Abstract: For several years there has been a research effort at La.L.I.C. aimed at describing how to use context to identify semantic information. From our point of view, all signs occurring in a text that we call textual context, must be taken into account to determine whether a specific semantic information is present in a sentence or not. We have developed the contextual exploration method which provides the framework for identifying specific semantic information contained in texts. In this paper, we lay the emphasis on internal contexts and give some examples of operational systems using contextual knowledge. We have developed the SECAT system which captures the aspectual value of a sentence in order to understand it and to infer other information that may ensue. We have also developed the SERAPHINE and the SAFIR systems which capture specific semantic information held in texts. The conceptual model and the general software architecture of a contextual exploration system are succinctly described.

Keywords: Text and context, linguistic context, identifying semantic information, contextual exploration method.

1 Introduction

Most of the natural language processing methods ([Fuchs 93], [Saint Dizier 95]) break the process down into several successive stages using several levels of representation: morphologic, syntactic, semantic then pragmatic analysis. These conceptual representational and understanding models of texts depend on contextual knowledge: each stage brings in some specific knowledge about the text being analysed.

The concept of linguistic context ([Choueka 85], [Slator 91], [Desclés 91]) is rather vague and not homogeneous. Operational methods taking context into account differ as they are based on varying contextual frameworks and contents. For a given text understanding model M our approach distinguishes between external context \( K_e \) and internal (linguistic)
context $K_i$ whence the contextualized models $K_e[M(K_i)]$. External context $K_e$ needs different kinds of knowledge about domain, user, conditions of the process, etc. Furthermore, each problem solving task (like for instance automatic, abstracting, knowledge modelling, etc.) has a specific external context whereas internal contexts $K_i$ need only linguistic knowledge about an analysed textual unit. Acquiring, designing and maintaining domain knowledge are time consuming and text understanding systems resorting to such knowledge are highly dependent on them ([Pugeault 95], [Zweigenbaum et ali. 95]). Furthermore, some texts deal with obscure domains on which little information have been gathered. Technological intelligence where new information is supplied by the analysed texts is a good example of this kind of problem. To acquire new concepts and build new relations between them, processing these texts may be necessary. As a result, the domain structure depends mainly on the linguistic units of these texts and their position in the text. In other words, it is text understanding which allows the reader to acquire domain knowledge and to build ontology. In this paper, we lay the emphasis on internal contexts and give some examples of operational systems using contextual knowledge.

2 Using internal context in linguistic

So far, context has mostly been used in linguistics to identify the syntactic category of a token. As a result, this kind of context is limited to a few (two or three) words around the token being processed. Syntactic parsers making use of contextual probability, or applying local grammar rules rely on this feature ([Church 88], [Brill 93]). But to identify semantic information, even partially, in a text, using local context is totally inadequate. For several years there has been a research effort at La.L.I.C. aimed at describing how to use context to identify semantic information. From our point of view, all signs occurring in a text that we call textual context, must be taken into account to determine whether a specific semantic information (see §3) is present in a sentence or not. Furthermore, taking textual context into account calls for a thorough linguistic analysis of grammatical words and a study of textual or discourse categories rather than building lexical data bases with semantic features [Rastier & ali 94] which are necessarily dependent on specific domain knowledge. We have developed a method called contextual exploration method which provides a framework for identifying specific semantic information contained in certain parts of text.

3 Context and the contextual exploration method

Several textual processing tasks, such as knowledge extraction or automatic summarising, may be solved by analysing exclusively linguistic units in the text, provided that their linguistic context is taken into account. For example, cognitive observations of professional summarizers [Endres-Niggemeyer 96] have shown that they use textual, structural, thematic and lexical markers in their search strategies. Furthermore, various linguistic works on text analysis ([Roulet 85, 87], [Charolles 88], [Adam 90]) have shown the interest in identifying and locating linguistic markers and their combinations in order to lend meaning to textual units. We have systematised these observations by taking into account all information about textual tokens (such as contextual word meaning, word location in the sentence, sentence or paragraph location in the text, structuration level of text, graphic signs used in titles) to assign semantic labels to sentences (but not necessarily to each sentence). All these observations about texts have led to some basic assumptions.

- Textual processing calls for identifying and studying the semantics of textual categories involved in texts which are independent of the text domain (medical, economical, technical, etc.). As a result, we have studied causality ([Jackiewicz
96], [Jackiewicz 97]), definition [Cartier 97], thematic announcement [Berri 96a et ali]; this list is not exhaustive and other research is underway.

- Studying textual categories entails the identification of specific linguistic indicators which are relevant clues for structuring semantic knowledge. But lexical identification is far from sufficient. The semantic processing of a linguistic unit depends on others linguistic clues which must be present in the same context in order to solve ambiguity caused by the phenomenon polysemy.

- For each textual category, the contextual exploration method suggests the same methodology: i) identify relevant semantic information and their linguistic indicators ii) identify the span of the textual context C needed to take polysemy into account iii) write procedural steps to find relevant linguistic clues by exploring context C in order to solve latent ambiguities. A contextual exploration system is therefore composed of:
  1) a data base of semantically relevant linguistic indicators;
  2) a data base of linguistic clue solving ambiguity affecting the indicators in their context;
  3) a set of decisions linked with linguistic indicators. A subset of these decisions, the procedural one, applies in the same conceptual representation system, the linguistic units system; Another subset applies in another conceptual representation system, the semantic units system (see §3.1).

  4) a data base of contextual exploration rules; the task of these rules is to identify linguistic indicators in order to trigger procedural or semantic decisions; we have formalised these rules in figure 1. In such rules, \( U_i \) and \( V_k \) are linguistic units and \( C_{ik} \) constitute the contexts which depend on both linguistic indicators and decisions. This is dealt with in more detail in (Berri et ali 95, Berri 96).

\[
\text{LET } U_i \text{ BE a linguistic indicator for the } D_j \text{ decision}
\]
\[
\text{IF } U_i \text{ occurs in a sentence } P
\]
\[
\text{AND IF linguistics clues } V_k \text{ occurs in } C_{ik} \text{ contexts}
\]
\[
\text{THEN perform } D_j \text{ decision.}
\]

Figure 1 : A contextual exploration rule.

- **The contextual exploration method, contextual grammars and indeterminacy**

  Contextual exploration method does not require a thorough syntactic parsing; syntactic tagging is generally sufficient. One must bear in mind that a contextual exploration system has nothing to do with a contextual grammar, like Chomsky's grammar. A contextual grammar rule is written like this:

\[
V+U+W \rightarrow a_1+...+a_n
\]

meaning that a unit \( U \) is interpreted as the sequence " \( a_1+...+a_n \) " provided that \( U \) is located in the context \( V-W \). On the one hand, concatenation and adjacency are two basic concepts of contextual grammars. On the other they are not able to deal with various representation systems. Both of these concepts are irrelevant in a contextual exploration system in which the basic idea is compilation, that is to say, the process of going from a linguistic representation system to a semantic representation system [Descles 90].
Sometimes, analysing the context of a linguistic indicator U, leads to indeterminacy because some contextual clues \(V\ldots W\) suggest taking the decision D, while others contextual clues \(V'\ldots W'\) suggest the decision D'. This is the only case where domain knowledge may be useful.

In the next sections, we describe several contextual systems using various kinds of contexts, ranging from the sentence to the full text, to provide semantic information.

### 4 Sentence and context

The SECAT system ([Maire-Reppert 90], [Berri 96]) has been developed to identify the aspectual values of a sentence in order to understand it and to infer other information that may ensue. Linguistic indicators like tense morphemes are generally insufficient to fully determine aspectual values of a sentence. Let us give an example:

(1a) *Le lendemain il était pris* (The day after, he was captured)

(1b) *Cinq minutes plus tard, le train déraillait* (Five minutes later, the train ran off track)

Depending on other explicit linguistic information provided by the context, one must infer various aspectual values. Hence, in following contexts:

(2a) *Malgré toutes les précautions, le lendemain il était pris* (In spite of all precautions, the day after he was captured)

(2b) *Malgré toutes les précautions, cinq minutes plus tard, le train déraillait* (In spite of all precautions, five minutes later the train ran off track)

(3a) *Sans toutes les précautions, le lendemain il était pris* (Without all precautions, the day after he was captured)

(3b) *Sans toutes les précautions, cinq minutes plus tard, le train déraillait* (Without all precautions, five minutes later the train ran off track)

we get the following aspectual values:

- "new state" inferred from the French "imparfait" for (2a) and (2b);
- "unreal" inferred from the French "imparfait" for (3a) and (3b).

From sentences (2a) and (2b) one infer that "*il a été pris*" (he was captured) or "*le train a déraillé*" (the train ran off track), and from sentences (3a) and (3b) one infers quite the opposite. Therefore, analysing the context, in this case the full sentence, is the only way to solve the indeterminacy. Depending on the linguistic clues *malgré* or *sans*, the relevant aspectual values, respectively "new state" or "unreal" may be assigned. Clearly, this simplistic heuristic may fail\(^1\) and more complex rules are needed to process larger text. For a detailed discussion, we refer the reader to [Maire-Reppert 90].

In the SECAT system, exploring the context of a linguistic indicator means exploring the clause or the whole sentence where such an indicator occurs. A brief summary of the general organisation of the SECAT system is as follows: i) identifying tense morphemes which are relevant indicators; ii) for each processed morpheme, exploring the context to find relevant clues and then applying contextual exploration rules; iii) depending on the information found, assigning an aspectual value or going on analysing the context (step ii)). Finally, temporal relations and enunciative registers are graphically displayed [Berri 96].

---

\(^1\) For example: *Malgré sa mauvaise note en philosophie et sans l'aide de son professeur, il était admis dans la classe supérieure* (In spite his bad mark in philosophy and without help from his teacher, he passed the exam).
5 Text and context

We have developed two systems, SERAPHIN [Berri & ali. 95] and SAFIR [Berri et ali. 96] to capture specific semantic information in texts. "Thematic announcement", "recapitulatory" or "causal argumentation" are the kind of semantic labels assigned to some text fragments. Such lexical or grammatical markers are highly ambiguous but their context, that is to say, the syntactico-semantic links between sentences, sentences and paragraphs location, provide relevant clues to solve ambiguity. Both systems provide one or several extracts, built with sentences picked up from the processed text and for each of the picked up sentence, its assigned semantic label is also provided. A first evaluation of twenty five extracts has been discussed in [Berri et ali 96], and more detailed results will be presented in [Minel et ali 97].
5.1 Contextual identification of conclusive sentences

In the framework of an automatic summarising task, conclusive sentences (or parts of text) are essential because they express explicitly what the reader, from the author's point of view, must memorise. Because unambiguous markers like "notre conclusion" (our conclusion) or ambiguous one like "donc" (consequently) are often mixed in a same part of text, we have built some heuristics rules using textual context to solve this problem. For example, the sentence:

(5) Ceci est notre troisième conclusion (That is our third conclusion) [Bertin Revue de l'Energie]

is a conclusive sentence, but gives no information about the conclusion itself. But the clue "ceci" (that) located at the beginning of the sentence triggers a wider textual exploration, in this case the preceding sentences P_p, where the system continues searching for clues, even ambiguous ones, which express some kind of conclusion. As a result, here is the selected sentence:

(6) Donc, pour que le développement de l'électronucléaire ait une influence significative, il faudra qu'il soit très important.

Furthermore, the clue "troisième" (third), which is called textual organisation clue, triggers the search of a specific part of text Ts, called textual sequencing ([Adam 90], [Charolles 88]) which may be spread out in several paragraphs in order to find conclusive indicators and clues bounded with "troisième" in Ts.

Finally, here is the final selection of the text2:

(7) Une remarque essentielle est que parmi les énergies actuellement utilisées à grande échelle seul le nucléaire et l'hydraulique ne sont pas concernés. Ceci est notre première conclusion. (....)

(8) Notre deuxième conclusion, est que, à cause de l'effet de serre, l'intérêt de développer l'électronucléaire est devenu évident à un certain nombre d'hommes politiques, d'industriels et de scientifiques de disciplines diverses. (....)

(9) Donc, pour que le développement de l'électronucléaire ait une influence significative, il faudra qu'il soit très important. Ceci est notre troisième conclusion. (....)

(....) (BERTIN Revue de l'Energie)

Let us point out that underlining or modal markers and their position in the text solve the ambiguity of connective markers in (7, 9) and such textual sequencing must be in the first or last section of the text to insure that these are global conclusions in the author's argumentation. As we said before, these linguistic indicators and clues (they appear in bold in these sentences) are not domain dependent and our observations confirm that they are used in technical, economical, and political fields.

5.2 Using context to distinguish between causality and agentivity

In the framework of the SAFIR project which aims at filtering textual information, we have pointed out the importance of the relation of cause and effect in this kind of task [Jackiewicz 1996]. The causality relation is used to organise systematically empirical facts and to give them some meaning. Textual information providing a relation of causality is precious both for pure knowledge and from a practical point of view. Causal knowledge must first be built,

---

2 The translation in English of a French paragraph does not make sense (from a linguistic point of view) because, the tense of verbs, lexical or grammatical clues are quite different in these two languages.
then applied in some practical ways. Accordingly, we have studied the means in natural language to express a causal link between two situations. We assume that some textual data illustrate both how the author builds this kind of relation (causal judgement supported by a body of arguments, statistical study) and the conditions of his knowledge about it. This is how the factual link between two situations is expressed by the author and how the core of the causality concept is formed. This allows us to distinguish it from the agentivity concept.

These two concepts, causality and agentivity, are closely related both on the conceptual level and on the discursive level. But, as the difference between them does not appear clearly in the verbal lexicon, we have showed that the textual context is a way to distinguish them. For example, French verbs like provoquer (to bring about), entraîner (to entail), contribuer à (to contribute) usually express causal situations, while verbs like créer (to create), gêner (to impede), aider (to help) refer to effects of agentive actions. It is not always possible to distinguish between causal sentences and agentive ones, relying only on such verbal data. Moreover, arguments accepted by these verbs do not provide any useful information because they are so varied in terms of syntactic structure and semantic content.

The attempt at understanding this difference results in building a causal link. Some arbitrariness and even deep indeterminacy may follow since the subjective uncertainty (due to insufficiency of our knowledge) may add to the objective indeterminacy existing in reality. Consequently, and especially in scientific texts, causality occurs with many kinds of terms setting out the causal link as certain, possible, unreal, proved to be true by some experiments, calculations, expert advice, etc. We have identified therefore these kinds of clues to find causal sentences in texts.

As shown in (10,11), verbs créer (to create) and perturber (to disrupt) bind together situations existing separately. This is a main characteristic feature of causality: a link is built between facts which might exist and be apprehended distinctly (this is not true in agentive sentences like Jean gêne le passage Jean impedes the passage). But this theoretical criterion is not useful from the natural language processing point of view. In contrast, the underlined contextual markers illustrates (risquer (to risk), conditional tense) the possibility of the causal link as well as the author’s way of taking it into account (les climatologues pensent que, climatologists think that).

5.3 Different kinds of context
Four kinds of contexts have been identified in SERAPHIN and SAFIR whenever an indicator occurs in a sentence S:

- The context $C_1$ must be explored to assign a semantic label to a sentence; the exploration is limited to the clause or the whole sentence S;
- The context $C_2$ must be explored to solve dangling anaphora; the exploration is limited to sentences (belonging to a paragraph P) located just before the sentence S containing the anaphoric clue; in any case, the beginning of paragraph P stops the exploration;
- The context $C_3$ must be explored to identify textual sequences; this context is constituted by one or several sentences before and after the sentence S. This context is limited, either by the beginning of a textual section (a set of SGML tagged paragraphs), or by a textual organisation clue;
- The context $C_4$ must be explored to identify textual segments formatted with special cues as bullet lists, enumerated symbols, etc.

6 General software architecture

We succinctly describe the conceptual model and the general software architecture of a contextual exploration system (Figure 2). This is dealt with in more detail in ([Berri et al. 95], [Berri 96]). Linguistic indicators and clues have been organised in non-disjunctive classes in a data base; contextual exploration rules have been broken down into production rules associated with tasks. Let us point out that the identification of some linguistic features, for example, dependencies between a verb and its arguments, remains unsolved, especially in heterogenous text. Consequently, heuristics based on textual observations have been developed.

Text processing is performed in several steps:

1) The source-text is marked in accordance with the SGML standard so that some selection rules that take into account the position of the sentence in the text and the dependence relations between various entities of the text (sections, sentences, etc.) can be used. Relying on these SGML tags and applying some heuristic rules (to deal with acronyms, references, etc.) real sentences are identified and a object model of the text is built.

2) One module identifies the contextual exploration clues. This process has to cope with two problems: the recognition of morphological variants and the discontinuity of composite markers.

3) One labelling module assigns relevant semantic labels to sentences or parts of texts, by applying several rules. A sentence may get several labels. Possible conflicts are solved by the following module.

4) For each specific textual task, a module chooses from the assigned semantic labels and takes into account users needs.

7 Conclusion

We have shown in this paper that a thorough study of the semantics of linguistic markers, in their textual context, provides a reliable means of identifying some discursive categories. Furthermore, this approach does not need any heavy linguistic resources and is compatible with an incremental development. On the other hand, the linguistic knowledge acquisition task requires a thorough study of textual corpora. Furthermore, we think that the contextual exploration method has cognitive implications. Such implications have been shown by using the same methodology with Bulgarian
texts ([Ivanova 94] [Guentchéva 94]) [Stankov 94]) and a study of English texts is underway. Therefore, this textual linguistic work relying on the systematic study of certain speech acts (how to define, how to conclude, etc.) in their real lexical and syntactic textual context, offers a plausible alternative to the existing theoretical approaches, that are more ambitious, but less effective.

![Diagram](image)

Figure 2: General software architecture.

### 8 Bibliography


