

# **A mark up language for tagging discourse and annotating documents in context sensitive interpretation environments**

**Graziella Tonfoni**

**The George Washington University  
Declassification Productivity Research Center (DPRC)**

**Ashburn, VA 20147**

**tonfoni@seas.gwu.edu**

**and**

**DSLO-University of Bologna-Italy**

**tonfoni@alma.unibo.it**

## **Abstract**

A mark up language for tagging discourse, and for converting discourse sequences into a written format, according to highly context sensitive procedures, will be illustrated as well as a system for document annotation. The context in which a given communicative intercourse has taken place needs to be made available to ensure consistent interpretation of both single sequences of discourse and global concepts that are carried along during the dialogue or conversation. Full visibility of different communicative intentions that reflect the evolution of conversations in time and space, as well as access to various modes of communicative actions and changing conditions of interpretation, is relevant and necessary, especially in contexts where interaction is based upon asynchronous communication. Interpretation shifts, that each sequence of a dialogue is likely to undergo, may create distortions in the interpretation of the overall intercourse and communicative action leading to the creation of the final document. Some contextual concretions, which may be based upon false assumptions, are particularly powerful. Persistent interpretative links

may be evoked and activated at any time, even if unintentionally. Special attention needs to be paid to ensure that undesirable links are not established and unintentional contextual concretions are not added. Use of a consistently applied, commonly shared conceptual tool for assigning context-sensitive interpretative values to each sequence of discourse holds great promise for avoidance of this problem. Appropriately packaged documents and parts of documents could carry along their own originating context of discourse in the form of attached information. Accurate illustrations using a fully functional set of tools I have developed are provided here to show how fuzziness and misinterpretation (caused by an absence of consistent interpretative clues about originating contextual conditions for discourse and conversations) may be significantly reduced or even eliminated.

## **Introduction**

A document comes from "somewhere in time and space and leads toward somewhere else"(Tonfoni,1998).

It may therefore be defined as a piece of information that has been derived from a


dynamically evolving information flow before it is converted into a stable form, e.g., hardcopy (Tonfoni 1996, 1998). Documents are derivative products of flows of conversations and various kinds of communicative intercourse, which may include a very high level of complexity and long duration. Documents and conversations, from which those documents were generated, are therefore two very tightly linked components that often play a crucial role in providing evidence for decision making. It is our claim that enhanced encoding procedures in the form of discourse tagging and labelling may be harmoniously linked by means of a consistent annotation system (Tonfoni 1998) to support accurate conversion of dialogues and discourse into a more stable format. This is what documentation is all about. The discourse tagging system presented here is based on and harmonious linked to an annotation system, which consists of a set of signs and symbols as follows:


- **Discourse tagging and document annotation signs:** to indicate the *communicative function* of a sequence of discourse, which is ultimately to become a piece of a document.
- **Discourse tagging and document annotation symbols:** to indicate the *communicative style* of each sequence of discourse, which is ultimately to become a piece of a document.
- **Discourse tagging and document annotation turn taking symbols:** to indicate *roles and interplay* between the discourse partners that are carried along during the information conversion process and successively attached to the resulting document.


Context sensitivity may be significantly enhanced by the consistent use of interpretation devices, designed to prevent fuzziness and misunderstanding from occurring. Some contextual links, if not properly handled, may be powerful enough to radically shift the scenario. The originating context may in fact be easily modified and completely distorted, even if unintentionally. Such links need to be accurately identified and then eliminated by repositioning, either by reassessing the originating context or assessing the new and intended one.

**A context sensitive mark up language for converting discourse sequences into document pieces.**

**Discourse tagging and document annotation signs:** The following represent the various communicative functions a document may convey, on a paragraph by paragraph basis, as a result of consistent conversation of discourse sequences into document pieces.

 **Square:** for an informative document or piece of a document, which carries information about a specific conversational event. Indicates that information conversion has been derived from an informative discourse sequence.

 **Square within the Square:** for a summary of a given document that has been produced to reinforce contextual consistency between the conversational context in which the discourse first occurred and its conversion into a larger document.

 **Frame:** for a document or piece of a document that is found to be analogous (in content) to other

documents which refer to previously stored information, some of which may still be available in discourse format. Normally, conversion occurs from discourse format into document format.



**Triangle:** for a memory and history generated out of a certain document. This is meant to establish topical continuity with background information, which may still only be available in its discourse format and still need to be converted into document format.



**Circle:** for a main concept conveyed by a certain document, which has been abstracted and linked to other documents, to show topical continuity. It is meant to reinforce topical word identification and to effectively link together documents with the same word and sequences of discourse prior to their conversion into document format. Both discourse and document (may) use the same topical words.



**Grouped Semicircles:** for main concepts, which are abstracted out of an originating document. Establishes both topical continuity and context consistency between the originating document, and a set of topical words and links to sequences of discourse prior to conversion into a document format. Here again, both discourse and document (may) use the same topical words.



**Semicircle:** for a locally identified concept, abstracted out of a piece of document and meant to reinforce context consistency by establishing further links to other documents. These links may be triggered by the same topical word. It is also meant to

trigger sequences of discourse prior to conversion into a document format and using the same topical words.



**Inscribed Arcs:** for indicating the need for an upgrade and/or update of a certain document. Indicates that a revision process is likely to occur, although it does not identify if it will be a major or minor revision. Revision may be based on conversion of discourse sequences into additional pieces of a document or into various alterations.



**Opened Text Space:** for indicating that an upgrade and/or update has indeed occurred and that the document has now reached a new revision state. It is meant to show that the structure of the previous document has been affected by discursive information, but does not identify if the revision has been a major or a minor one..



**Right Triangle:** for a comment made about a given document or piece of a document, for the case where more contextual information is needed. This information is not available in a document format and has to be derived from other external relevant sources based upon topical continuity.

**Discourse tagging and document annotation symbols** are used to indicate communicative intentions and styles, locally within the discourse, such that discourse sequences can be consistently and accurately interpreted with the the additional information they provide. They are particularly useful for showing contributions made by individuals in either synchronous or asynchronous conversations, and for supporting co-ordination and information conversion for production of a document.

These information-containing elements may be conveniently incorporated into the final document to provide clues about the nature of the original information conversion process. Document annotation symbols, therefore, represent different modes of information conversion (from a discourse) which may be packaged with the originating context, and activated at a later time. They may be combined and used dynamically for further information conversion purposes, such as further discussions, because they effectively indicate transitional states within a discourse that may be evolving in time and space. They are of the following types:



**Describe:** from Latin *describo*: write around.

It means complementing the original discourse or document by adding as much relevant information as may be found from previous discourse, without any specific constraints.

It may also indicate the need for further information to be put together in discourse format or in document format. It is represented by a spiral, which starts from its middle point, to indicate a flow of information from topical words, towards an expanding topic and linking with other information. The other information can come from different discourse sequences or from other relevant documents or pieces of documents.



**Define:** from Latin *definio*: put limits.

It means complementing the document or the discourse with limited information about a defined topical word, which has been previously selected and identified as the most relevant. The point in the middle of the square represents relevance. The concept indicates that there is a real need to incorporate available, highly specific information about a relevant discourse or document.



**Narrate:** from Latin *narro*: tell the story.

It means complementing the discourse or the document with various facts and events (from the originating context) by following a logical and chronological order. They may be used either in the form of discourse or in the form of the document itself. In other words, it indicates a set of major points or facts representing different diachronic stages, which are closely linked.



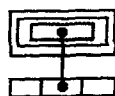
**Point out:** take a single point out of a story chain.

It means to isolate a specific event or fact (among those reported as discourse or occurring within a document) by focusing on just one sequence or a piece of document, and adding more detailed information. Information is added by as a significant expansion and linked with other relevant discourse sequences or documents.



**Explain:** from Latin *explano*: unwrap, open up.

It means that facts and reasons are given to support interpretation of an event, within a certain discourse or document. Explanation may start by indicating the originating cause and proceeding logically toward the effects or by starting with the effects and going backwards towards the cause, depending on which approach is found to be the most effective.



**Regress:** from Latin *regredior*: go back.

It means that more information about a topic, presented during a given discourse sequence or within a document, is absolutely necessary for understanding. Information may come in verbal format and then be converted into document format. It represents a topic-oriented process and an in depth information expansion, which is activated only for the precise topic being considered.



**Inform:** from Latin *informo*: put into shape, shape up.

It means that any discourse and document is the result of an information packaging process, and that the specific discourse and document under consideration is organised in the most unconstrained way, as the result of many information conversion operations.

It leads toward two different kinds of further specification, which are respectively conveyed by the "inform synthetically" and the "inform analytically" indication.



"inform synthetically" means departing from a larger discourse or document

and proceeding toward a summary (related to a specific topic) which is the most relevant one emerging from the originating discourse and document .



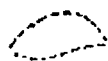
"inform analytically" means departing from a limited discourse and document and expanding toward further discourse sequences and documents by adding more information, which needs to be then converted into the final form of a document, and is not available yet.



**Reformulate:** from Latin *reformo/reformulo*: change

shape and reshape.

It means changing the style, which was previously adopted, either in the discourse or in the document, and substituting one form of information packaging with a different but related (same discourse or document) packaging. It may turn produce (a more or less) radical transformation of the originating context, according to a precisely defined request or set of requests.



**Express:** from Latin *exprimo*: push out and press out.

It means adding personal opinions and individual feelings related to facts and

events during a discourse or within a document. It indicates the most subjective mode of information organisation, which may be clearly influenced by and bound to personal evaluations, judgements and emotional states.

**Discourse tagging and document annotation turn taking symbols** are meant to define the mode of interpretation of a certain discourse and of accessing a certain document, requested at each given time; they are provided to facilitate accurate context transport.

They are the following ones:



**Major Scale:** it shows that literal interpretation is needed and that those sequences of discourse or pieces of documents indicated and marked off, should be extracted and quoted literally, the way they were first intended to be.



**Minor Scale:** it shows that accurate interpretation may need a further process of abstraction and that sequences of discourse and pieces of documents indicated and marked off, may need significant interpretation processes due to heavy context constraints.



**Open or Unsaturated Rhythm:** it shows that accessing the discourse and document at the present stage may lead to an incomplete interpretation of those facts and events, which are presented. It is meant to suggest accessing a broader discourse and larger document and acquiring many and various kinds of sources, some of which may not yet be available.





**Tight or Saturated Rhythm:** it shows that accessing the document will lead the user toward complete interpretation of those facts and events, which are presented. It suggests

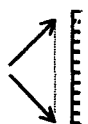
that the user should hold fast to the interpretation provided, though access to other sources of supportive evidence is also available.


**Discourse tagging and document annotation amplifier symbols** come last and may be added only after the previously illustrated ones have been used. They apply to larger discourse sequences and documentation territories and indicate specific operations, which are to be performed in order to connect conversational actions and sets of documents, which have been previously encoded and accurately stored.

They are as follows:


 **Choose:** it is meant to represent the dynamic process of first identifying and then deciding between different contexts for interpretation, which are mutually exclusive, given a certain set of conversations, which have occurred and documents, which have been derived accordingly, but seem to have different possible evolutions.


 **Identify:** it is meant to represent a definition of a more specific context within a broader context, for interpretation of a set of conversations, which have occurred and for documents that have been derived. It naturally occurs before "search" and "select".


 **Search:** it is meant to represent the dynamic process of choosing among different contexts for interpretation of a set of conversations, which have occurred and documents, which have been derived and are many and compatible as to find the most appropriate one.


 **Select:** it is meant to represent multiple contexts that may evolve (either synchronously or

asynchronously) and may be modified, once a certain decision making process has been performed. It is based upon a certain discourse that was converted into a document and stored.

 **Copy/Replicate:** it is meant to represent the dynamic process of duplication and repetition of a context, which if lost, would affect understanding and accuracy of interpretation of events and facts that have been organised, first in the form of discourse and then converted into a document.

 **Ahead:** it is meant to represent the progression of a conversation to become a document or set of documents, which are linked together by context consistency or harmoniously shifting contexts.

 **Back:** it is meant to represent the need to go back to delete and replace the originating context, which has radically shifted, in the course of various transition states, during an ongoing conversation such that, if not eliminated, would indeed affect consistent interpretation of a whole set of documents, which are based upon it.

 **Conflict:** it is meant to represent an emerging inconsistency and incompatibility between various context attributions to a set of conversations to be converted into documents, the context needs to be cleared as to proceed toward any further interpretation.

The discourse tagging and document annotation system we have illustrated here (including its various components) may be applied at different layers and at various levels of complexity.

Following this perspective, discourse context may be enhanced through visual clues or symbols to provide a very powerful means to monitor inherent complexity of any communicative intercourse and to reduce possible distortions which may occur.

### **Conclusions**

This system for tagging discourse provides consistent and harmonious linkages to the original context in the form of a mark up language for visually annotating documentation. It has been extensively and intensively tested in many and various context and languages. The acronym, CPP -- TRS stands for Communicative Positioning Program -- Text Representation Systems.

What is therefore indicated is that icons representing precise operations performed upon texts, both in their verbal and written dimensions, carry the effective intentionality to be encapsulated. Encoding and pre-programming a document means precisely complementing a document with all of the instructions that are necessary to enhance understanding of context. Visual programming of a document (based on

previous discourse) provides the ability to categorise and classify information and to carry the full details of a context into the (delicate) process of information conversion to ensure the most reliable final product, e.g., the document.

### **Acknowledgements**

The author sincerely thanks Dr. Richard Scotti, Director of the DPRC, for reading this paper, prior to its submittal, and more for having captured the relevance of the most important features of her system and approach, creating the appropriate conditions for producing practical applications of it in the real world.

### **Basic references**

- Tonfoni, G.**, 1996, *Communication Patterns and Textual Forms*, Intellect, Exeter, U.K. .
- Tonfoni, G.**, 1998, *Information Design: The Knowledge Architect's Toolkit*, Scarecrow Press, Lanham, Maryland, U.S.