

Clues for the Identification of Implicit Information in Multimodal Referring Actions – DRAFT VERSION

Frédéric Landragin

LORIA
Campus scientifique – BP 239
F-54506 Vandœuvre-lès-Nancy CEDEX
FRANCE
Frederic.Landragin@loria.fr

Abstract

The implicit is an imprecise and heterogeneous notion that plays a role, not only in the global comprehension of utterances, but also in the interpretation of reduced phenomena like multimodal referring actions, which combine visual perception, language, and gesture. The identification of the intended referents relies on the correct identification of implicit information that is communicated with the multimodal utterance. The implicit can be linked either to the conjoined use of multiple modalities, to the cognitive status of referents and reference contexts, to the dialogue history, to inferences which occur during the interpretation process, and to interpretative effects for future utterances. For each of these types of the implicit, we focus on the determination of clues that allow the specification of reference domains (structured sub-contexts for reference resolution). We show that this notion of reference domain can integrate all these aspects with computational objectives, thus laying the foundations to a future algorithm for implicit identification.

1 Introduction

With the development of speech and multimodal interfaces, it is easier to imagine to what a spontaneous human-computer interaction may tend towards. If it is pleasant to hear a system react with “*which one?*” to a request like “*remove the blue object*”, and if it is all the more pleasant to see an avatar or a robot pointing out an object and asking “*this one?*”, we want here to claim that such reactions show a good recognition of the user’s message and intention but not efficient interpretation abilities. Why? Because in usual communication situations the user’s utterance is not intentionally ambiguous but relies on a presumption of success to be interpreted, and because this presumption is based on implicit information (not present in the transmitted code). Following Relevance Theory (Sperber & Wilson, 1995), communication is seen as ostensive and inferential. Inferences have to be made from the ostensive clues to identify the implicit. In man-machine dialogue, the correct interpretation implies the best implicit identification capabilities. What is the nature of the implicit, and how can a system identify it? We focus here on the nature of the implicit that occurs during the resolution of reference in a multimodal context (man-machine dialogue with a visual support). We propose a classification of the implicit and of clues that the system may exploit to identify it. We show how the model of ‘reference domains’ constitutes an efficient framework for the formalization of all these heterogeneous pieces of information. Then we conclude about the design of multimodal dialogue systems.

2 Implicit and Multimodal Referring Actions

(Grice, 1975) distinguishes two levels of meaning: ‘what is said’ and ‘what is implicated’. The first corresponds to what is explicitly stated in the verbal utterance, and the second to the implicit that is implied from it. The main idea is that the proposition conveyed by the utterance is built from what is said, and that its interpretation requires implicatures. Considering that some information is not said but explicitly communicated, (Sperber & Wilson, 1995) and (Bach, 1994) add a middle level of meaning. This information (‘making as if to say’) is called implicature by Bach and completes the proposition conveyed by the utterance. Sperber & Wilson call explicature the resultant enriched proposition. Their cognitive theory of relevance emphasizes the exploitation of ostensive clues for implicit identification, but includes no computational model for that, as we need for man-machine dialogue. With more computational objectives, (Grosz & Sidner, 1986) propose a model of implicit identification through the notions of intentional structure and attentional state. The problem is that these notions are more explicative than operational. The authors say themselves that the two structures are related: the first is a primary factor in determining the second, and the second helps constrain the first. Their computation is by consequence difficult. (Geurts, 1999) formalizes with DRS (Discourse Representation Structure) the implicit linked to presuppositions. Nevertheless, this important extension of the theory of (Kamp & Reyle, 1993) follows the same restriction to linguistic considerations. The lack is in the exploitation of the visual context. In dialogue with visual support, the implicit depends on the verbal utterance as well as on extra-linguistic characteristics. Visual context is taken into account by (Beun & Cremers, 1998) with focus spaces, and by (Kievit, Piwek, Beun & Bunt, 2001) with visual salience. Beun & Cremers present a computational model of identification of focus spaces in dialogue with visual support. One important point is that they determine a strong clue to detect a change of focus space: redundancy. They show that the presence of redundant information in the utterance denotes the user’s intention to reconsider the context. Such a model is very near to what we imagine for a dialogue system. It consists in a formalization of the implicit using data structures, and in clues for their identification. To complete this point of view, the DenK system (Kievit et al., 2001) exploits salience as a visual clue to choose between objects when searching for referents. Then salience constitutes a useful implicit information. Another important point of the DenK system is the identification of the implicit in a multimodal context, including speech and gesture. A weak point is that inferences are ignored. We want here to try to put together all these aspects of the implicit in a multimodal context.

The work we pursue in multimodal reference resolution (Landragin, Salmon-Alt & Romary, 2002) shows that each referring action implies the activation of a reference domain. This subset of contextual information can come from visual perception, language or gesture, or can be linked to the dialogue history or the task’s constraints. The utterance’s components allow to extract the referents from this sub-context and to prepare the interpretation of a future reference. For example, “*the red triangle*” includes two properties that must be discriminative in a reference domain that must include one or more ‘not-red triangles’. The use of the referring expression, and particularly the linguistic constraint conveyed by the determiner, is then justified. A further referring expression like “*the other triangles*” may be interpreted in the same domain, denoting a continuity in the reference sequence. On the other hand, the use of a demonstrative determiner (“*this triangle*” associated to a gesture) implies that another triangle is present in the reference domain, in order to justify the spatial contrast denoted by this determiner and the gesture. The same domain will be used for the interpretation of “*the other one*”. Some reference domains may come from perceptual grouping (see (Landragin et al., 2002) for the computation of the Gestalt criteria to build visual domains). Some may come from the user’s gesture (Landragin, 2002), others from the

task's constraints. All of them are structured in the same way. They include a grouping factor (being in the same referring expression, being in the same perceptual group, etc.), and one or more partitions of elements, each partition being characterized by a differentiation criterion ('red' vs. 'not-red', 'focused' vs. 'not-focused', etc.). This unified framework allows to confront the various contexts, and to model the implicit whatever its origin between perception, speech and gesture.

3 A Classification of Implicit Information

Some implicit information intervenes during the specification of the semantic form of the current utterance, so before its pragmatic interpretation (section 3.1). This implicit may be distinguished from the one used for the pragmatic interpretation, that corresponds to constraints for referents identification (sections 3.2, 3.3 and 3.4). A third kind of implicit intervenes after the pragmatic interpretation and is important for the dialogue continuation (section 3.5).

3.1 Implicit Linked to the Conjoined Use of Multiple Modalities

In order to build the logical form corresponding to the enriched proposition conveyed by the utterance, we have to precise the nature of the implicit link between modalities. First, we consider the presumption that a gesture is associated to the linguistic expression. A simple clue is the presence of a demonstrative determiner. This clue is particularly strong when all anaphora are impossible considering the linguistic context. Another strong clue is a temporal synchronicity between speech and gesture. Second, we consider the implicit link between the pointed objects and the referents. The interpretation may be generic: "*I love these cars*" associated to a gesture pointing out a Ferrari. The main clues are linguistic: the aspect of the predicate (unlike 'love', 'possess' and 'destroy' have a punctual aspect and lead to a specific interpretation); the presence of a numeral or a quantifier (both of them force the specific interpretation). Our idea here is that the presence of a gesture does not change anything to the possible ambiguity between specific and generic interpretation. The very common utterance "*this N*" associated to a gesture pointing out one N can always refer to all N of the corresponding type. Even for a specific interpretation, the pointed object may differ from the referent: "*he has a big head*" associated to a gesture pointing out a hat. The clue here is an incoherence between speech and gesture. Third, the association of speech and gesture is also an implicit association of criteria for searching referents. Following componential semantics, a lot of criteria like category, properties, cardinality, and spatial focusing due to the gesture, work together in the reference resolution process. An example of clue is the association of a definite determiner and a gesture, that denotes the extraction of a particular referent in the domain delimited by the gesture (Landragin, 2002). The use of reference domains is a way to formalize the implicit link between modalities, because the differentiation criteria can complete each other in order to merge domains built from the gesture and from the linguistic form.

3.2 Implicit Linked to the Cognitive Status of Referents and Domains

The resolution of reference to objects requires to test each object of the situation. Some objects may be more accessible than others. For example, (Gündel, Hedberg & Zacharski, 1993) propose a classification of cognitive status of entities invoked in the interaction. Considering the form of the verbal referring expression, the referent must be focused ("*My neighbor has a dog. It kept me awake*"); activated in the low-term memory ("*My neighbor has a dog. This dog kept me awake*"); familiar in the long-term memory ("*I couldn't sleep last night. That dog kept me awake*"); etc. This implicit has been built from the previous utterances. Another implicit is linked to the visual accessibility of entities. This is the role of salience. The more an object is salient, the more it is

accessible for a referring action. A system based on implicit identification must then include a model of salience. Moreover, we said that a referring action implies the activation of an implicit reference domain. The accessibility of such domains is also an important aspect in the interpretation, and constitutes the implicit linked to the cognitive status of the interpretation context. The model of reference domains, that includes a model of salience and the management of a stack of domains, appears as a good way to formalize this kind of the implicit.

3.3 Implicit Linked to the Dialogue History

When following the same referring strategy, the user gives implicitly indications to the system. For example, in the corpus presented in (Wolff, De Angeli & Romary, 1998) and corresponding to a tidying task, two strategies may be identified: tidying guided by the category of the objects, and tidying guided by the visual perception of objects. These strategies correspond to scripts (or frames). If the user always follows one of them, his referring actions may be easier to understand. These scripts can be formalized in reference domains: the grouping factor corresponds to the strategy and the differentiation criteria are linked to the steps.

3.4 Implicit Linked to Inferences for the Interpretation

Following (Ducrot, 1991), the term ‘implication’ groups together all inferences that are computed from the logic form of the utterance and from the context. It includes presuppositions and implicatures. Presuppositions are important in the reference resolution process because they reduce the possibilities. For example, the interpretation of “*tidy these objects*” will only consider the objects that are not tidied yet. Implicatures occur when the user does not follow the cooperative principle of communication (Grice, 1975). For example, he may assert “*these objects have to be tidied*” instead of ordering, with the same communicative intention. The speech act is implicit and has to be identified by the system. In this manner, some utterances may include an ironic intention (or other emotive characteristics) that might be identified, too. The problem is very large because the information needed to understand is boundless. A solution is to take into account only the information linked to the concepts of the utterance and of the visual context. That is what is done in the model of reference domains.

3.5 Implicit Effects for Future Utterances

The last kind of the implicit consists in interpretative effects for future utterances. For example, “*the first*” may be followed by “*the second*”. In French, “*l’un*” will be necessarily followed by “*l’autre*”. These baiting effects are linked to the way information is presented in the utterance (referents are presented taking into account the future possibilities of anaphora). Clues are deduced from the referring expression and also from visual perception or from task constraints. For example, a spatial disposition of objects or a particular sequence of actions due to the task may incite to referring expressions like “*the next one*”. Once more, reference domains provide a structure for the formalization of such heterogeneous constraints.

4 Towards the Computation of Clues

What are the computational consequences of these phenomena? First, the necessity to manage mental representations of the entities that are in the user’s mind. Without them, it seems impossible to manage cognitive status and implicit domains corresponding to attentional state. Second, the necessity to clearly define algorithms for the exploitation of the clues that lead to

implicit identification. Considering the referring expression, constraints are formulated about the referents and the possible reference domains. These constraints filter the potential referents and allow to identify the correct one, whose cognitive status is then set as focused. A third consequence is methodological. It is very difficult to validate an implicit notion like mental representations. Psycholinguistic experimentations involving sequences of references constitute a solution. The interpretation of expressions like “*the other one*” can prove the existence of an implicit reference domain. The possibilities of such a methodology are numerous.

5 Conclusion

To conclude, we think that a simple microphone associated to algorithms based on implicit identification and on the exploitation of clues like a determiner or characteristics of the visual context will be more efficient than a lot of devices associated to recognition and fusion algorithms. Technical aspects must not mask semantics and pragmatics concerns. The approach of underdetermined semantics and the one characterized by a fine treatment of grammatical words like determiners (as we show with our examples) are just beginning. We hope that the notion of reference domain, that corresponds to unified structures modeling heterogeneous mental representations, will be useful to the design of multimodal systems, first for the interpretation of multimodal referring expressions, and then for the comprehension of the whole utterance.

References

- Bach, K. (1994). Conversational Implicature. *Mind and Language*, 9, 124-162.
- Beun, R.-J., & Cremers, A. H. M. (1998). Object Reference in a Shared Domain of Conversation. *Pragmatics and Cognition*, 6 (1/2), 121-152.
- Ducrot, O. (1991). *Dire et ne pas dire*. Paris: Hermann.
- Geurts B. (1999). *Presuppositions and Pronouns*. London: Elsevier.
- Grice, H. P. (1975). Logic and Conversation. In: Cole, P., & Morgan, J. (Eds.) *Syntax and Semantics* (vol. 3, pp. 41-58). Academic Press.
- Grosz, B. J., & Sidner, C. L. (1986). Attention, Intentions and the Structure of Discourse. *Computational Linguistics*, 12 (3), 175-204.
- Gündel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive Status and the Form of Referring Expressions in Discourse. *Language*, 69 (2), 274-307.
- Kamp, H., & Reyle, U. (1993). *From Discourse to Logic*. Dordrecht: Kluwer.
- Kievit, L., Piwek, P., Beun, R.-J., & Bunt, H. (2001). Multimodal Cooperative Resolution of Referential Expressions in the DenK System. In: Bunt, H., & Beun, R.-J. (Eds.) *Cooperative Multimodal Communication* (pp. 197-214). Berlin & Heidelberg: Springer.
- Landragin, F. (2002). The Role of Gesture in Multimodal Referring Actions. In: *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces*. Pittsburgh.
- Landragin, F., Salmon-Alt, S., & Romary, L. (2002). Ancrege référentiel en situation de dialogue. *Traitement Automatique des Langues*, 43 (2), 99-129.
- Sperber, D., & Wilson, D. (1995). *Relevance. Communication and Cognition*. Oxford UK & Cambridge USA: Blackwell.
- Wolff, F., De Angeli, A., & Romary, L. (1998). Acting on a Visual World: The Role of Perception in Multimodal HCI. In: *Proceedings of the AAAI Workshop on Multimodal Representation*.