PALinkA: A highly customisable tool for discourse annotation

Constantin Orăsan
Research Group in Computational Linguistics
School of Humanities, Languages and Social Sciences
University of Wolverhampton
United Kingdom
C.Orasan@wlv.ac.uk

Abstract

Annotation of discourse phenomena is a notoriously difficult task which cannot be carried out without the help of annotation tools. In this paper we present a Perspicuous and Adjustable Links Annotator (PALinkA), a tool successfully used in several of our projects. We also briefly describe three types of discourse annotations applied using the tool.

1 Introduction

Annotation of discourse phenomena is a notoriously difficult task which cannot be carried out without the help of annotation tools. In this paper, we present an annotation tool successfully employed in three tasks which capture various discourse phenomena. In addition it proved useful in several other simpler tasks. Even though the annotation still needs to be done by humans, the features of the tool facilitate the process.

The structure of this paper is as follows: In Section 2 we discuss some of the requirements of annotation tools. Several such tools are discussed in Section 3 explaining why we decided to develop our own annotator. A brief description of it is presented in Section 4, followed by a three case studies briefly showing how the tool was used for marking different discourse phenomena. The article finishes with conclusions indicating ways to further develop the tool.

2 Requirements of annotation tools

In recent years the need to produce reusable corpora led to an increasing use of XML encoding in annotation. As a result, the annotation cannot be applied using simple text editors. In addition, the discourse annotation is usually complicated requiring specialised tools. In this section, we present the most important characteristics of a discourse annotation tool.

An annotation tool needs to be easy to use; with a minimum time required to learn how it works. It should also hide unnecessary details from the annotator (e.g. XML tags which are not directly linked to the task).

Usually the annotators are linguists with little or no experience of computers or annotation schemes. Because of this, an annotation tool has to be designed so that humans provide the information in a very simple and friendly way. In addition, the tool needs to ensure that no illegal information is introduced during the process (e.g. illegal XML constructions, wrong values for the attributes, etc.).

Last, but not least, it is desirable that a tool can be used for more than one task, so the annotators do not need to learn a new tool every time the task is changed. Moreover, in projects which build corpora in several languages, one way to ensure consistency between the annotations in the different languages is by using the same tool. Therefore, it is desired that a tool is language independent.

PALinkA, the tool presented in this paper meets all these requirements, being appropriate for discourse annotation.

3 Existing annotation tools

A large number of the existing annotation tools are for specific purposes only (e.g. for coreference (Garside and Rayson, 1997; Orăsan, 2000), for Rhetorical Structure Theory (Marcu, RSTTool)). Due to space limits we will not refer to them. In this section we briefly present few tools which can be used for a wide range of annotations tasks.

Day et. al. (1998) present Alembic Workbench, a tool developed by MITRE Corporation and used in the MUC conferences. The tool is highly customisable and features machine learning algorithms which facilitate the annotation process. Unfortunately the support seems to be discontinued and the documentation how to use the machine learning algorithms is sparse. When we tried to process texts with rich annotation it became slow.

Other tools which can be used to annotate a large range of discourse phenomena are MATE (McKelvie et al., 2001), ATLAS (Laprun et al., 2002) and MMAX (Müller and Strube, 2001). All these tools provide advanced frameworks for annotating text and speech, allowing customisation according to the task. They are very powerful, but they also require advanced computing knowledge in order to install and take full advantage of the facilities they provide. We consider that the installation and customisation process needs to be simple, so that people without much knowledge about computers can use them.

In the next section, we present PALinkA, a tool which requires little computing knowledge to install and customise, and can be employed in a large number of annotation tasks.

4 Perspicuous and Adjustable Links Annotator (PALinkA)

Our corpus annotated for coreference (Mitkov et al., 2000) was produced using Coreferential Links Annotator (CLinkA) (Orăsan, 2000). Even though the tool was useful for the annotation, we noticed that it has limitations. For example it does not allow to annotate texts which already contained other type of annotation and the annotation scheme it built in the tool which means that it cannot be changed.

We started to develop PALinkA as a replacement of CLinkA, trying to address its shortcomings. Soon we realised that it is easy to make a multipurpose annotation tool, which can be *adjusted* to the requirements of the task, without losing its ease of use, keeping it *perspicuous*.

The underlying idea of PALinkA is that it is possible to decompose most of the annotation tasks using three types of basic operations:

- Insertion of information not explicitly marked in the text (e.g. ellipsis, zero pronouns)
- Marking of the elements in a text (e.g. noun phrases, utterances, sentences)
- Marking the links between the elements (e.g. coreferential links)

We should emphasise that these three operation do not correspond to only three XML tags. The number of tags which can be inserted in a text is unlimited, for each one being possible to specify its name and attributes. However, for each tag it is necessary to define the type of operation attached to it, so that the tool will know how to handle it. For example, for missing information the tool will insert a marker in the text, whereas for a link it will ask the annotator to specify the referred element. The set of tags which can be used to annotate is loaded from a preferences file. Figure 1 shows a small part of a preferences file used to annotate coreference. It could look complicated for a non-expert in computers, but its syntax relies on a limited number of rules which are described in the documentation.

```
[MARKER]
;<EXP ID="#" COMMENT="">...</EXP>
NAME:EXP
BGCOLOR:23,255,254
FGCOLOR:123,111,10
ATTR:ID=# ;unique id
ATTR:COMMENT=!
INSERT_BEFORE:[
INSERT_AFTER:]
```

Figure 1: Part of the preferences file used to annotate the coreference

As can be seen in Figure 2 in the main screen of the tool does not display the XML tags, so the text can be easily read. In order to identify the tags present in the text, the user can specify colours to display the annotated text and can require to have the boundaries explicitly marked (in our example

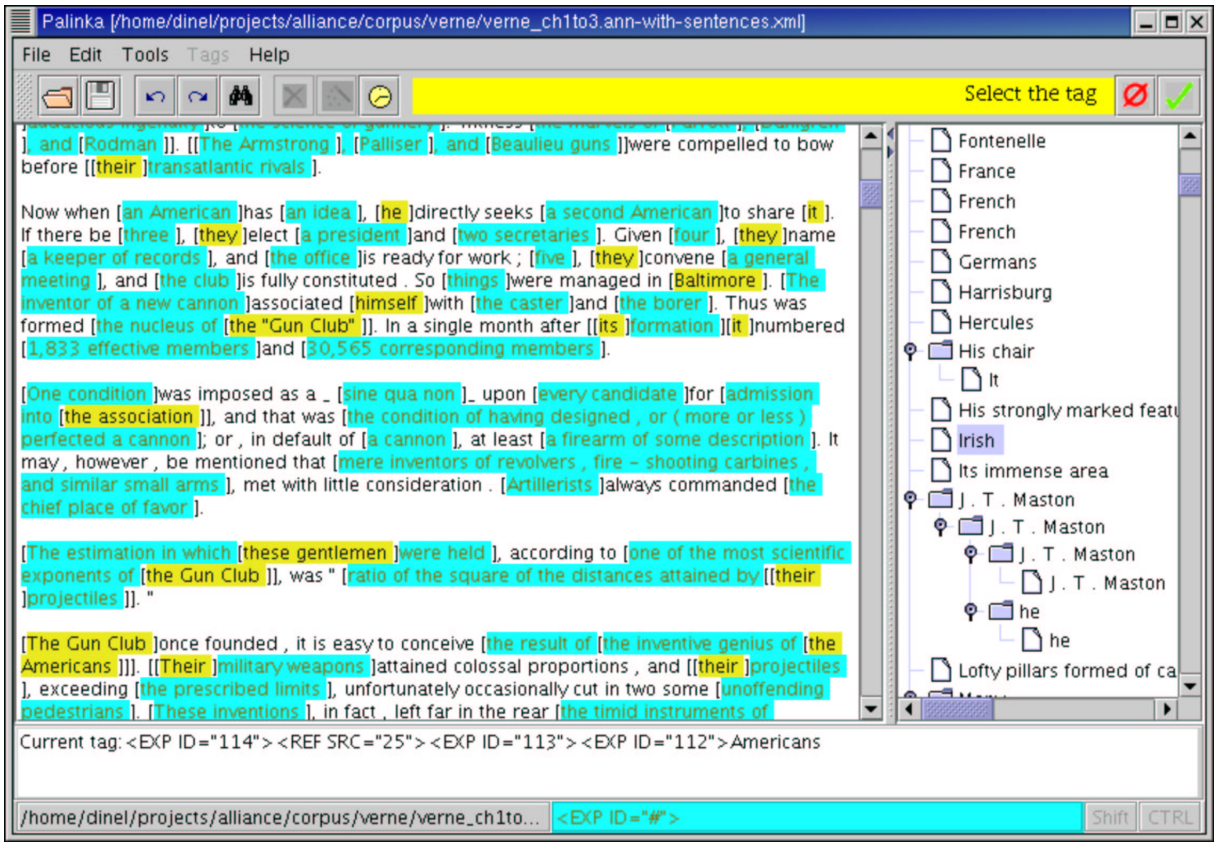

Figure 2: The main screen of the tool during annotation of coreference

with square brackets). PALinkA can be used to add annotation to files which already contain some sort of annotation. However, if the existing annotation is not relevant for the task, it does not appear on the screen at all.

The annotation process is kept as simple as possible; the boundaries of tags and the links between them being indicated with the help of the mouse. The tags which need to be linked require a unique ID. These IDs are generated and managed by the program allowing the annotator to concentrate on the annotation process. In addition to this, the tool has all the normal operations an annotation tool has: it can insert, delete or change tags.

The output of the program is a well-formed XML, the tool making sure that the human annotator does not produce invalid XML constructions. At present the tool supports only in-line annotation, but in the long term, we intent to offer the possibility of producing stand-off annotation.

Given that PALinkA is implemented in Java it is platform independent, running on any computer with a Virtual Java Machine installed. The tool is also language independent. In order to keep the tool as flexible as possible, it does not has a tokeniser. Instead, the tokens in the input text have to be explicitly marked using XML.

Due to space restrictions, we cannot present all the features of PALinkA and how it operates in more detail. More information can be found at the project's web page: http://clg.wlv.ac.uk/projects/PALinkA/. At the same address it is possible to download the tool for free.

5 Case studies

In this section, we show how PALinkA was used to create annotated corpora for coreference resolution, automatic summarisation and centering theory. We finish the section with few examples of simpler annotation tasks where PALinkA proved useful.

5.1 Coreference annotation

Annotating coreference is a notoriously time-consuming and labor-intensive task. In this task, the annotators have to mark the coreferential links between entities in a text. Usually, each entity receives a unique ID, and a link between two entities is marked using these IDs. These IDs are automatically managed by PALinkA. Some of the links refer to more than one entity. This fact can also be encoded using the tool.

For this annotation we extended the scheme presented in (Tutin et al., 2000). Even though this scheme is not similar with the one used in the MUC, it can be easily converted to the MUC scheme. PALinkA is currently used in the Alliance Project[1] to produce coreferentially annotated corpora for English and French.

The coreferential chains can be quickly identified by using the entities' tree in the right hand side of the screen (see Figure 2) or by highlighting them.

5.2 Annotation for automatic summarisation

Automatic summarisation is not part of the discourse analysis field, but it can use discourse information in order to produce high quality summaries. A corpus of news was annotated with information useful for automatic summarisation (Hasler et al., 2003). In addition to indicating the importance of each sentence, we enhanced the corpus with additional information which allows to measure the *conciseness* and the *coherence* of summaries. In order to be able to measure the conciseness of a summary, we indicated in the important sentences which parts can be removed without losing important information. For coherence, we used a simplified version of the coreference annotation task. For each important sentence containing a referential expression with the antecedent in another sentence, we indicated the link between sentences.

As for other tasks, the tool eased the annotation thanks to its friendly interface. In addition, PALinkA has two features which made the task much easier. One of these features indicates how much of the text is marked with a certain tag. We asked our annotators to mark 15% of the text as essential and another 15% as important. Using PALinkA it was possible to keep these length restrictions.

The time necessary to annotate a text was another parameter we wanted to record. With PALinkA it is possible to record this time. If the annotator needs to take a break during the process, this can be indicated by pressing the Pause button, in this way recording the actual time required by the annotation.

The corpus annotated for automatic summarisation is part of the Computer-Aided Summarisation Tool (CAST) project.[2]

5.3 Annotating centering

Centering Theory (CT) characterises the local coherence of a text on the basis of the discourse entities in a text and the way in which they are introduced (Grosz et al., 1995). CT was developed and demonstrated on simple texts. In order to test if the theory holds for real texts and gain insights into how the theory can be applied to them, 60 news reports and encyclopedic texts were annotated by several annotators. The number of annotated texts may seem small, but given the difficulty of the annotation and the fact that six versions of Centering Theory were marked for each text, it is impossible to produce large corpora.

In Centering Theory the discourse consists of a sequence of *utterances*. Each utterance has several *forward looking centers* and at most one *backward looking center*. One of the forward looking center is called *preferred center* and indicates the topic of the utterance. Due to space limits, Centering Theory cannot be discussed here, more details can be found in (Grosz et al., 1995; Walker et al., 1998).

The main difficulty when annotating centering comes from the number of embedded tags which have to be marked. Each utterance contains several centers, some of these also embedding other centers. Given this richness of tags the main advantage of using PALinkA is that it hides the XML tags, using colours for each tag. In addition to this, it is possible to configure the program to mark the beginning and end of each tag using a character chosen by the user. This feature proved also useful for coreference annotation. It is possible to notice it in Figure 2

[1] More details about the Alliance project are available at: http://clg.wlv.ac.uk/projects/Alliance/

[2] http://clg.wlv.ac.uk/projects/CAST/

where the boundaries of each NP are marked by square brackets. The user friendly interface facilitate the annotation process and does not distract the annotator with technical details.

5.4 Other tasks

In addition to annotating the aforementioned discourse phenomena, the tool was also employed in several other simpler tasks. It proved useful to annotate named entities in a corpus of Romanian news, mark noun phrases, prepositional phrases and their attachment in Romanian texts. The tool was also used to post-edit the output of automatic programs which identify the layout of scientific articles (e.g. headings, footnotes, citations).

6 Conclusions and future work

In this paper, we briefly presented a multipurpose annotation tool used in several of our projects which annotated the structure of the discourse. The tool is freely available for research purposes at http://clg.wlv.ac.uk/projects/PALinkA/.

In the future we intend add two new features to PALinkA. The first one will enable automating certain tasks with the possibility of post-editing the output of the automatic methods. We are currently working on an API which will allow the users to plug their modules into PALinkA. However, given that these modules will have to be written in Java, this function will be available only for programmers.

The second feature which we want to add to the system is to allow annotation of cross-document links. Such an option will prove very useful for cross-document coreference research.

7 Acknowledgements

The development of this tool was supported by the Arts and Humanities Research Board (AHRB) through the CAST project and by the British Council through the Alliance Project.

References

David Day, John Aberdeen, Sasha Caskey, Lynette Hirschman, Patricia Robinson, and Marc Vilain. 1998. Alembic workbench corpus development tool. In *Proceedings of the First International Conference on Language Resource&Evaluation*, pages 1021 – 1028.

Roger Garside and Paul Rayson. 1997. Higher-level annotation tools. In Roger Garside, Geoffrey Leech, and Anthony McEnery, editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pages 179 – 193. Addison Wesley Longman.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203 – 225.

Laura Hasler, Constantin Orăsan, and Ruslan Mitkov. 2003. Building better corpora for summarisation. In *Proceedings of Corpus Linguistics 2003*, pages 309 – 319, Lancaster, UK, March.

Christophe Laprun, Jonathan G. Fiscus, John Garofolo, and Sylvain Pajot. 2002. A practical introduction to ATLAS. In *Proceedings of LREC2003*, pages 1928 – 19932, Las Palmas de Gran Canaria, Spain.

Daniel Marcu. RSTTool. RST Annotation Tool. Availabe at: http://www.isi.edu/licensed-sw/RSTTool/index.html.

David McKelvie, Amy Isard, Andreas Mengel, Morten B. Moeller, Michael Grosse, and Marion Klein. 2001. The MATE Workbench - an annotation tool for XML coded speech corpora. *Speech Communication*, 33(1–2):97 – 112.

Ruslan Mitkov, Richard Evans, Constantin Orăsan, Cătălina Barbu, Lisa Jones, and Violeta Sotirova. 2000. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*, pages 49–58, Lancaster, UK.

Christoph Müller and Michael 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*, pages 45 – 50, Seattle, Washington, 5th August.

Constantin Orăsan. 2000. CLinkA a coreferential links annotator. In *Proceedings of LREC'2000*, pages 491 – 496, Athens, Greece.

Agnes Tutin, Francois Trouilleux, Catherine Clouzot, Eric Gaussier, Annie Zaenen, Stephanie Rayot, and Georges Antoniadis. 2000. Annotating a large corpus with anaphoric links. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2000)*, pages 28 – 38, Lancaster, UK, 16th – 18th November.

Marilyn A. Walker, Aravind K. Joshi, and Ellen Prince, editors. 1998. *Centering Theory in Discourse*. Oxford University Press.