Trameur: A 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. 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

Treebanks, or parsed text corpora that annotate syntactic or semantic sentence structure, are widely appreciated in language studies. These corpora are getting more and more complex and consist of several layers of annotations. Such multifaceted data collections present many challenges for the development of specific NLP (Natural Language Processing) tools, statistical modelling and data visualisation techniques. In terms of statistical text analysis, most research has been carried out using either unannotated texts or tagged texts, annotated with parts of speech only (Lebart et al., 1998). As for corpora with richer annotations such as clauses, grammatical functions and dependency links, recent research projects have been focused on the development of treebank querying methods with fast search algorithms (Cunningham et al., 2002; Mírovský and Ondruška, 2002; Götze and Dipper, 2006). However, for annotated corpus analysis, it is often necessary to precisely locate different types of linguistic schemes under study, with simultaneous access to multiple corpus layers and their interactions statistics.

To face these challenges, we develop an integrated system for statistical analysis of annotated text data called Trameur (Fleury, 2013a). Trameur manages multiple layers of linguistic annotations and allows statistical exploration of complex linguistic features and embedded dependency relations. Additionally, Trameur incorporates advanced NLP processing and text mapping features. This framework has been developed to allow multiple corpus analyses within a single graphical user interface, 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 structures for annotated text processing

Several software packages dedicated to multi-level text processing apply pipeline architecture to support reproducible analyses of text annotations stored in specific formats (Eberle et al., 2012; Eckart et al., 2012). We provide here a description of the specific data structures used by Trameur (text segmentation, partition, statistical tables, etc.) to implement incremental textual resources for treebanks.

In Trameur, a formal representation of text segmentation relies upon two systems of units: text containers and contents (or items) (Fleury, 2013a). Text containers reflect the structure of different corpus parts, while contents represent systems of textual units (lexemes, graphemes, etc.) and their frequency variations which can be examined on a statistical basis. Based on computerized procedures, the systems of containers and contents allow automatic counts in the form of large statistical tables. These tables can be further processed by quantitative methods to detect salient points of statistical distribution in the corpus and build the premises of variation analysis.

Containers and contents are essential elements of double-tracking text segmentation process: (1) Cutting text into containers (parts, textual zones, sections, chapters, paragraphs, sentences, etc.) allows creating a system of text spans that can be further compared on a quantitative basis. (2) Text segmentation into items (or contents) isolates distinct textual units within a text corpus. This text segmentation is closely interconnected with the other two operations: annotation of items (attributing a distinct tag, such as lemma or grammatical category, to each isolated item), and typing of isolated items (judged similar because of their intrinsic nature or on the basis of corpus annotations). Thus, each segmented text is a succession of isolated items that can always be reunited to restore the initial text sequence. Each item is further attached to a given type.

After text segmentation in Trameur, the items are numbered to create a system of coordinates, where each item is spotted by its corresponding sequence number. In Trameur, this system is called Thread (in French: *trame*). The same system of coordinates is used to define and locate text spans (or textual zones) formed by a series of consecutive items between positions x_1 and x_2, or by a certain number of textual zones of this type. The definition of a Thread allows describing various systems of textual zones corresponding to the text containers (corpus parts, sections, chapters, paragraphs, sentences, etc.). The descriptions of the containers are united within a specific data structure called Frame.

The Thread/Frame structure allows defining Selection as any object corresponding to the selected subset of Thread items used in the corpus exploration. There are several types of Selection: (1) Selection corresponding to the contents (certain Thread items, such as occurrences of the same token, lemma, or items corresponding to the same regular expression operated on one or several annotation levels); (2) Selection corresponding to certain containers (textual zones, paragraphs, sections, etc.) composed of contiguous item sequences, selected according to their positions in the text; (3) Selection resulting from highlighting operations performed by a human expert on the basis of some complex criteria, possibly difficult to describe on a formal basis; (4) A textual zone discovered following the results of a statistical calculation; (5) A sub-set of textual units selected following a statistical calculation that highlights a specific distribution of items (contents) within a corpus.

The following demonstration of these principles is based on 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] The data set used in this demonstration comes from Rhapsodie v1 micro-syntax annotations. This first release is limited to 13 annotations.[2] The corpus data is structured as follows: all Rhapsodie texts are first identified by codes (column 1). Each text is further divided into separate units (column 2) and segmented into tokens (column 3). The remaining columns display linguistic annotations of the tokens, including microsyntax (rection, dependency and constituancy) (Fleury, 2013b).

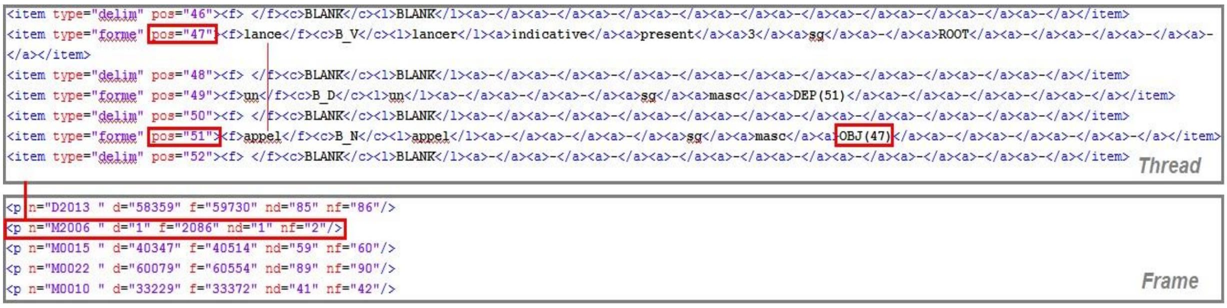
Figure 1 shows the Rhapsodie data set transformed from the spreadsheet format into a Trameur base file using regular expressions. The data structure is composed of two parts: (1) a Thread, which is a list of items with position identifiers; (2) a Frame, which is a list of corpus partitions defined on the Thread. Each partition has a name and a list of named constituents identified through their first and

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

[2] Rhapsodie v3 is the latest online release with micro-syntax, macro-syntax and prosody annotations. It is currently under study in a number of on-going research projects that will be revealed in forthcoming publications.

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

Dependency 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 1). All linguistic annotations from the Rhapsodie sample file are preserved in the Trameur base file. These annotations can be displayed and edited in the graphical user interface (see Figure 2): a-00001:Token, a-00002 Lemma, a-00003 P.O.S, a-00004 Mode, a-00005 Tense, a-00006 Person, a-00007 Number, a-00008 Gender, a-00009 Type_rection(Gov_rection), a-00010 Type_para(Gov_para), a-00011 Type_inher(Gov_inher), a-00012 Type_junc(Gov_junc), a-00013 Type_junc-inher(Gov_junc-inher).

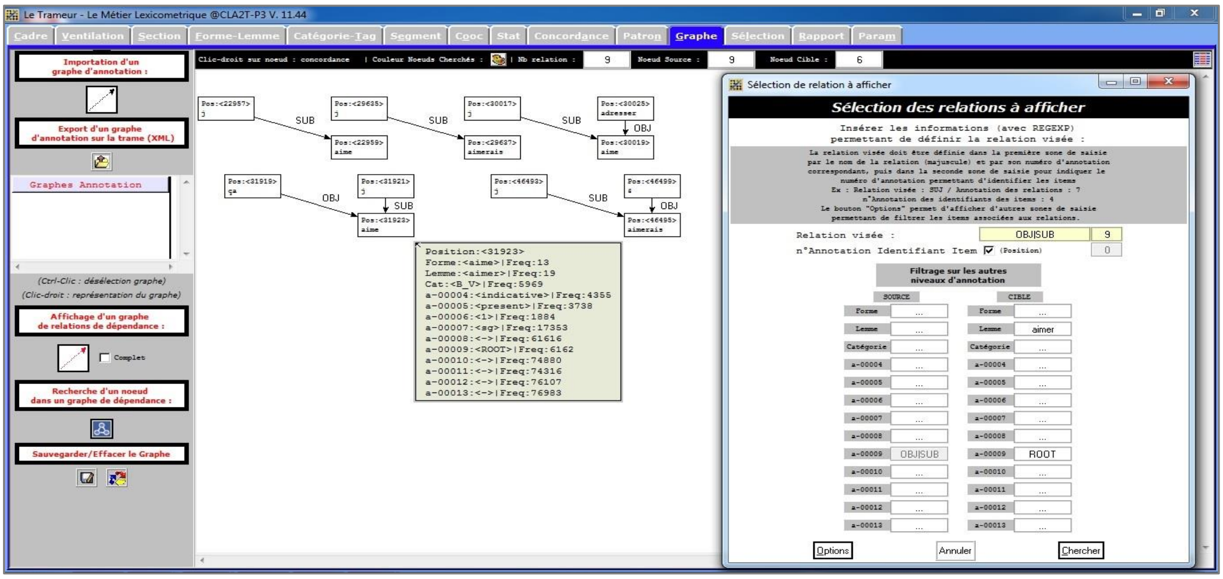


```
<item type="delim" pos="46"><f> </f><c>BLANK</c><l>BLANK</l><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
<item type="forme" pos="47"><f>lance</f><c>B_V</c><l>lancer</l><a>indicative</a><a>present</a><a>3</a><a>sg</a><a>-</a><a>ROOT</a><a>-</a><a>-</a><a>-</a><a>-
</a></item>
<item type="delim" pos="48"><f> </f><c>BLANK</c><l>BLANK</l><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
<item type="forme" pos="49"><f>un</f><c>B_D</c><l>un</l><a>-</a><a>-</a><a>-</a><a>sg</a><a>masc</a><a>DEP(51)</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
<item type="delim" pos="50"><f> </f><c>BLANK</c><l>BLANK</l><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
<item type="forme" pos="51"><f>appel</f><c>B_N</c><l>appel</l><a>-</a><a>-</a><a>-</a><a>sg</a><a>masc</a><a>OBJ(47)</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
<item type="delim" pos="52"><f> </f><c>BLANK</c><l>BLANK</l><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a><a>-</a></item>
```

```
<p n="D2013 " d="58359" f="59730" nd="85" nf="86"/>
<p n="M2006 " d="1" f="2086" nd="1" nf="2"/>
<p n="M0015 " d="40347" f="40514" nd="59" nf="60"/>
<p n="M0022 " d="60079" f="60554" nd="89" nf="90"/>
<p n="M0010 " d="33229" f="33372" nd="41" nf="42"/>
```

Figure 1: Rhapsodie data in a Trameur base file.

2.2 Exploration of dependency relations



Le Trameur - Le Métier Lexicometrique @CLA2T-P3 V. 11.44
Cadre
Ventilation
Section
Forme-Lemme
Catégorie-Tag
Segment
Cooc
Stat
Concordance
Patron
Graphe
Sélection
Rapport
Param
Importation d'un graphe d'annotation :
Export d'un graphe d'annotation sur la trame (XML)
Graphes Annotation
(Ctrl-Clic : désélection graphe)
(Clic-droit : représentation du graphe)
Affichage d'un graphe de relations de dépendance :
Complet
Recherche d'un noeud dans un graphe de dépendance :
Sauvegarder/Effacer le Graphe
Pos:<22957> j
SUB
Pos:<22959> aime
Pos:<29635> j
SUB
Pos:<29637> aimerais
Pos:<30017> j
SUB
Pos:<30025> adresser
OBJ
Pos:<30019> aime
Pos:<31919> ça
OBJ
Pos:<31921> j
SUB
Pos:<31923> aime
Pos:<46493> j
SUB
Pos:<46499> s
OBJ
Pos:<46495> aimerais
Sélection de relation à afficher
Sélection des relations à afficher
Options
Annuler
Chercher

Clic-droit sur noeud : concordance | Couleur Noeuds Cherchés : | Nb relation : 9 Noeud Source : 9 Noeud Cible : 6

```
Position:<31923>
Forme:<aime>|Freq:13
Lemme:<aimer>|Freq:19
Cat:<B_V>|Freq:5969
a-00004:<indicative>|Freq:4355
a-00005:<present>|Freq:3738
a-00006:<1>|Freq:1884
a-00007:<sg>|Freq:17353
a-00008:<->|Freq:61616
a-00009:<ROOT>|Freq:6162
a-00010:<->|Freq:74880
a-00011:<->|Freq:74316
a-00012:<->|Freq:76107
a-00013:<->|Freq:76983
```

Insérer les informations (avec REGEXP)
permettant de définir la relation visée :

La relation visée doit être définie dans la première zone de saisie par le nom de la relation (majuscule) et par son numéro d'annotation correspondant, puis dans la seconde zone de saisie pour indiquer le numéro d'annotation permettant d'identifier les items
Ex : Relation visée : SUJ / Annotation des relations : 7
n°Annotation des identifiants des items : 4
Le bouton "Options" permet d'afficher d'autres zones de saisie permettant de filtrer les items associées aux relations.

Relation visée : OBJ|SUB 9
n°Annotation Identifiant Item ☑ (Position) 0

Filtrage sur les autres niveaux d'annotation

SOURCE		CIBLE	
Forme	...	Forme	...
Lemme	...	Lemme	aimer
Catégorie	...	Catégorie	...
a-00004	...	a-00004	...
a-00005	...	a-00005	...
a-00006	...	a-00006	...
a-00007	...	a-00007	...
a-00008	...	a-00008	...
a-00009	OBJ|SUB	a-00009	ROOT
a-00010	...	a-00010	...
a-00011	...	a-00011	...
a-00012	...	a-00012	...
a-00013	...	a-00013	...

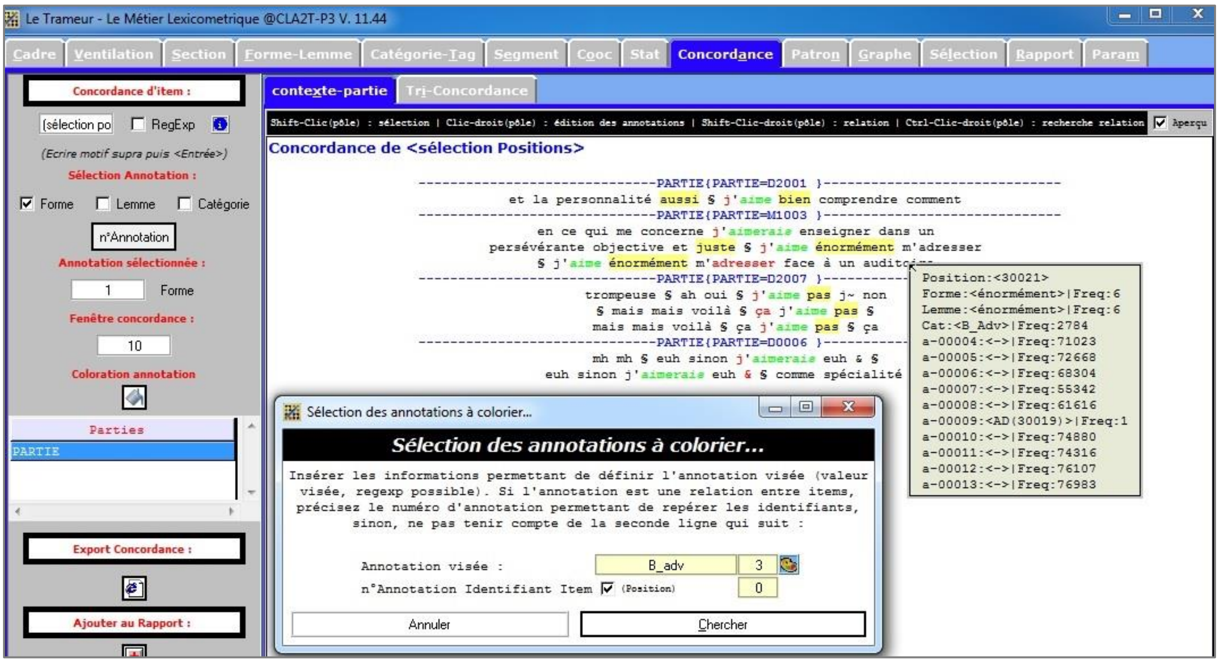
Figure 2: Trameur. Graph of a double dependency relation.

Dependency relations emerging from linguistic annotations can be explored, filtered and displayed using dependency graphs, text maps with context return, concordances, co-occurrence statistics, etc. All these features with related screenshots are presented in the on-line user guide.[3]

Figure 2 displays a sample graph of a double dependency 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 Type_rection(Gov_rection) is ROOT. For each item in context, a pop-up window displays all annotation levels with corresponding tags (or “no tag” signs <—>) and their frequencies (Freq) on a given annotation level.

Figure 3 displays the nodes of the same graph in a concordance window. A selection tool is used to highlight annotation n°3 in concordance lines (value = `B_adv`).

[3] Trameur (user guide): `http://www.tal.univ-paris3.fr/trameur/leMetierLexicometrique.htm`



[sélection po
☑ Forme
1 Forme
10
PARTIE

Shift-Clic(pôle) : sélection | Clic-droit(pôle) : édition des annotations | Shift-Clic-droit(pôle) : relation | Ctrl-Clic-droit(pôle) : recherche relation ☑ Aperçu

Concordance de <sélection Positions>

```
------------------------------PARTIE{PARTIE=D2001 }------------------------------
           et la personnalité aussi § j'aime bien comprendre comment
------------------------------PARTIE{PARTIE=M1003 }------------------------------
              en ce qui me concerne j'aimerais enseigner dans un
         persévérante objective et juste § j'aime énormément m'adresser
              § j'aime énormément m'adresser face à un auditoire
------------------------------PARTIE{PARTIE=D2007 }------------------------------
                    trompeuse § ah oui § j'aime pas j~ non
                     § mais mais voilà § ça j'aime pas §
                     mais mais voilà § ça j'aime pas § ça
------------------------------PARTIE{PARTIE=D0006 }------------------------------
                     mh mh § euh sinon j'aimerais euh & §
                euh sinon j'aimerais euh & § comme spécialité
```

```
Position:<30021>
Forme:<énormément>|Freq:6
Lemme:<énormément>|Freq:6
Cat:<B_Adv>|Freq:2784
a-00004:<->|Freq:71023
a-00005:<->|Freq:72668
a-00006:<->|Freq:68304
a-00007:<->|Freq:55342
a-00008:<->|Freq:61616
a-00009:<AD(30019)>|Freq:1
a-00010:<->|Freq:74880
a-00011:<->|Freq:74316
a-00012:<->|Freq:76107
a-00013:<->|Freq:76983
```

Sélection des annotations à colorier...

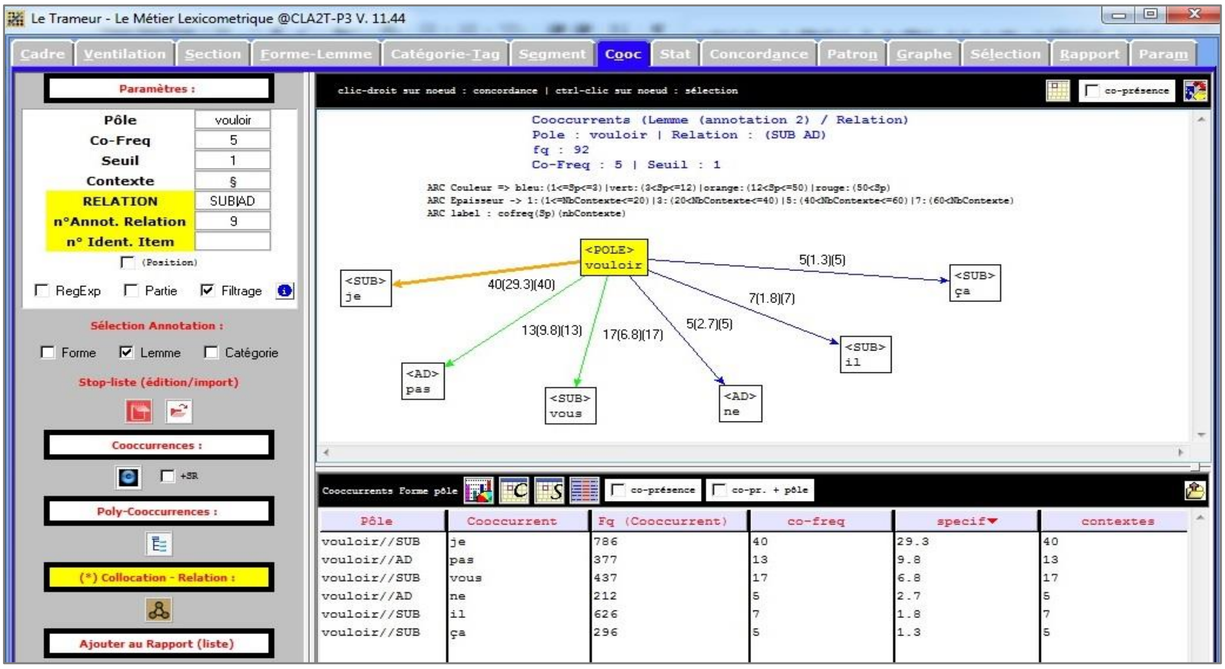
Sélection des annotations à colorier...

Insérer les informations permettant de définir l'annotation visée (valeur visée, regexp possible). Si l'annotation est une relation entre items, précisez le numéro d'annotation permettant de repérer les identifiants, sinon, ne pas tenir compte de la seconde ligne qui suit :

Annotation visée : B_adv 3
n°Annotation Identifiant Item ☑ (Position) 0

Figure 3: Trameur. Concordance window with selected dependency relations.

Figure 4 displays a co-occurrence graph for the lemma “vouloir” constrained by the dependency RELATION set by the regular expression SUB|AD (SUB or AD).



Pôle	vouloir
Co-Freq	5
Seuil	1
Contexte	§
RELATION	SUB|AD
n°Annot. Relation	9
n° Ident. Item	

□ (Position)
□ RegExp □ Partie ☑ Filtrage
□ Forme ☑ Lemme □ Catégorie

clic-droit sur noeud : concordance | ctrl-clic sur noeud : sélection □ co-présence

```
Cooccurrents (Lemme (annotation 2) / Relation)
Pole : vouloir | Relation : (SUB AD)
fq : 92
Co-Freq : 5 | Seuil : 1
```

```
ARC Couleur => bleu:(1<=Sp<=3)|vert:(3<Sp<=12)|orange:(12<Sp<=50)|rouge:(50<Sp)
ARC Epaisseur -> 1:(1<=NbContexte<=20)|3:(20<NbContexte<=40)|5:(40<NbContexte<=60)|7:(60<NbContexte)
ARC label : cofreq(Sp)(nbContexte)
```

□ co-présence □ co-pr. + pôle

Pôle	Cooccurrent	Fq (Cooccurrent)	co-freq	specif▼	contextes
vouloir//SUB	je	786	40	29.3	40
vouloir//AD	pas	377	13	9.8	13
vouloir//SUB	vous	437	17	6.8	17
vouloir//AD	ne	212	5	2.7	5
vouloir//SUB	il	626	7	1.8	7
vouloir//SUB	ça	296	5	1.3	5

Figure 4: Trameur. Co-occurrence graph constrained by selected dependency relations.

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: PDT2.0[4], GATE[5], ANNIS[6], Macaon[7]. 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 of dependency relations within a single graphical user interface.

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

In Trameur, quantitative analyses are applied to textual resources, built upon a Thread and an evolutionary Frame. These analyses can progressively enrich these resources with new text divisions or annotations, without deleting previous information. Following this approach, it becomes possible to implement incremental textual resources for treebanks.

It is the use of the Thread/Frame data model implemented in Trameur that allows different combinations of analyses when processing multifaceted linguistic annotations. We believe that this evolutionary 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`

Hamish Cunningham, Diana Maynard, Kalina Bontcheva and Valentin Tablan. 2002. *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, July 2002.

Kurt Eberle, Kerstin Eckart, Ulrich Heid and Boris Haselbach, 2012. *A Tool/Database Interface for Multi-Level Analyses. Proceedings of the eighth conference on International Language Resources and Evaluation (LREC 2012)*. Istanbul (Turkey), May 2012.

Kerstin Eckart, Arndt Riester and Katrin Schweitzer, 2012. *A Discourse Information Radio News Database for Linguistic Analysis. Linked Data in Linguistics*. Springer.

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` (v1 base)
`http://www.tal.univ-paris3.fr/trameur/bases/rhapsodie2trameur-v3.pdf` (v3 base)

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.

Michael Götze and Stefanie Dipper. 2006. *ANNIS: Complex Multilevel Annotations in a Linguistic Database. Proceedings of the 5th Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing (NLPXML-2006)*. Trento (Italy), April 2006.

Ludvic Lebart, André Salem and Lisette Barry. 1998. *Exploring Textual Data*. Kluwer Academic Publishers.

Jiří Mírovský and Roman Ondruška. 2002. *NetGraph System: Searching through the Prague Dependency Treebank. Prague Bulletin of Mathematical Linguistics*, 77, MFF UK, Prague, Czech Republic, Prague, 2002.

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), 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, June 2014.

[4] PDT (The Prague Dependency Treebank 2.0): `ufal.mff.cuni.cz/pdt2.0/`

[5] GATE (General Architecture for Text Engineering): `http://gate.ac.uk/gate/`

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

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