

Alignement de Treebank pour le *Trameur* (v1)

18/02/2015 21:40:22

Serge Fleury

(document de travail)

Références

Le Trameur, manuel d'utilisation

<http://www.tal.univ-paris3.fr/trameur/leMetierLexicometrique.pdf>

Dans cette documentation, la partie « *Relations de dépendance entre les items de Trame (via leurs annotations)* » présente l'état des développements actuellement disponibles dans le *Trameur* pour l'exploitation des relations de dépendance.

Le Trameur. Propositions de description et d'implémentation des objets textométriques

<http://www.tal.univ-paris3.fr/trameur/trameur-propositions-definitions-objets-textometriques.pdf>

Ce document met au jour une description des objets textométriques et les méthodes mises en œuvre dans le *Trameur* pour travailler sur et avec ces objets dans une perspective textométrique. On y détaille aussi les opérations permises sur une *base textométrique* : format des données textuelles, modification dynamique de la *Trame*, correction ou ajout d'annotation etc.

Ressources PARTUT

ParTUT is a project for the development of a multilingual parallel treebank for Italian, English and French. The aim of this work is twofold: building an aligned parallel treebank for Italian, English and French, by extending and applying a single treebank schema to other languages, and studying how the schema can be used to address issues typically related to parallel corpora. The annotation and tools used for the development of this resource are those of the [Turin University Treebank \(TUT\)](#), a collection of Italian sentences annotated at a morpho-syntactic, syntactic and (to a lesser extent) semantic level, with dependency-oriented representation format.

<http://www.di.unito.it/~tutreeb/partut.html>

1. Préambule

Ce document commence par décrire le processus de transcodage des données issues du projet *Partut* (<http://www.di.unito.it/~tutreeb/partut.html>) sous la forme d'une base textométrique importable dans le *Trameur*.

La base résultante contient un alignement de treebank (1 volet EN et un volet FR, les 2 étant multiannotés).

Ce document présente ensuite les différentes fonctionnalités mises en œuvre pour traiter ce type de données via le *Trameur*.

La base ParTUT2Trameur est accessible à cette adresse :

<http://www.tal.univ-paris3.fr/trameur/bases/baseTrameurFromPartut.zip>

2. La Base ParTUT2Trameur (v1)

2.1 Données ParTUT

Les données traitées pour constituer la base sont disponibles à cette adresse :

<http://www.di.unito.it/~tutreeb/corpora/Par-TUT/FrEn.zip>

Le projet ParTUT offre des ressources sur 3 langues (EN, FR et ITA).

Seules les ressources EN et FR ont été utilisées ici.

Les données ParTUT utilisées ici sont au format CONL.

La figure suivante donne à voir un extrait du fichier FR : un empilement de phrases (2 sont visibles ci-dessous) ; les mots de chaque phrase sont décrits par une ligne (10 colonnes) , ces informations seront utilisées pour construire les annotations de chaque forme graphique.

| | | | | | | | | | |
|----|------------------------|-----------|-------|------------------------------|-------------------------------|----------------|---------------|---|---|
| 1 | <u>Paternité</u> | PATERNITÉ | NOUN | NOUN | COMMON F SING | 0 | TOP | _ | _ |
| 2 | - #\- | PUNCT | PUNCT | _ | 1 | SEPARATOR | _ | _ | |
| 3 | <u>Partage</u> | PARTAGE | NOUN | NOUN | COMMON M SING PARTAGER TRANS | 1 | APPOSITION | _ | _ |
| 4 | <u>Des</u> DE | PREP | PREP | MONO | 1 | RMOD | _ | _ | |
| 5 | <u>Des</u> LE | ART | ART | DEF ALLVAL PL | 4 | ARG | _ | _ | |
| 6 | <u>Conditions</u> | CONDITION | NOUN | NOUN | COMMON F SING | 5 | ARG | _ | _ |
| 7 | <u>Initiales</u> | INITIALE | ADJ | ADJ | QUALIF ALLVAL PL | 6 | CONTIN+DENOM | _ | _ |
| 8 | <u>À</u> À | PREP | PREP | MONO | 6 | RMOD | _ | _ | |
| 9 | <u>l'</u> LE | ART | ART | DEF ALLVAL SING | 8 | ARG | _ | _ | |
| 10 | <u>Identique</u> | IDENTIQUE | ADJ | ADJ | QUALIF ALLVAL SING | 9 | ARG | _ | _ |
| 11 | <u>2.0 >2.0></u> | NUM | NUM | _ | 1 | APPOSITION | _ | _ | |
| 12 | . #\. | PUNCT | PUNCT | _ | 1 | END | _ | _ | |
| 1 | <u>Creative</u> | CREATIVE | NOUN | NOUN | PROPER | 4 | SUBJ | _ | _ |
| 2 | <u>Commons</u> | COMMONS | NOUN | NOUN | PROPER | 1 | CONTIN+DENOM | _ | _ |
| 3 | <u>n'</u> NE | ADV | ADV | NEG | 4 | RMOD | _ | _ | |
| 4 | <u>est</u> ÊTRE | VERB | VERB | MAIN IND PRES INTRANS 3 SING | 0 | TOP | _ | _ | |
| 5 | <u>pas</u> PAS | ADV | ADV | NEG | 4 | RMOD | _ | _ | |
| 6 | <u>un</u> UN | ART | ART | INDEF M SING | 4 | PREDCOMPL+SUBJ | _ | _ | |
| 7 | <u>cabinet</u> | CABINET | NOUN | NOUN | COMMON M SING | 6 | ARG | _ | _ |
| 8 | <u>d'</u> DE | PREP | PREP | MONO | 7 | RMOD | _ | _ | |
| 9 | <u>avocats</u> | AVOCAT | NOUN | NOUN | COMMON M PL | 8 | ARG | _ | _ |
| 10 | <u>et</u> ET | CONJ | CONJ | COORD COORD | 4 | COORD+BASE | _ | _ | |
| 11 | <u>ne</u> NE | ADV | ADV | NEG | 12 | RMOD | _ | _ | |
| 12 | <u>fournit</u> | FOURNIR | VERB | VERB | MAIN IND REMPAST TRANS 3 SING | 10 | COORD2ND+BASE | _ | _ |
| 13 | <u>pas</u> PAS | ADV | ADV | NEG | 12 | RMOD | _ | _ | |
| 14 | <u>de</u> DE | ART | ART | INDEF ALLVAL ALLVAL | 12 | OBJ | _ | _ | |
| 15 | <u>services</u> | SERVICE | NOUN | NOUN | COMMON M PL | 14 | ARG | _ | _ |
| 16 | <u>de</u> DE | PREP | PREP | MONO | 15 | RMOD | _ | _ | |
| 17 | <u>conseil</u> | CONSEIL | NOUN | NOUN | COMMON M SING | 16 | ARG | _ | _ |
| 18 | <u>juridique</u> | JURIDIQUE | ADJ | ADJ | QUALIF ALLVAL SING | 17 | RMOD | _ | _ |
| 19 | . #\. | PUNCT | PUNCT | _ | 4 | END | _ | _ | |

2.2 Descriptif des données ParTUT

Nous insérons ci-dessous les documents fournis sur la page du projet parTUT

The annotation guidelines are the same as those used for TUT:

[Syntactic categories](#)

the Part of Speech tagset of the TUT corpus

[Labels of the edges](#)

the list of the grammatical relations labelling the dependency edges of the second release of TUT corpus

For language-specific annotation criteria for English and French, see the document below:

[Linguistic notes](#)

2.3 Transcodage des données ParTUT

Les 2 fichiers au format CONL du volet EN et du volet FR sont transcodés sous la forme d'une base textométrique via un script perl.

Le fichier résultant contient :

- une trame textométrique (une segmentation annotée) composée tout d'abord du volet EN puis du volet FR.
- un cadre textométrique décrivant la partition construite pour contraster les 2 volets.

2.4 La base *ParTUT2Trameur*

Au final, chaque item de la *Trame* (i.e chaque mot) est décrit par 9 annotations :

L' une des des personnes] qui vient d' être assassinée au au Sri Lanka est M. Kumar Ponnambalam , qui avait rendu visite au au Parlement euk

```
Position:<45514>
Forme:<personnes>|Freq:12
Lemme:<PERSONNE>|Freq:39
Cat:<NOUN>|Freq:10750
a-00004:<4>|Freq:1633
a-00005:<NOUN>|Freq:10750
a-00006:<COMMON|F|PL>|Freq:667
a-00007:<ARG(45512)>|Freq:1
a-00008:<_>|Freq:45869
a-00009:<_>|Freq:45869
```

La figure précédente met au jour l'item en position 45 514 et ses 9 annotations. Et la suivante, la zone miroir (dans le volet EN) :

One of the people assassinated very recently in Sri Lanka was Mr Kumar Ponnambalam , who had visited the European Parliament just

```
Position:<371>
Forme:<people>|Freq:31
Lemme:<PEOPLE>|Freq:36
Cat:<NOUN>|Freq:10750
a-00004:<3>|Freq:1819
a-00005:<NOUN>|Freq:10750
a-00006:<COMMON|N|PL>|Freq:1136
a-00007:<ARG(369)>|Freq:1
a-00008:<_>|Freq:45869
a-00009:<_>|Freq:45869
```

| n°Annotation | Annotation trameur | Contenu parTUT |
|--------------|--------------------|-------------------|
| 1 | Forme | Token |
| 2 | Lemme | Lemma |
| 3 | Cat | POS |
| 4 | a-00004 | Id phrase |
| 5 | a-00005 | (à préciser) |
| 6 | a-00006 | (à préciser) |
| 7 | a-00007 | dépendance |
| 8 | a-00008 | (à préciser) |
| 9 | a-00009 | (à préciser) |

2.5 Le Cadre textométrique

La base finale contient 1 seule partition permettant de contraster les 2 volets de l'alignement :

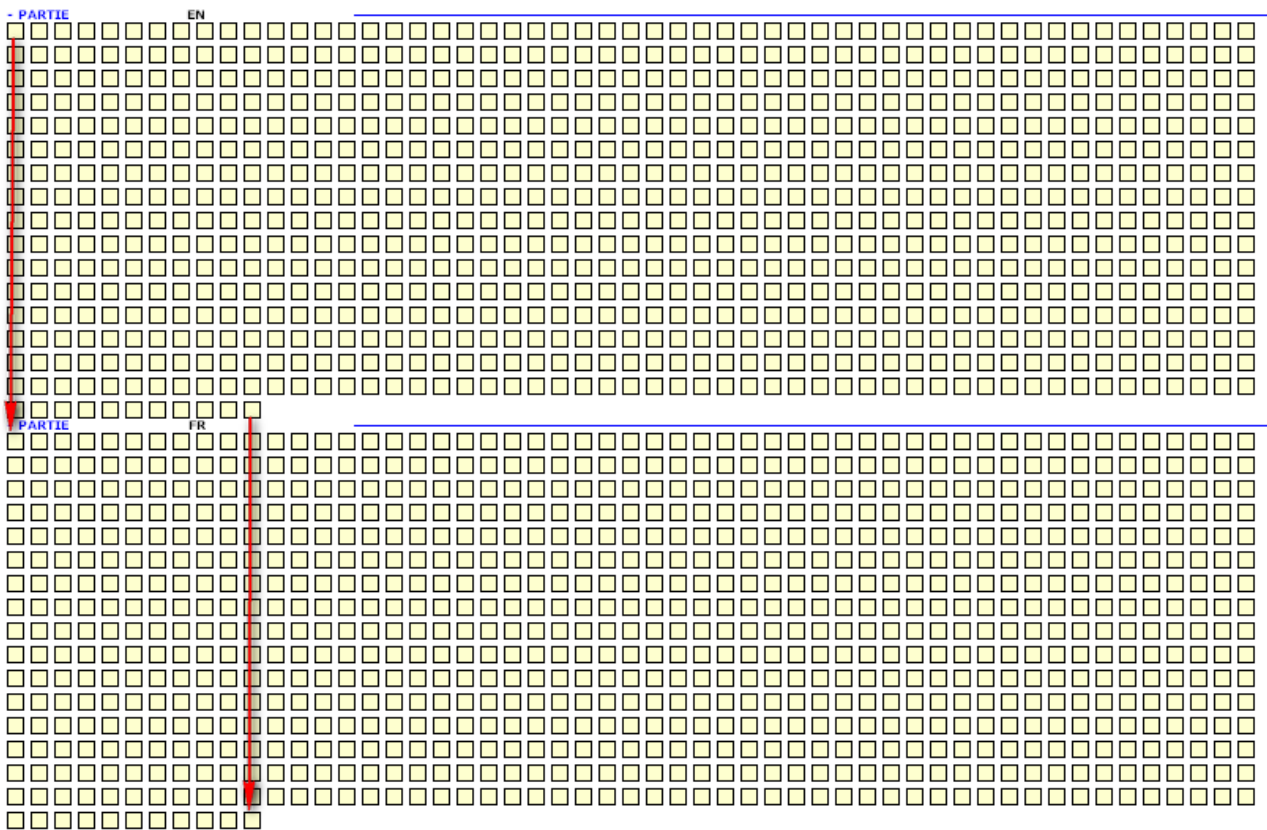
(Partie = EN ; Partie = FR)

2.6 Sections

Le processus de transcodage intègre aussi un marquage de sections :

- après chaque phrase, un caractère délimiteur de section (§) est introduit pour permettre de construire dans le *Trameur* une représentation cartographique de la base sous la forme d'une *carte des sections* des phrases

La carte des phases (avec mise au jour des volets) :



Les 2 volets sont alignés : chaque section-phrase du volet EN a un miroir dans le volet FR

----- PARTIE=EN -----

Resumption of the session . § I **declare resumed** the session of the European Parliament adjourned the session . § I **declare** resumed **the** session of the European Parliament adjourned on I would like once again to **wish you** a happy new year in the hope would like once again to **wish you a** happy new year in the hope that year in the hope that you **enjoyed a** pleasant festive period . § You festive period . § You have **requested a** debate on this subject in the course I should like to **observe a** minute 's silence , as a number of Members The House rose and **observed a** minute 's silence) . § Madam President Mr Kumar Ponnambalam , who had **visited the** European Parliament just a few months ago you , Madam President , to **write a** letter to the Sri Lankan President **expressing** Madam President , to **write a** letter **to** the Sri Lankan President **expressing** Parliament 's to the Sri Lankan President **expressing** Parliament 's regret at his and the other violent violent deaths in Sri Lanka and **urging her** to **do** everything she possibly can to Sri Lanka and **urging** her to **do everything** she possibly can to **seek a** peaceful **do** everything she possibly can to **seek a** peaceful reconciliation to a very difficult situation of the type you have just suggested **would** be entirely appropriate . § If

(les premiers contextes dans le volet EN)

----- PARTIE=FR -----

Reprise de la session . § Je **déclare** de la session . § Je **déclare reprise** la session du du Parlement européen qui la session . § Je **déclare** reprise **la** session du du Parlement européen qui avait le vendredi 17 décembre dernier et je **vous renouvelle** tous mes vux en **espérant** que décembre dernier et je **vous renouvelle** tous **mes** vux en **espérant** que vous avez **passé** vous **renouvelle** tous mes vux en **espérant que** vous avez **passé** de bonnes vacances . vux en **espérant** que vous avez **passé de** bonnes vacances . § Vous avez bonnes vacances . § Vous avez **souhaité un** débat à ce sujet dans les prochains , comme un certain nombre de collègues **me l' ont demandé** , que nous **observions** comme un certain nombre de collègues **me l' ont demandé** , que nous **observions** une de collègues **me l' ont demandé** , **que** nous **observions** une minute de silence pour l' ont **demandé** , que nous **observions une** minute de silence pour toutes les victimes qui ont été touchés . § Je **vous invite** à vous lever pour cette minute (Le Parlement , debout , **observe une** minute de silence) . § par la presse et par la télévision **que** plusieurs attentats à la bombe et crimes vous pas , Madame la Présidente , **qu' il** conviendrait d' **écrire** une lettre au Présidente , qu' il conviendrait d' **écrire une** lettre au au président du du Sri qu' il conviendrait d' **écrire** une lettre **au** au président du du Sri Lanka pour au président du du Sri Lanka pour **lui communiquer** que le Parlement **déplore** les morts du du Sri Lanka pour lui **communiquer que** le Parlement **déplore** les morts violentes , pour lui **communiquer** que le Parlement **déplore les** morts violentes , dont celle de M. celle de M. Ponnambalam , et pour **l' inviter** instamment à **faire** tout ce qui pour l' **inviter** instamment à **faire** tout **ce** qui est en son pouvoir pour **chercher**

(les premiers contextes dans le volet FR)