Trameur: A Framework for Annotated Text Corpora Exploration

Serge Fleury, Maria Zimina

Trameur: Framework for Annotated Text Corpora Exploration

Serge Fleury
Sorbonne Nouvelle – Paris 3
SYLED-CLA2T, EA2290
75005 Paris, France
serge.fleury@univ-paris3.fr

Maria Zimina
Paris Diderot – Sorbonne Paris Cité
CLILLAC-ARP, EA 3967
75205 Paris, cedex 13, France
maria.zimina@eila.univ-paris-diderot.fr

Abstract

Corpus resources with complex linguistic annotations are becoming increasingly important in the work of language specialists with different professional backgrounds. Regardless of their computer skills, they often need to perform extensive corpus research, including Natural Language Processing (NLP), statistical modelling and data visualisation. Our software system, called Trameur, aims at making these analyses possible within a single graphical user interface. It relies upon a specific data modelling framework presented in this paper.

1 Introduction

Annotated text corpora are currently used by language specialists with different professional backgrounds. As a rule, linguists are familiar with standard text processing tools, software engineers use Natural Language Processing (NLP) tools and statisticians work with statistics software. Corpus resources are getting more and more complex and consist of several layers of annotations for multifaceted data collections. To face these challenges, we develop an integrated system for complex annotated text data analyses called Trameur (Fleury, 2013a).

Trameur manages multiple layers of linguistic annotations and allows exploring complex linguistic features and embedded dependence relations. Additionally, Trameur incorporates advanced NLP processing, statistical analysis and text mapping features. Thus, Trameur has been developed to allow multiple corpus analyses within a single graphical user interface.

This framework system was developed to make complex analyses of annotated corpora accessible to any corpus linguist, without extensive programing skills and knowledge of statistical modelling tools.

2 Corpus research with Trameur

Trameur was successfully tested for monolingual text processing within several research projects in corpus linguistics and discourse analysis (Branca-Rosoff et al., 2012; Née et al., 2012).

On-going research demonstrates its potential for processing parallel and comparable text data in distant languages (Zimina and Fleury, 2014).

2.1 Data structure

Figure 1 presents a data sample from the Rhapsodie corpus of spoken French annotated for prosody and syntax (Gerdes et al., 2012).

The data file is available online in tabular form and can be displayed in a spreadsheet.[1] It is a list of linguistic features dealing with syntax and prosody, and a list of annotated tokens from Rhapsodie texts. All texts are identified by codes (column 1). Each text is further divided into separate interlocutory units (column 2) and segmented into tokens (column 3). The remaining columns display linguistic annotations of the tokens (columns 4-21):

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	TextID	TreeID	TokenID	Token	Lemma	POS	Mode	Tense	Person	Number	Gender	Gov_rection	Type_rection	Gov_para	Type_para	Gov_inher	Type_inher	Gov_junc	Type_jun	Gov_junc-inhe	Type_junc-inher
2	M2006	1	1	bonjour	bonjour	B_I						0	root								
3	M2006	1	2																		
4	M2006	1	3	Eric	Eric	B_N				sg	masc	0	root								
5	M2006																				
6	M2006	2	1	bonjour	bonjour	B_N				sg	masc	0	root								
7	M2006	2	2																		
8	M2006	2	3	à	à	B_Pre						1	dep								
9	M2006	2	4																		
10	M2006	2	5	tous	tous	B_Pro			3	pl	masc	3	dep								
11	M2006																				
12	M2006	3	1	nouvelle	nouveau	B_Adj				sg	fem	3	dep								
13	M2006	3	2																		
14	M2006	3	3	nuit	nuit	B_N				sg	fem	0	root								
15	M2006	3	4																		
16	M2006	3	5	de	de	B_Pre						3	dep								
17	M2006	3	6																		
18	M2006	3	7	pillage	pillage	B_N				sg	masc	5	dep								
19	M2006	3	8																		
20	M2006	3	9	et	et	B_I													5	junc	
21	M2006	3	10																		
22	M2006	3	11	d	de	B_Pre								5	para_coord	3	dep_inherited		9	junc	

Figure 1. Data sample from Rhapsodie

Figure 2 shows the same data set transformed from the spreadsheet format into a Trameur base file using regular expressions (Fleury, 2013b). The data structure is composed of two parts:

- A Thread (in French: *trame*), which is a list of items with position identifiers.
- A Frame, which is a list of corpus partitions defined on the Thread. Each partition has a name and a list of named constituents identified through their first and last token positions on the Thread.

Thus, each annotated token from the Rhapsodie corpus becomes an item identified by its position on the Thread.

Dependence relations among items are annotated as follows: RELATION(TARGET), where:

- RELATION is a character string corresponding to the name of the relation
- TARGET is a numerical value of the position identifier on the Thread (see Figure 2).

All linguistic annotations from the Rhapsodie sample file are preserved in the Trameur base file. These annotations can be further edited in the graphical user interface.

Figure 2. Rhapsodie data in Trameur base file

[1] The Rhapsodie corpus texts are available for download from: `http://www.projet-rhapsodie.fr`

2.2 Exploring dependence relations

Dependence relations emerging from linguistic annotations can be explored, filtered and displayed in several ways:

- Using graphs of dependence relations
- Using concordances and context return
- Using co-occurrence statistics
- Using visual tools (graphs) and text mapping

A detailed description of these features with related screenshots is available on-line (Fleury, 2013).

Figure 3 displays a sample graph of a double dependence relation from Rhapsodie. It is set by the regular expression OBJ|SUB (OBJ or SUB), where the relation target for the lemma is “aimer” and the annotation n°9 is ROOT:

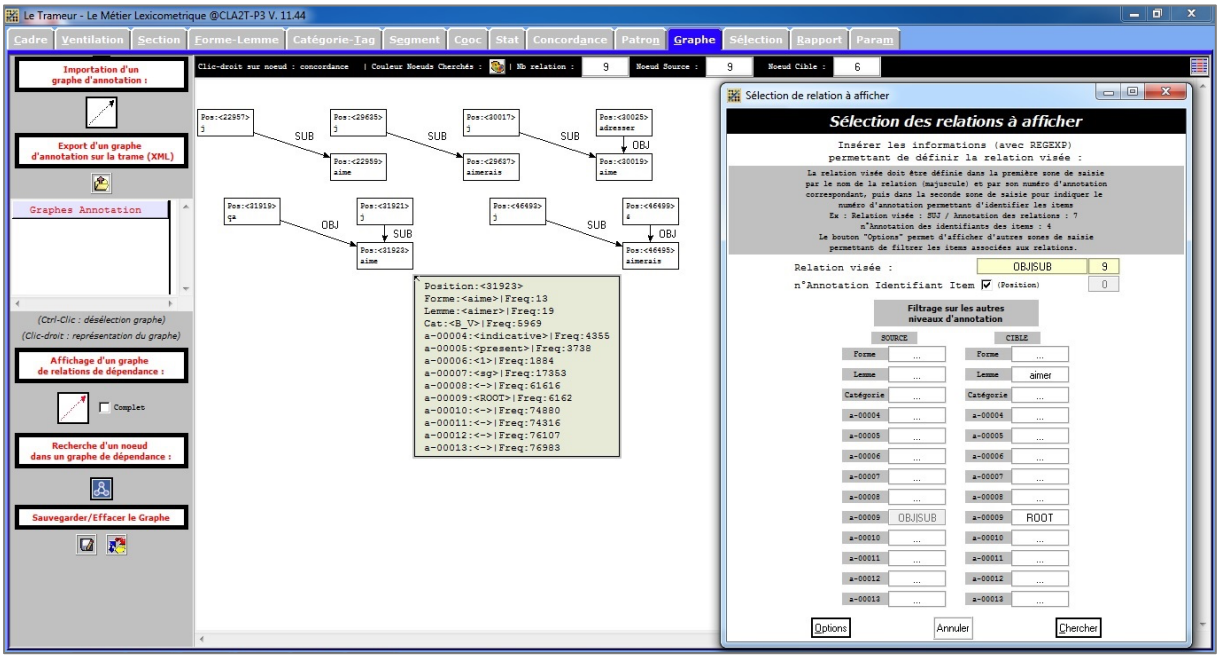

Figure 3. Graph of a double dependence relation

Figure 4 displays the nodes of the same graph in a concordance window. Selected annotations are automatically highlighted.

3 Conclusion

Trameur is a tool for exploration of complex multi-layer linguistic corpora with different annotations. It is built upon a Thread/Frame data model using XML. The software is distributed with a graphical user interface. A reference package for Windows is available for free download from the official website: `http://www.tal.univ-paris3.fr/trameur`.

Several other systems have been already developed for processing annotated corpora, for example: ANNIS[2] and Macaon[3]. However, the novelty of Trameur consists in expanding a multi-layered data model to all stages of corpus exploration, including text mapping features and statistical analysis within a single graphical user interface.

Following this integrated approach, Trameur can be used, for example, to build complex linguistic units, analyse their dependence relations and reveal their characteristic attractions using text statistics.

---

[2] ANNIS (ANNotation of Information Structure): `http://www.sfb632.uni-potsdam.de/annis/`

[3] MACAON Project: `http://macaon.lif.univ-mrs.fr`

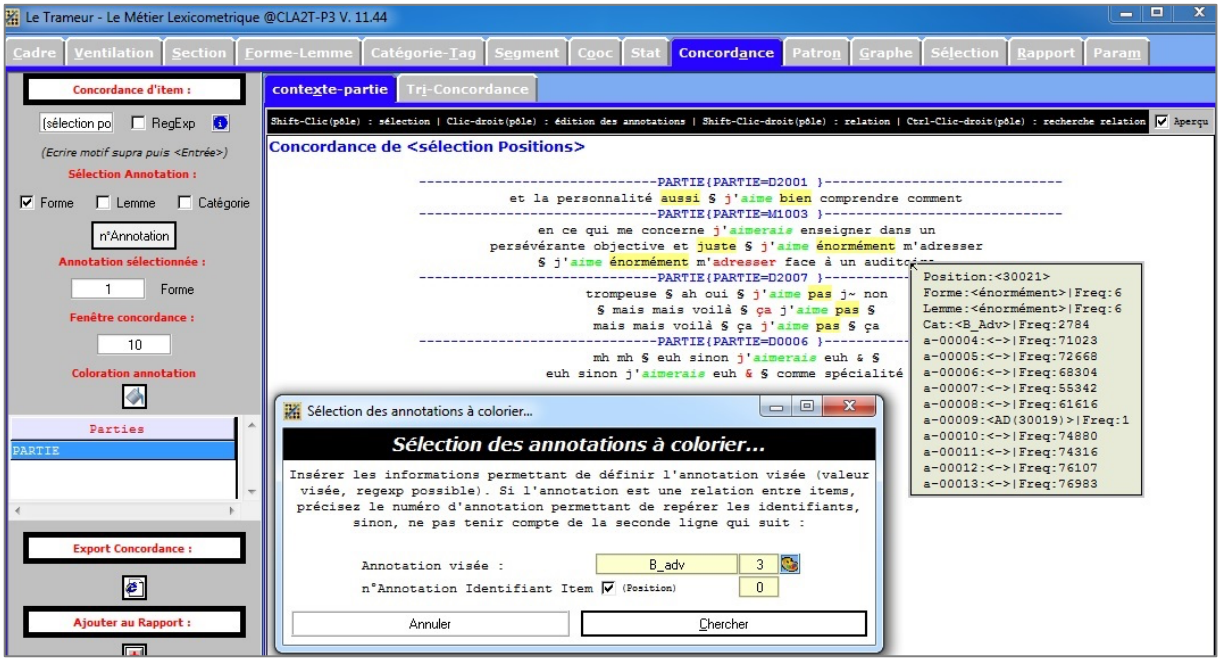

Figure 4. Access to dependence relations in a concordance window

Thus, the use of the Thread/Frame data model implemented in Trameur allows different combinations of analyses when processing multifaceted linguistic annotations. We believe that this framework offers new perspectives for corpus research.

References

Sonia Branca-Rosoff, Serge Fleury, Florence Lefeuvre and Mat Pires. 2012. *Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP 2000). Sorbonne nouvelle – Paris 3*. Online publication: `http://cfpp2000.univ-paris3.fr/CFPP2000.pdf`

Serge Fleury. 2013a. *Le Trameur. Propositions de description et d'implémentation des objets textométriques. Sorbonne nouvelle – Paris 3.* Online publication: `http://www.tal.univ-paris3.fr/trameur/trameur-propositions-definitions-objets-textometriques.pdf`

Serge Fleury. 2013b. *Annotations Rhapsodie pour le Trameur*. *Sorbonne nouvelle – Paris 3*. Online publication: `http://www.tal.univ-paris3.fr/trameur/bases/rhapsodie2trameur.pdf`

Emilie Née, Erin MacMurray and Serge Fleury. 2012. Textometric Explorations of Writing Processes: A Discursive and Genetic Approach to the Study of Drafts. Journées internationales d'Analyse statistique des Données Textuelles (JADT 2012). Liège (Belgium), 13-15 June 2012.

Kim Gerdes, Sylvain Kahane, Anne Lacheret, Arthur Truong, and Paola Pietrandrea. 2012. *Intonosyntactic data structures: The Rhapsodie treebank of spoken French*. *Proceedings of the Linguistic Annotation Workshop, COLING 2012.* Jeju, Republic of Korea, July 2012.

Emilie Née, Erin MacMurray and Serge Fleury. 2012. *Textometric Explorations of Writing Processes*: *A Discursive and Genetic Approach to the Study of Drafts*. *Journées internationales d'Analyse statistique des Données Textuelles (JADT 2012)*. Liège (Belgium), 13-15 June 2012.

Maria Zimina and Serge Fleury. 2014. *Approche systémique de la résonance textuelle multilingue. Journées internationales d'Analyse statistique des Données Textuelles (JADT 2014).* Paris, 3-6 June 2014.