Discourse structure annotation: Creating reference corpora

Alexandre Labadié, Patrice Enjalbert, Yann Mathet, Antoine Widlöcher

Discourse structure annotation: Creating reference corpora

Alexandre Labadié, Patrice Enjalbert, Yann Mathet, Antoine Widlöcher

GREYC, Université de Caen
{firstname.lastname}@info.unicaen.fr

Abstract

In this paper we address the problem of defining standards for discourse-level text annotation. We propose the URS metamodel as a common formalism in which existing models and models-to-be could be (re)formulated. This meta-model should permit to create standard annotated corpora in a consistent way, i.e. allowing alignment of various discourse structures. We also present the annotation platform Glozz supporting the URS metamodel. The approach is illustrated by an on-going campaign devoted to both topical and rhetorical structure of texts.

1. Introduction

One can currently observe an increasing interest for discourse structure analysis in the NLP community, both for applicative purposes (improvement of document indexation, text summarization, document browsing...) and corpus based linguistic studies. A great variety of phenomena and models are investigated: lexical cohesion (Morris and Hirst, 1991), coreference chaining, discourse moves à la Swales (Swales, 2004), discourse framing, rhetorical structure within models such as RST (Mann and Thompson, 1988), SDRT (Asher, 1993), LDM (Polanyi and van den Berg, 1996), etc. If it is clear that no "universal model" is (at least for the time being) to be expected or even challenged, it is also very clear that all these phenomena cooperate to produce the structure of discourse, and the various models capture different aspects of this structure.

Several works and scientific events in the NLP community reveal the need for standards, guidelines, methodologies and tools to ensure the long-term availability and interoperability of resources. As well as LRT workshop, let us just mention, for example, the XBRAC workshop (Witt et al., 2004) or the ACL workshop on Discourse Annotation (Webber and Bryon, 2004). Standardization projects, such as *Lirics* for example, also exist but mainly focus on morphosyntactic informations. A consequence is an urging need for the elaboration of some common ground on which models can be compared and combined, and their interactions observed.

The result should be 1) an abstract framework in which many or most (if not all) discourse models could be (re)formulated, 2) a text annotation format implementing this framework and a set of tools allowing to mine these annotations, and 3) a set of reference corpora relative to different models coded in this common format. The work presented here is a first attempt in this direction[1].

2. The Unit-Relation-Scheme metamodel

Ad hoc models, dedicated to specific linguistic objects, make it difficult to study interactions between discourse phenomena, and to compile and compare annotations. Originally coming from (Widlöcher, 2008), the Unit-Relation-Scheme (URS) metamodel is an endeavor to propose a common formalism in which existing and models-to-be could be (re)formulated. This metamodel relies on three classes of *elements*: *units*, *relations* and *schemes*.

2.1. Metamodel and models

Within the general framework defined by the meta-model, specific models can be expressed. Such a specific model declares available *elements* as well as their expected properties. All *elements* (units, relations or schemes) are characterized by a *type* and a variable number of *features*. Type names, and available features for a given type, depend on a user-defined specific model. Feature values can be free or limited to predefined values.

2.2. Elements

Unit A *unit* is a textual segment, i.e. a text sequence of any size from one character to the whole document. For example, in a discursive perspective, a *topic segment* could be represented by a unit, whose feature set could represent its topic (most of the time a small piece of free text).

Relation A *relation* designates a link (oriented or not) between any combination of two *elements*.

RST- or SDRT- discourse relations, such as *elaboration* (directed) or *contrast* (undirected) are typical examples. They might relate units (such as single clauses) or schemes (representing "complex" discourse objects).

Scheme Both previous elements are quite common in most of existing models (even if designated otherwise). The scheme is less conventional and more specific to the URS metamodel. A scheme is a complex recurring textual configuration, or pattern that can link together any number of units, relations or even other schemes.

Enumerative structures (Ho-Dac et al., 2009) provide a good example of schemes. They are composed of a set of consecutive *items* units, whose enumeration is usually embedded in a larger structure, introduced by a *header*, thereby linked to the block of items by an *introduction* relation. Furthermore, *similarity* or *contrast* relations link items to one another.

[1] ANNODIS project, supported by the Agence Nationale de la Recherche (Péry-Woodley et al., 2009)

3. Glozz annotation platform

Despite the availability of various annotation tools such as UAM Corpus Tool[2], Gates's manual annotation module (Cunningham et al., 2002), Wordfreak (Morton and LaCivita, 2003), Protégé's *plugin* Knowtator (Ogren, 2006), MMAX (Müller and Strube, 2001), PALinkA (Orăsan, 2003) or RSTTool (O'Donnell, 2000), it must be noticed that annotation tools are often devoted to a particular theory or to a specific class of linguistic objects. Consequently, different requirements, in particular in terms of abstraction and genericity, are over-all not satisfied.

The general-purpose Glozz platform[3] (Widlöcher and Mathet, 2009) takes these constraints into account and implements the URS metamodel, with no prior hypothesis on the studied and annotated phenomena. It provides a graphical and highly configurable annotation environment (figure 1), usable for corpus exploration of various linguistic phaenomena.

Annotators are given an access to textual content and to any linguistic information which may result from preliminary (manual or automatic) annotations, and may produce new annotations.

The annotations must conform to the categories defined by

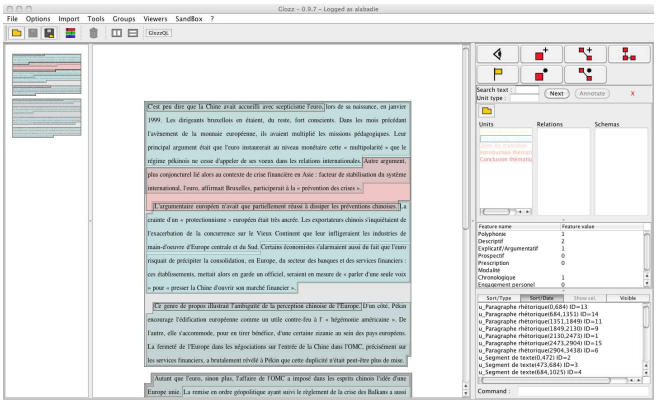

Figure 1: Glozz interface

the meta-model and presented above. In addition, they have to conform to an *annotation model* beforehand defined for the annotation campaign, which specifies allowed linguistic objects and expected feature sets. This specific linguistic model is described by an XML document, which configures the Glozz application.

Annotations are stored in a XML standoff format. Glozz also provides various mining tools. In particular, the Glozz Query Language (GlozzQL) makes it possible to express constraints on the annotations and their properties, in order to select and easily observe relevant objects, according to the specific aims of the campaign.

4. Applications: rhetorical and topical structure

Be it in NLP or in corpus linguistics, there is a lack of reference corpora in the area of discourse structure analysis. Using the URS metamodel and its annotation platform Glozz, several annotation campaigns are in progress as contributions in order to fill this gap. One was already evoked in section 2, relative to Enumerative Structures The one described here is twofold and concerns the dual notions of *rhetorical* and *topical* structure of texts. Beyond investigation of some specific aspects of these structures (to be described below) an important goal is to study the way they *combine*.

4.1. Rhetorical tagging

Rhetorical structure can be considered from two general viewpoints. One, which can be characterised as bottom-up and relation-oriented, aims at discovering discourse relations between elementary text units (at propositional level); it is notably represented by the RST, SDRT or LDM models.

The other approach, the one we want to investigate, rather coarse-grained and segment-oriented, is interested in discovering segments of texts filling different communicative functions. Such segments can be defined in different ways. One, following (Swales, 2004) is based on the notion of discourse moves, which are "parts of the message" specific to some specific genres. For example, in scientific articles, we find segments relative to: context of the study, aim and hypotheses, experiment, results and discussion. In NLP, such a model has been notably worked out in S.Teufel's pioneer work (Teufel and Moens, 1999) and, in a slightly different perspective, by (Biber et al., 2007). Another approach, in a sense more "universalist" is the classical *Narrative-Description-Argument* model (Adam, 2005), (Smith, 2001).

The model we adopted for the annotation campaign is rather of the latter type and we adopted the term of *discourse mode* proposed by C. Smith to denote the different "rhetoric types". The corpus in view is composed of journalistic texts from *Le Monde*. This choice is due both to applicative goals (summary, document browsing) and to the linguistic quality of articles in this newspaper.

However we won't ignore that, as Swales points out, the specificities of a genre have to be taken into account, and decided to adapt the Narrative-Description-Argument trilogy to fit our corpus. Moreover, we make the observation that several discourse modes are generally simultaneously present in a portion of text: for example description is intertwined with argumentation, or with narration.

The task of *rhetorical tagging* can then be described as follows. We make the hypothesis that paragraphs can be considered as relevant textual units. Rhetorical tagging of paragraphs consist in identifying which are the main discourse modes, insisting in the plural for the reasons above.

Annotators have to allocate a score to seven fields representing the seven discourse modes we decided to keep for this campaign. Paragraphs are already delimited and annotators only have to allocate the scores.

The discourse modes we chose can be divided into two main dimensions: The representational (or ideational) and the interpersonal one[4].

[2] http://www.wagsoft.com/CorpusTool/.

[3] http://www.glozz.org/

[4] The term "representational" is inspired by Adam's terminology and "ideational" by Halliday's one.

The representational or ideational dimension concerns the semantic content of the message, the representations construed by the reader. The different fields are strongly related to different modes of internal coherence. The four ideational fields are:

Description: Indicates the weight of factual information in the paragraph.

Argumentation / Explanation : Represents to which extent the paragraph is about convincing or explaining something to the reader. We considered that the mechanisms of argumentation and explanation are the same, even if the goals are not.

Chronology: Indicates the weight of chronological information in the paragraph.

Prospection: Represents to which extent the paragraphs project the reader into the future.
The interpersonal dimension concerns the relation between the writer and the reader in the communicative process. The three interpersonal fields are:

Personal commitment: Indicates if the paragraph holds some of the author personal opinion.

Prescription: Represents to which extent the paragraph is about advising or instructing the reader to do something.

Polyphony: Indicates the weight of directly or indirectly reported speech in the paragraph. Other people view reformulated by the author are included in polyphony.

Each of the seven fields is given a score between 0 and 2. **0:** The discourse mode is absent or too marginally present to be important; **1:** The discourse mode is present, but not key to understand the information conveyed by the paragraph; **2:** The discourse mode is key to understand the paragraph.

4.2. Topic zoning

As topic segmentation is quite a popular task in NLP, a wide variety of methods have been used to challenge it. Most basically admit that a topic segment is a text unit which is thematically consistent and thematically distinct from the previous and next segments. Some concentrate on finding boundaries (Beeferman et al., 1999), others try to regroup consistent part of the text (Choi, 2000). But, be it for training or evaluation, these methods mostly use corpora made from aggregated small texts, assuming that finding boundaries between small texts is similar to finding topic boundaries inside a text. This assumption is regularly criticized (Bestgen and Pirard, 2006), (Labadié and Prince, 2008). This lack of corpora and annotation processes leads us to propose our own topic zoning annotation model.
Our topic zoning annotation model is based on the hypothesis that, in a well constructed text, abrupt topic boundaries are more the exception than the rule. So we introduced the notion of transition zones between topics that help the reader to slip from a topic to another.
The "perfect" transition zone should be composed of two parts: the conclusion of the previous topic and the introduction of the beginning one. But, even in well formed text, there are cases of abrupt boundaries or incomplete transition zones. The topic zoning annotation process consist in delimiting four kinds of unit:

Topic segments: Even if we assumed that transition between topics should be "fuzzy", we need to delimitate topic segments. These segments should represent the main topics of the text, and, if relevant, some subtopics. We decided to have an top-down approach and not to go deeper than the first subtopic level to stay close to our goal of a global representation of the text structure.

Introductions: Whenever annotators meet with a clear introduction of the segment topic at the beginning of a segment, they should delimitate it.

Conclusions: As for introductions, they should delimitate conclusions each time they meet with obvious clues.

Transition zones: When there is a conclusion followed by an introduction, the transition zone is clearly identified. But, sometimes, one or both of them could be absent. Still, the author might have made a transition between the two topic. The transition zone unit is here to identify these parts of the text where a reader feels that there is a transition, but without clues of introduction or conclusion.

4.3. Mining the structures

The presented annotation models are currently used in an annotation campaign. 5 annotators are working on a total of 30 texts from the daily newspaper *Le Monde* that range from 418 words for the shortest to 3745 words for the longest. In the end, each text will be annotated three times on both tasks (rhetorical tagging and topic zoning).
Goals of this annotation campaign are multiple:

About agreement: In both rhetorical tagging and topic zoning, the subjectivity of the annotator can influence the result. As each text is tagged three times, we aim to measure agreement between annotators on both these tasks. By doing so we can evaluate if such a reference corpus is worth building at a bigger scale.

On rhetorical tagging: One big challenge is to constitute collections of linguistic indices responsible for the different modes, a task for which manual annotation will give a firm and indeed inescapable ground. Another goal is to evaluate our hypothesis of intertwining of modes and nevertheless to see how some global segments may appear.

On topic zoning: Apart from the agreement measure between annotators and the information about the subjectivity of topic segmentation we will gain through it, the corpus could be used mainly for three task: - The topic segment part could be used as a base reference to evaluate and / or train automatic methods for topic segmentation / detection. Such methods are mostly trained or evaluated on corpora of aggregated texts due to the lack of such resources, and so results on actual topic segmentation are somewhat disappointing.
- Most of unsupervised topic segmentation methods, such as $c99$ (Choi, 2000) are based on lexical cohesion[5]. The topic segments of this corpus can be used to evaluate the

[5] As defined by (Morris and Hirst, 1991)

lexical cohesion of human built segments and their distance with each others. This can help us to study to which extend the lexical cohesion principle is part of the segmentation process.
- The transition zone part will help us to validate (or invalidate) our hypothesis that boundaries between topic segments are more fuzzy text fragments than abrupt cuts. If validated, we would use this corpus to learn patterns (lexical, syntactic, etc.) specific to these zones.

On both tasks: But, the most interesting result is probably the link between rhetorical and topical structures. The link between "what is said" (the topical structure) and "how it is said" (the rhetorical structure) is intuitively admitted, but not measured. If these two phenomena are linked with each other, this corpus will allow us to see it.

5. Conclusion

We presented here our attempt to establish a standard in discourse annotation allowing to compare many discourse level linguistic phenomena. Through the use of the URS metamodel and its dedicated tool Glozz in an actual annotation campaign, we gave some hints that it can be used as a standard. For the time being, only french corpora are being annotated, but we hope to see annotations project in other languages.

6. References

J. M. Adam. 2005. *la linguistique textuelle. Introduction l'analyse textuelle des discours*. Armand Colin.

N. Asher. 1993. *Reference to Abstract Objects in Discourse: A Philosophical Semantics for Natural Language Metaphysics*.

D. Beeferman, A. Berger, and J. Lafferty. 1999. Statistical models for text segmentation. In *Machine Learning*, volume 34, pages 177–210.

Y. Bestgen and S. Pirard. 2006. Comment évaluer les algorithmes de segmentation automatiques ? essai de construction d'un matriel de réfrence. *Actes de TALN'06*.

D. Biber, U. Connot, and T. A. Upton. 2007. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. John Benjamins Publishing Co.

F. Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. *Proceeding of NAACL-00*, pages 26–33.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Meeting of the Association for Computational Linguistics*, pages 168–175, Philadelphia, USA.

L.-M. Ho-Dac, C. Fabre, M.-P. Pry-Woodley, and J. Rebeyrolle. 2009. A top-down approach to discourse-level annotation. *Corpus Linguistics Conference*.

A. Labadié and V. Prince. 2008. Finding text boundaries and finding topic boundaries : two different task ? *Proceedings of GoTAL 2008*.

W. C. Mann and S. A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. pages 243–281.

J. Morris and G. Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:20–48.

T. Morton and J. LaCivita. 2003. Wordfreak: An open tool for linguistic annotation. In *Proceedings of Human Language Technology (HLT) and North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 17–18, Edmonton, Canada.

C. Müller and M. Strube. 2001. Mmax: A tool for the annotation of multi-modal corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Seattle, WA, Etats-Unis.

M. O'Donnell. 2000. Rsttool 2.4 – a markup tool for rhetorical structure theory. In *Proceedings of the International Natural Language Generation Conference (INLG'2000)*, pages 253 – 256, Mitzpe Ramon, Israel, 13-16 June 2000.

P. V. Ogren. 2006. Knowtator: A protg plug-in for annotated corpus construction. In *Proceedings of Human Language Technology (HLT) and North American Chapter of the Association for Computational Linguistics (NAACL)*, New-York, tats-Unis.

C. Orăsan. 2003. Palinka: a highly customizable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog*, pages 39–43, Sapporo, Japan, July, 5-6.

L. Polanyi and M. H. van den Berg. 1996. Discourse structure and discourse interpretation. In *University of Amsterdam*, pages 113–131.

M.-P. Péry-Woodley, N. Asher, F. Benamara, M. Bras, P. Enjalbert, C. Fabre, S. Ferrari, L.-M. Ho-Dac, A. Le Draoulec, Y. Mathet, P. Muller, L. Prévot, J. Rebeyrolle, M. Vergez-Couret, L. Vieu, and A. Widlcher. 2009. Annodis: une approche outille de l'annotation de structures discursives en corpus. *Actes de TALN'09*.

C. S. Smith. 2001. Discourse modes: aspectual entities and tense interpretation. *Cahiers de Grammaire*, 26:183–206.

J. Swales. 2004. *Research Genres: Exploration and Application*. Cambridge University Press.

S. Teufel and M. Moens. 1999. Discourse-level argumentation in scientific articles: human and automatic annotation. *In Proceedings of ACL-99 Workshop "Towards Standards and Tools for Discourse Tagging"*, pages 84–93.

B. Webber and D. Bryon, editors. 2004. *Proc. of the ACL 2004 Workshop on Discourse Annotation.*, Barcelone, Espagne.

A. Widlöcher and Y. Mathet. 2009. La plate-forme glozz: environnement d'annotation et d'exploration de corpus. *Actes de TALN'09*.

A. Widlöcher. 2008. *Analyse macro-sémantique des structures rhétoriques du discours - Cadre théorique et modèle opératoire*. Ph.D. thesis, Université de Caen Basse-Normandie, 17 octobre.

A. Witt, U. Heid, H. S. Thompson, J. Carletta, and P. Wittenburg, editors. 2004. *Workshop on XML-based richly annotated corpora (XBRAC)*, Lisbonne, Portugal, 29 mai. Confrence LREC 2004.