vee \alpha_n\} \bullet \{\beta_1 \vee \ldots \vee \beta_n\}> \bullet \{\kappa_1 \vee \ldots \vee \kappa_n\}>$.

•: operation of concatenation.

d_2 is the domain of the information system, including the information classification into the text; this universe is constituted by the previous and the representation of processes in duration (or event). It represents the domain of the global channel,

$d_2 : <<\{\{\kappa_1 \wedge \ldots \wedge \kappa_n\}>. \} \bullet \{o_1 \wedge \ldots \wedge o_n\}> \bullet \{\nu\}>$.

If we observe more precisely, the domains are distinct to the situations. The situations represent the superposition of the different domains at the level of an unique Information Structure. Considering the previous order among the domains, we have a relation when one entity is selected to classify in another domain the first entity. One entity of a domain predicts on one entity of a less or more precise domain. In this way, we preserve the distinction between the domains as sets of possible entities, facts or events, and the realized situations. A situation is a frame in the world where two domains are in relation (by a classification). This is why "M. Smith, 59 kg" describes both the weight of a patient and the operation of measure of a patient weight.

3.2 From primitives to ontology

The channel limits the quantity of entities and their classification. Only the conveyed structures encapsulate information (the text economy characterizes some surface syntactic rules allowing the recognition of the Information Structure in the text). The Information Structure is a "reconstructed" expression like these : " M. Smith, kinetic, concentrations, 5,56 µg/l, T0, 12h32 "or "M. Smith, weigh, 52kg, 10/12/04".

How and why only this sort of structure conveys a new content and allows its recognition: the information flow analysis isolates five entities that compose the Information Structure but can't characterize exactly the structure of each entity and its exact role in the representation of the fact. These five entities are a more precise characterization than the usual two or three articulations on the sentence, usually proposed (Steedman and Kruijff-Korbayovà, 2003).

We follow the general frame of the Situation Theory (Devlin, 1991). We introduce the following new concepts :

□ : situation supporting a relation. A situation is considered as a segmentation in the world founded on a predicative relation.

□: condition. A condition is a pre-requisite to interpret and refer to a previously interpreted situation.

R: relation. The relation is systematically attested by a recognized structure (a diagram or a linear syntactic realization) in text. The formula s_D □ [Ra (b, c) ↾ d] characterize the situation and the parametric predicative relation between two arguments.

We consider as an INDIVIDUAL any relation between a name (a) and a set of channels parameterized by a global channel (or the name of the activity). The individual is not a characteristic of the proper name, but of the situation that "supports" the relation between a set of possible events on the domain of the adaptation and the domain characterized by the referent of the proper name. The relation is submitted to the representation proposed by the global channel (i.e. the diagrammatic representation).

INDIVIDUAL:

s_D □ [Ra (Φ) ↾ν]

An entity on time is characterized by the relation between a distributed channel and a precise type (that limits the measured things into the object), under the parameter of the inclusion of this object into the individual.

ENTITY IN TIME:

s_D □ [Rb (κ_i, β_i) ↾a]

The channel represents a process in duration in relation to the individual. The relation is limited by the parameters of the lexicons of types. This is an event (Steedman, 2002) or the linguistic representation of an abstract or conceptual process.

EVENT:

s_D □ [Rν (a, κ), (ο, (b_1, …., b_n) ↾ (Γ → Δ)].

The property is defined by a relation, marked by the type α. It's a relation between an individual that satisfies this property and a lexicon of specifications. The property is static and is something that has a relevance under the parameter of one event.

PROPERTY:

s_D □ [Rα. (a, (β_1,…, β_n)) ↾ κ]

The specification allows the representation of a temporary state of the event, under the parameter of a certain property. The relation is between an event and an entity on time.

SPECIFICATION:

s_D □ [Rβ (κ, b_i) ↾ α]

4 CONCLUSION

I have presented how a logic model can be fruitfully applied into a real world corpus in a way to extract the non-linguistic properties of an information structure. The interest of a corpus-directed approach is its ability to associate heterogeneous entities in the representation of a process. The corpus is considered as

an ideal type of the regular process: the information system can be both mathematically characterized and empirically isolated in context.

The model of the flow has consequences both on the linguistic analysis and on the characterization of the reasoning at distance.

Our work respects some theories of the language (Harris, 1988) and can be considered as a context representation. Considering that the context constraints the utterance and specifies the meaning of the linguistic entities and structures, we entail from the operations of the information flow the syntactic specific operations of the information structure (Muskens, 2003) and the under-specification of the semantic.

Following the corpus representation, the text is the model of a patient representation containing each constant (or distributed channel) assigned to any individual. The text is not informational: this is a regular frame that accepts any new information. Then, following situated and distributed cognition, it can be considered as a knowledge representation.

The information system is a part of the activity of adaptation. Accordingly to information flow theory, the reasoning at distance is founded on uncertainty and non-monotony because the text is a partial representation of the individuals in the world. There is no correlation between the precision of the information (and the collections of information) and the persistency of the uncertainty. The reasoning at distance integrates the dimensions of the local situation of interpretation in time (Recanati, 1996). The propositional dimension of the interpretation and the pragmatic perspective will be considered in the course of the activity.

References

Agree P. (1997), *Computation and Human Experience*, Cambridge, Cambridge University Press.

Barwise J. & Seligman J. (1997) *Information flow: the logic of Distributed Systems*, Cambridge University Press.

D'Andrade (1989), "Cultural Cognition", in R. Posner, *Foundations of Cognitive Sciences*, MIT Press

Devlin K. (1991) *Logic and Information*, Cambridge University Press.

Dretske, F R. G.. (1981), *Knowledge and information flow*, Cambridge : Cambridge University Press.

Guinzburg J. and Cooper R., (2004), "Clarification Ellipsis and the Nature of Contextual Updates in Dialogue", in *Linguistic and Philosophy*, **27**, number 3, Kluwer pp. 297-365.

Halliday M. A. K & Hasan R. (1985), *Language context and text: Aspects of language in a social-semiotic perspective.* Oxford, UK: Oxford University Press

Harris Z. (1988), *Language and Information*, New York : Columbia University Press.

Hutchins E. (1995), *Cognition in the wild*, Cambridge, MIT Press.

Jackendoff R., (1990), *Languages of the mind*, Bradford/MIT Press

Link, G. (2002), "The Logical Analysis of plurals and Mass-terms, in P. Portner and B. H. Partee, *Formal Semantics: the essential Readings*, Blackwell Publishing, pp. 127-146.

Muskens R. (2003) "Language, Lambdas, and Logic". In Geert-Jan Kruijff and Richard Oehrle, editors, *Resource Sensitivity in Binding and Anaphora*, Studies in Linguistics and Philosophy, pages 23-54. Kluwer

Recanati F., (1996) "Domains of discourse" ,in *Linguistic and philosophy*, **19**, Kluwer, pp. 445-75

Steedman M. (2002), "Plans, affordances, and Combinatory Grammar", Draft 2 may 17th –a revised version is to appear in *Linguistic and Philosophy*, **25**

Steedman M.& Kruijff-Korbayovà I. (2003), "Discourse and Information Structure", *Journal of Logic, Language and Information*, Vol. 12, 249-259.

Van Der Lubbe J. A. (1997), *Information Theory*, Cambridge : Cambridge University Press.

Weber M., (1910/1995), *Economie et Société*, Paris, Plon.

APPENDIX A