

Séminaire interne, 18 mars 2016, Université d'Artois

Analyse quantitative de corpus annotés en traduction

Maria ZIMINA

CLILLAC-ARP, Université Paris Diderot-Paris 7, Paris, France

mzimina@eila.univ-paris-diderot.fr



Plan : première partie

Analyse contrastive sur **corpus multilingues**

- Alignements multilingues, bi-texte
- Limites de l'approche centrée sur le standard TMX (Translation Memory eXchange).
- Approche textométrique : modèle de données ***Trame/Cadre***
- Bi-texte : interactions via *ressources textuelles incrémentales*
- *Le Trameur* <http://www.tal.univ-paris3.fr/trameur/> => démonstrations

« **Contraster** » en linguistique : comparer et/ou différencier ?

- comparaison des langues les plus parlées
- mises en contact des dialectes
- analyse des niveaux de langue différents
- analyse contrastive des différents types de discours
- contrastes entre des formulations langagières et non langagières (sémiotique)
- ...

Objets de recherche complexes...

Multiplés cadres théoriques

- les approches de la grammaire traditionnelle
- les procédures structuralistes
- les principes de la grammaire générative
- les approches de la linguistique cognitive
- La linguistique contrastive
- Les théories de la traduction
- L'analyse de discours
- ...

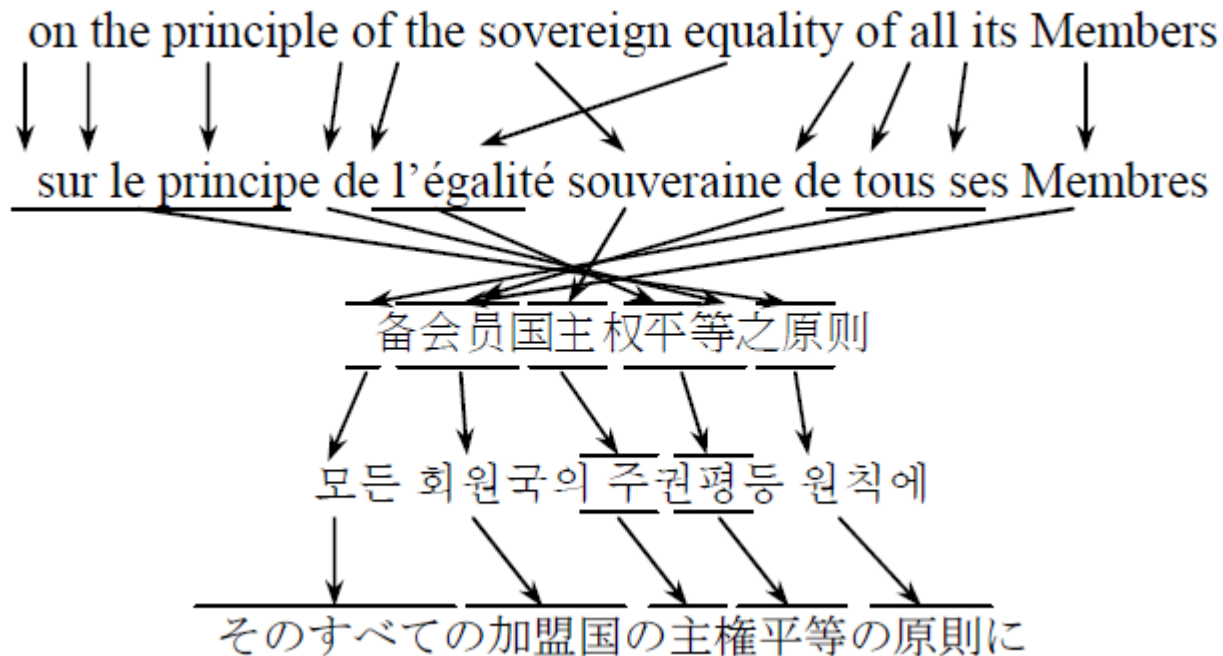
Annotations...

Corpus multilingues

- Typologies textuelles
- Corpus parallèles/comparables
- Unidirectionnels/bidirectionnels
- Synchroniques/diachroniques
- ...

Choix et conséquences...

Alignements multilingues



CHO Joon-Hyung, 2009

correspondances sous-phrastiques ??

What is **Bi-text**?

"One way to describe bi-text [...] is to say that it is ST and TT as they co-exist in the translator's mind at the moment of translating. Another way of putting it is to say that a bi-text is not two texts but a single text in two dimensions, each of which is a language. [...]"

To semioticians, by the way, I submit that bi-text falls into the same paradigm as intertext, in that it is a construct of two or more related texts."

Bi-text, a new concept in translation theory.

Brian Harris, 1988



Version de l'Iliade bilingue, texte en grec et traduction latine (Biblioteca Apostolica Vaticana)

Approche traditionnelle : TMX

```
<?xml version="1.0" encoding="UTF-8"?>
<tmx version="1.4">
  <header
    creationtool="Stingray"
    creationtoolversion="1.0-1"
    srclang="en"
    adminlang="en"
    datatype="xml"
    o-tmf="XLIFF 1.2"
    segtype="block"/>
  <body>
    <tu tuid="818265400-0-48">
      <tuv xml:lang="en"><seg>TMX Files [*.tmx]</seg></tuv>
      <tuv xml:lang="zh-CN"><seg>TMX文件[*.tmx]</seg></tuv>
    </tu>
    <tu tuid="818265400-0-49">
      <tuv xml:lang="en"><seg>All Files [*.*)</seg></tuv>
      <tuv xml:lang="zh-CN"><seg>全部文件[*.*)</seg></tuv>
    </tu>
    <tu tuid="818265400-0-50">
      <tuv xml:lang="en"><seg>Selected file is not a TMX document.</seg></tuv>
      <tuv xml:lang="zh-CN"><seg>已选定的文件不是(翻译存储交换)TMX文件.</seg></tuv>
    </tu>
  </body>
</tmx>
```

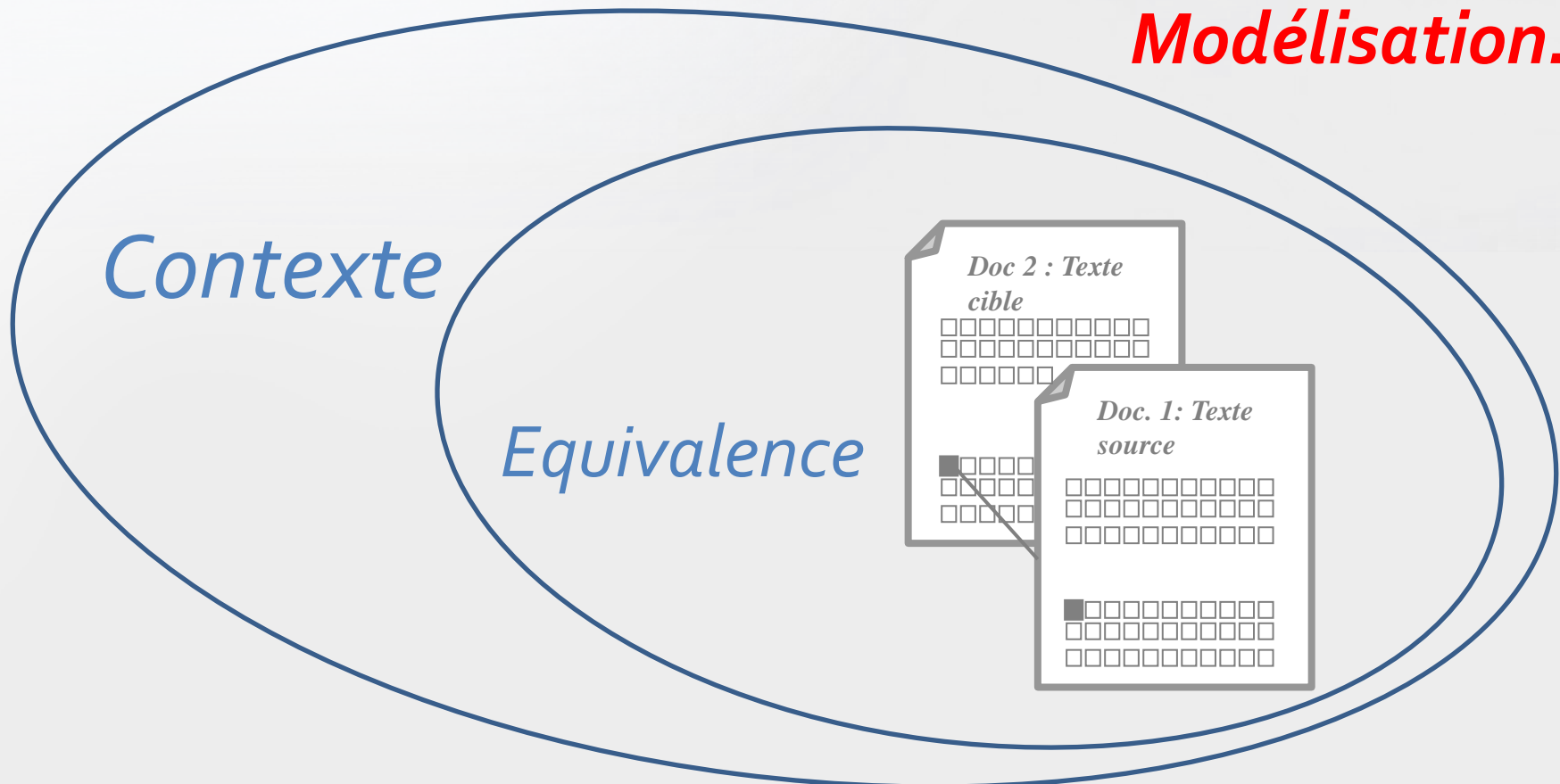
**Translation Memory eXchange (TMX) :
tableau de correspondances**

<http://www.maxprograms.com/articles/tmx.html>

Bi-texte informatisé

"If meaning is function in context, [...] then equivalence of meaning is equivalence of function in context" (Halliday, 1992)

Modélisation...

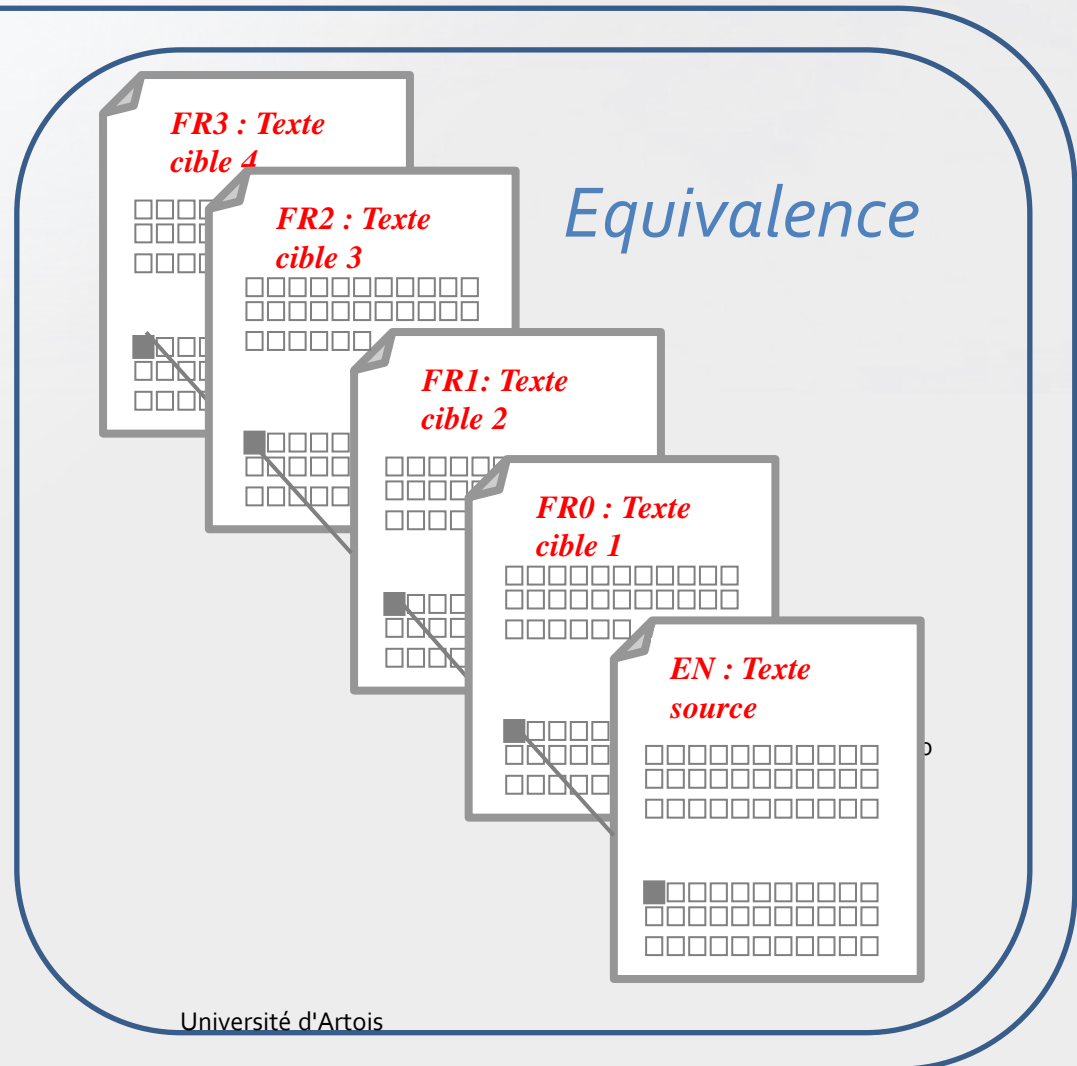


Bi-texte *Obama* (EN / FR₀FR₁FR₂FR₃)

Contexte

EN: le discours original en anglais prononcé par B. Obama le 20 janvier 2009 à Washington, publié sur le site de *The New York Times*

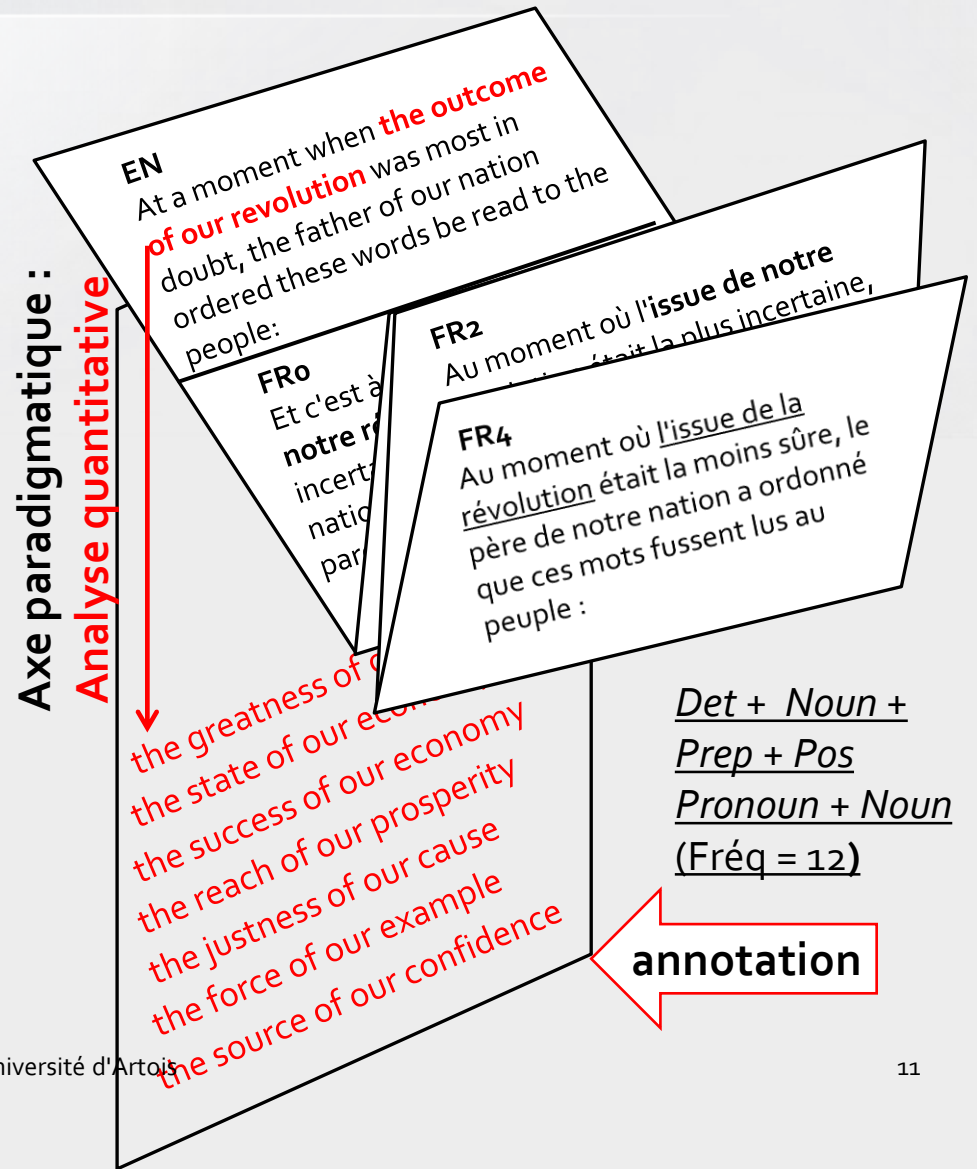
FR₀₋₁₋₂₋₃: différentes traductions françaises de ce discours (*Maison Blanche*, journaux et sites français...)



Limites de l'approche centrée sur le TMX

Translation Memory eXchange

<p>EN ← → FRo</p> <p>At a moment when the outcome of our revolution was most in doubt, the father of our nation ordered these words be read to the people:</p>	<p>FRo</p> <p>Et c'est à l'heure où l'issue de notre révolution était le plus incertaine que le père de notre nation ordonna que les paroles suivantes fussent lues à la population :</p>
<p>"Let it be told to the future world that in the depth of winter, when nothing but hope and virtue could survive, that the city and the country, alarmed at one common danger, came forth to meet it. "</p>	<p>" Qu'il soit dit au monde à venir (...) qu'au plus profond de l'hiver, alors que rien ne pouvait survivre hormis l'espoir et la vertu, que cette ville et ce pays, alertés par un danger commun, se levèrent à sa rencontre. "</p>



Modèle de données *Trame/Cadre* issu des recherches menées en *textométrie*

La *textométrie* regroupe l'ensemble des méthodes quantitatives permettant d'opérer des **réorganisations formelles de la séquence textuelle** et des **analyses statistiques** portant sur l'ensemble des unités textuelles d'un corpus.

Architecture *Trame* / *Cadre*

Le texte dans le Trameur

1. Une segmentation
2. Systèmes de parties

on "découpe" des parts

Un texte est segmenté
en unités
(les mots en général)

Pour chaque mot

etc...

P.O.S

lemmes

Formes

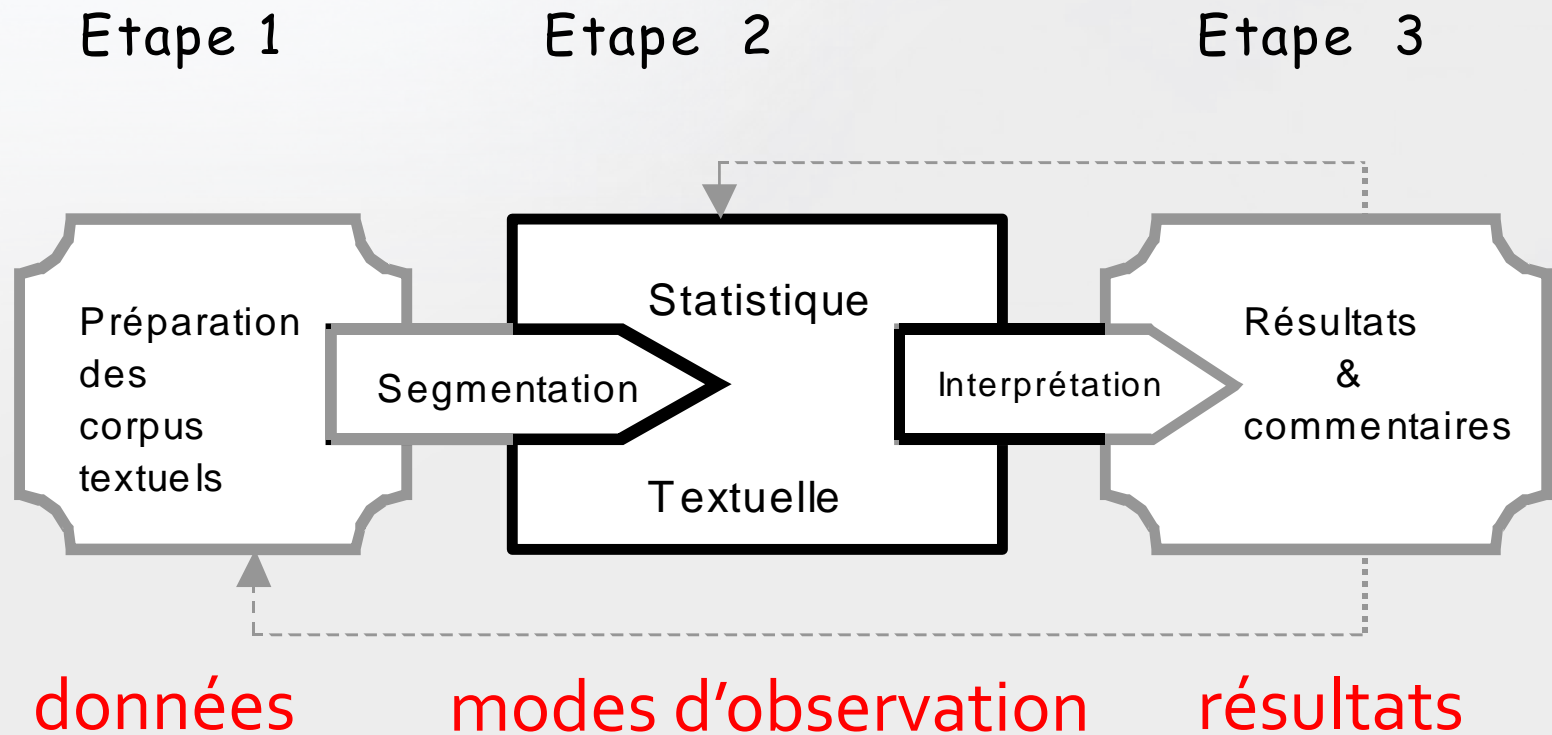
On compte des « contenus » dans des « contenants »

```
<trame>
<codage>utf-8</codage>
<delimiteur><! [CDATA[. , ; ! ? / _ - ' ( ) [ ] { } $ % ! * > < = +
« » → ] ] ></delimiteur>
<items>
<item type="delim" pos="1"><f>RETURN</f><c>DELIM</c><l>RETURN</l></item>
<item type="delim" pos="2"><f> .</f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="3"><f>My</f><c>PP$</c><l>my</l></item>
<item type="delim" pos="4"><f> .</f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="5"><f>fellow</f><c>JJ</c><l>fellow</l></item>
<item type="delim" pos="6"><f> .</f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="7"><f>citizens</f><c>NNS</c><l>citizen</l></item>
.../
<item type="forme" pos="30913"><f>aux</f><c>PRP_det</c><l>au</l></item>
<item type="delim" pos="30914"><f> .</f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="30915"><f>générations</f><c>NOM</c><l>génération</l></item>
<item type="delim" pos="30916"><f> .</f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="30917"><f>futures</f><c>ADJ</c><l>futur</l></item>
<item type="delim" pos="30918"><f> .</f><c>DELIM</c><l> .</l></item>
<item type="delim" pos="30919"><f> .</f><c>DELIM</c><l>BLANK</l></item>
<item type="delim" pos="30920"><f>RETURN</f><c>DELIM</c><l>RETURN</l></item>
<item type="delim" pos="30921"><f>$</f><c>DELIM</c><l>$</l></item>
<item type="delim" pos="30922"><f>RETURN</f><c>DELIM</c><l>RETURN</l></item>
<item type="delim" pos="30923"><f>$</f><c>DELIM</c><l>$</l></item>
<item type="delim" pos="30924"><f>RETURN</f><c>DELIM</c><l>RETURN</l></item>
<item type="delim" pos="30925"><f>$</f><c>DELIM</c><l>$</l></item>
<item type="delim" pos="30926"><f> .</f><c>DELIM</c><l>BLANK</l></item>
</items>
</trame>
<cadre>
<access>
<partition nom="volet">
<p n="EN" d="1" f="5467" nd="1" nf="2"/>
<p n="FR0" d="5467" f="11835" nd="3" nf="4"/>
<p n="FR1" d="11835" f="18423" nd="5" nf="6"/>
<p n="FR2" d="18423" f="24667" nd="7" nf="8"/>
<p n="FR3" d="24667" f="30926" nd="9" nf="10"/>
</partition>
</access>
</cadre>
</baselexicométrique>
```

Élément de la
Trame

Cadre

Textométrie : procédures

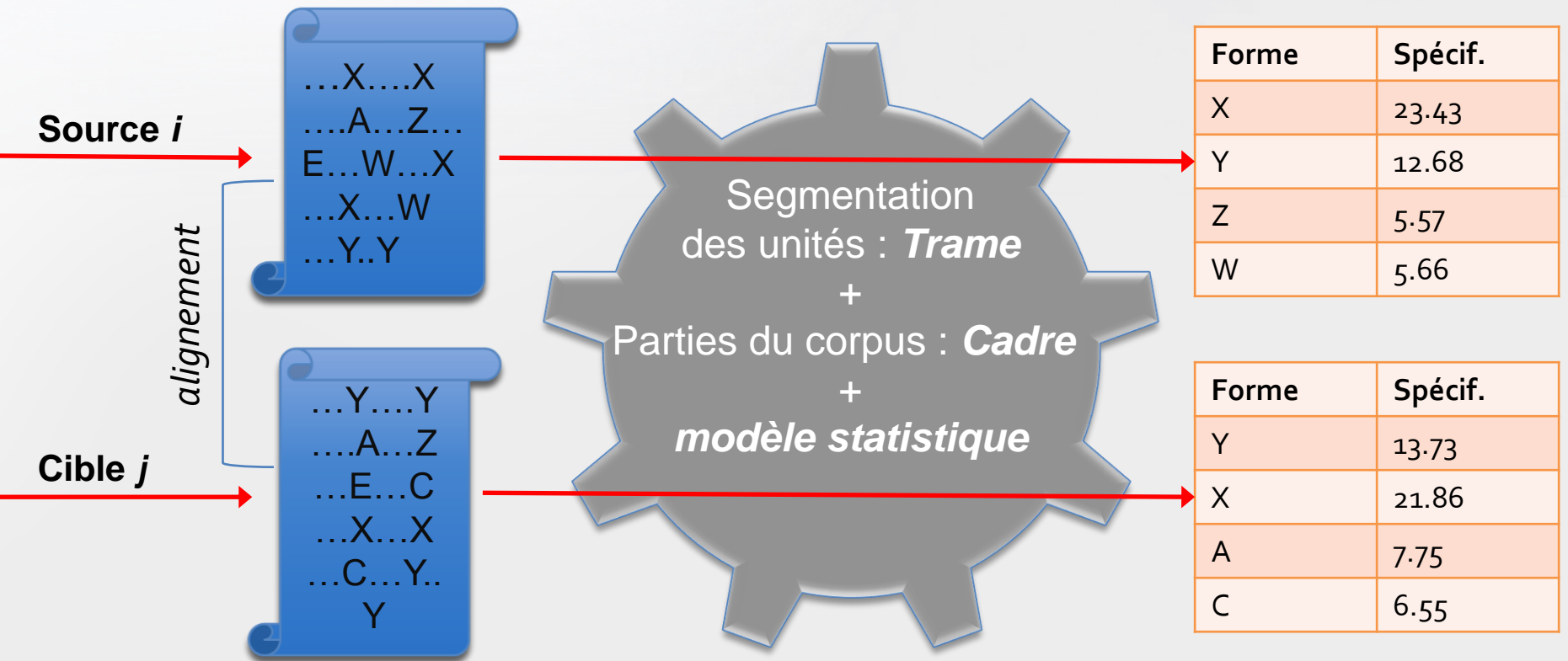


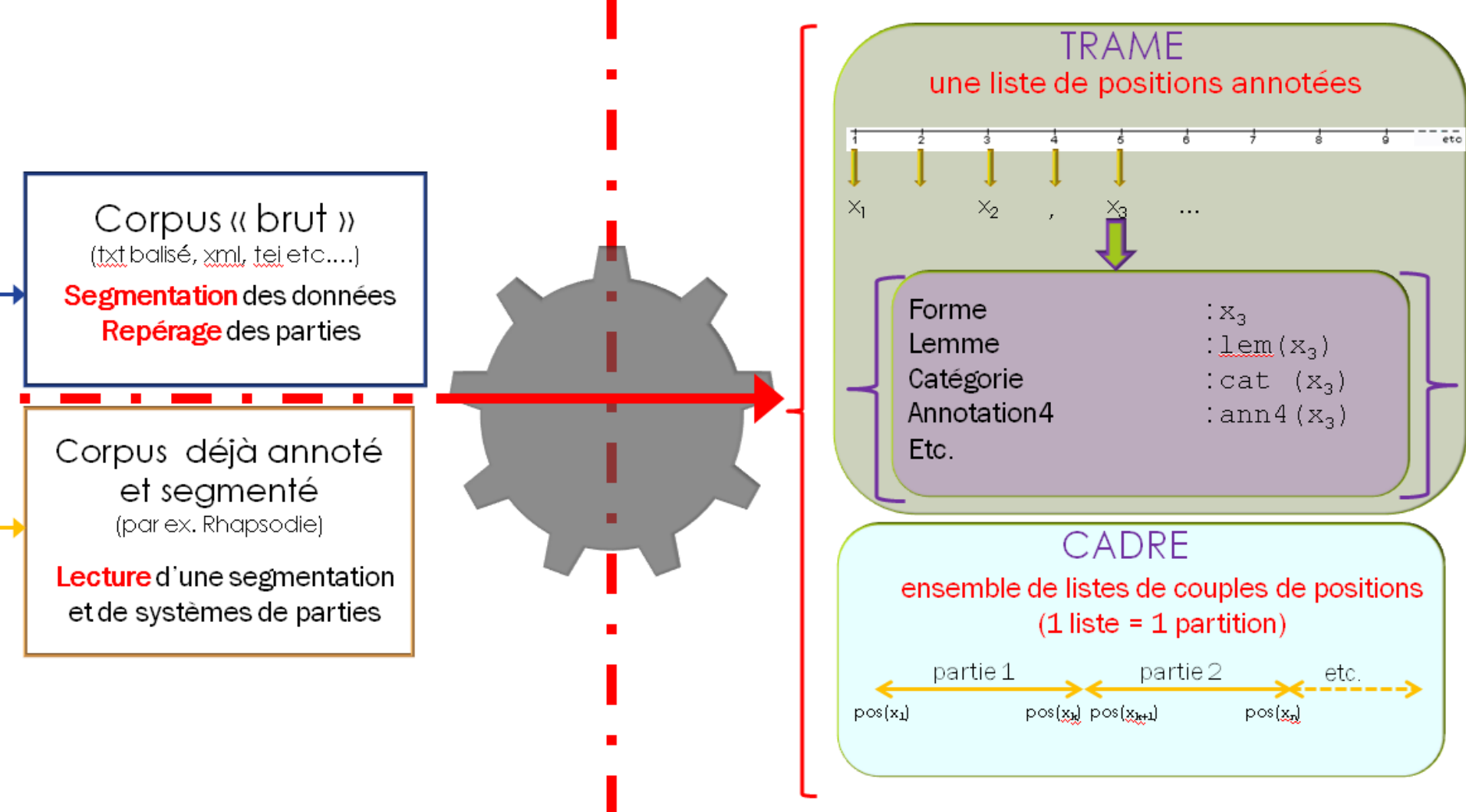
Lamalle *et al.*,
2006

Analyse quantitative du bi-texte

Moteur textométrique

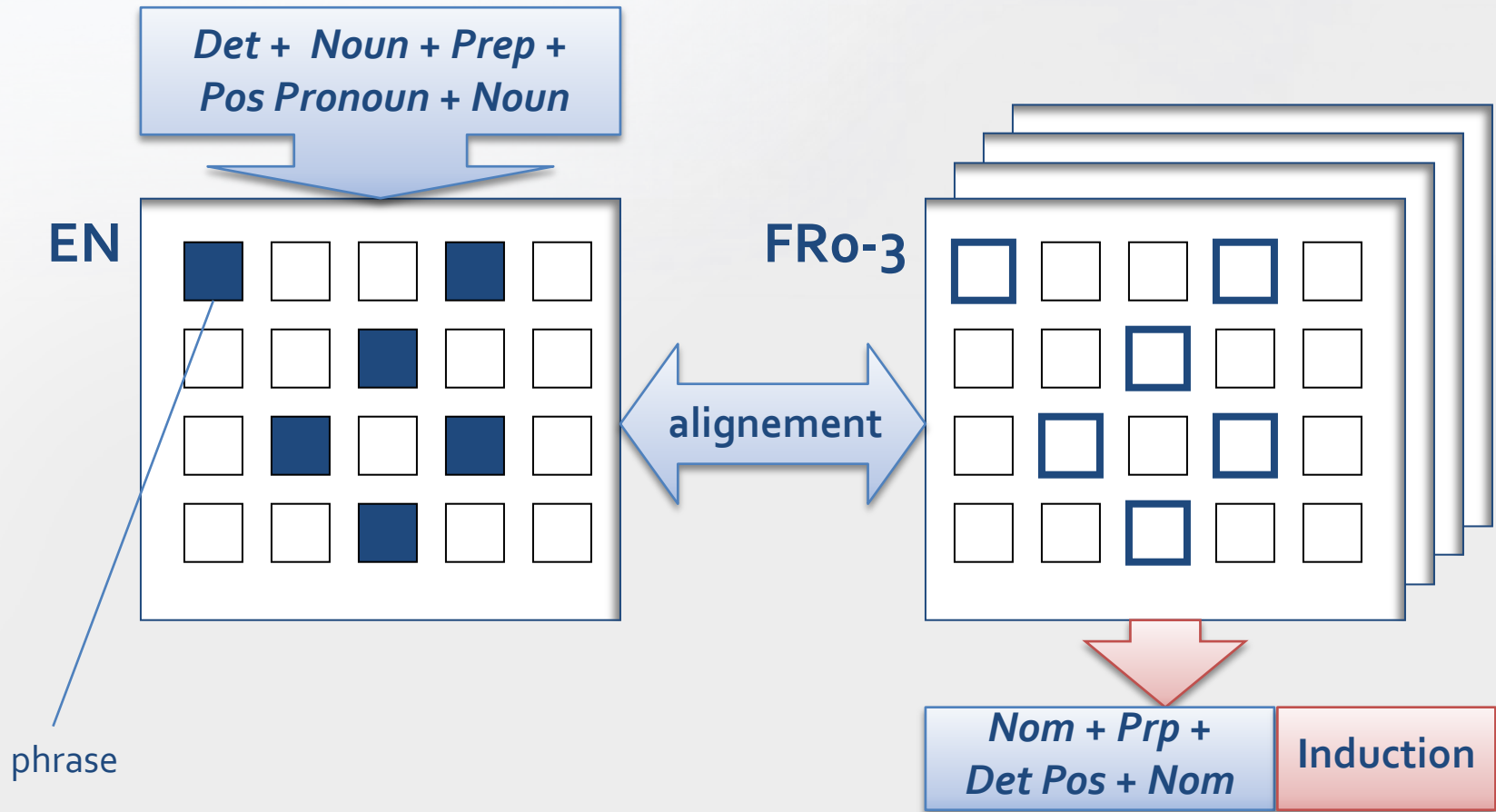
Tableaux statistiques
calculés dynamiquement





Correspondances induites dynamiquement

Corpus *Obama*



Alignements multiples

Corpus *Obama*

- *the success of our economy*

- *réussite de notre économie*

- *succès de notre économie*

- *the justness of our cause*

- *justesse de notre cause*

- *the meaning of our liberty*

- *sens de notre liberté*

- *signification de notre liberté*

Bi-texte : base textométrique dans **Le Trameur**

- volet EN

- volet FR0

- volet FR1

- volet FR2

- volet FR3

Forme	Ind-Sp	FQ	fq
NOM PRP DET POS NOM	21	83	34
NOM PRP DET_POS	21	90	35
PRP DET_POS	19	151	43
DET_ART NOM PRP DET_POS NOM	19	52	26
DET_POS	18	308	60
DET_ART NOM PRP DET_POS	18	56	26
PRP DET_POS NOM	18	133	39
DET_POS NOM	18	272	56

Det + Noun + Prep +
Pos Pronoun + Noun

sur marqueur de page : sélection 5 sections | Shift-control-clic sur marqueur de page : sélection 25 sections (1 ligne)

N° Sect. : 62: (2519, 2574) Annotation : 1 Aperçu : 50 Nb Volets 5 BiText

The success of our economy has always depended not just on the size of our gross domestic product, but on the reach of our prosperity;
\$

La réussite de notre économie a toujours dépendu non seulement du niveau de notre produit intérieur brut, mais aussi de l'étendue de notre prospérité ;
\$

Notre réussite économique n'a pas été dépendante uniquement du montant de notre produit intérieur brut, mais également de l'étendue de notre prospérité,
\$

Le succès de notre économie n'est pas uniquement fonction de la taille de notre produit intérieur brut. Il dépend aussi de l'étendue de notre prospérité,
\$

Le succès de notre économie a toujours dépendu, non seulement de la taille de notre PIB, mais de l'étendue de notre prospérité,
\$

Systeme d'annotations dynamiques

The screenshot displays a software interface for dynamic annotations. On the left, a window titled "Correction Item" is open, showing the following fields and values:

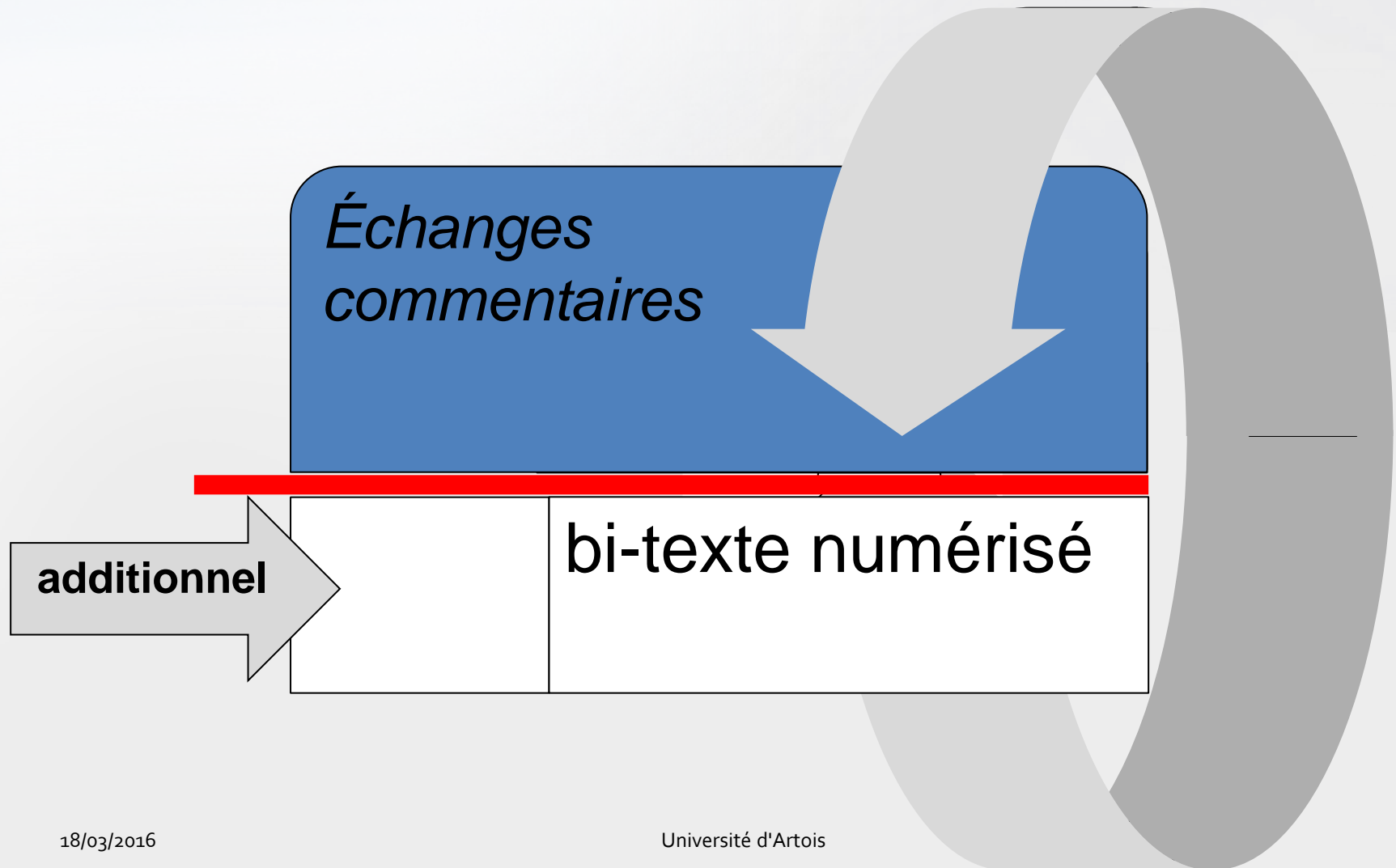
- Forme..... La
- Lemme..... le
- Catégorie..... DET_ART
- commentaire..... à modifier...

Buttons for "Annuler" and "Enregistrer" are visible at the bottom of the dialog. The main interface below features a status bar with the following information: "Nb L. Sections sélectionnées : 0 N° Sect. : 62: (2519,2574) Annotation : 1 Aperçu : 50 Nb Volet 5 BiText".

The main workspace is divided into five vertical panels, each displaying a different translation of the same source text: "The success of our economy has always depended not just on the size of our gross domestic product, but on the reach of our prosperity; \$".

- Panel 1: Original English text.
- Panel 2: "La réussite de notre économie a toujours dépendu non seulement du niveau de notre produit intérieur brut, mais aussi de l'étendue de notre prospérité ; \$"
- Panel 3: "Notre réussite économique n'a pas été dépendante uniquement du montant de notre produit intérieur brut, mais également de l'étendue de notre prospérité, \$"
- Panel 4: "Le succès de notre économie n'est pas uniquement fonction de la taille de notre produit intérieur brut. Il dépend aussi de l'étendue de notre prospérité, \$"
- Panel 5: "Le succès de notre économie a toujours dépendu, non seulement de la taille de notre PIB, mais de l'étendue de notre prospérité, \$"

Ressources incrémentales, interactions



Conclusions (première partie)

- Les **mémoires de traduction** axés sur le format **TMX** sont statiques.
- Le modèle de données textométrique *Trame/Cadre* permet de **représenter dynamiquement des correspondances sous-phrastiques**.
- Ce modèle permet d'**interagir avec le bi-texte** à plusieurs niveaux d'analyse (mots, syntagmes, phrases, cadres discursifs, etc.).

Plan : deuxième partie

LTD (Lecture Textométrique Différentielle) de corpus annotés

- Annotations multiples : corpus d'erreurs de traduction **ER-TRAD-SP** (N. Kubler *et al.*, 2015)
- LTD : principes de différenciation
- Lecture bilingue à l'écran
- Détection des **difficultés spécifiques** des apprentis traducteurs : *démonstrations*

Corpus d'erreurs de traduction **ER-TRAD-SP** (N. Kübler, M. Pecman, A. Mestivier-Volanschi)

- Les traductions en français sont réalisées par les étudiants de Master 1 ILTS de l'UFR EILA (Paris 7) en collaboration avec des étudiants et les enseignants de l'UFR des Sciences de la Terre et des Planètes (STEP) de la même université
- Chaque année (projet initié en 2013), les traductions portent sur une sélection d'une vingtaine de **d'articles scientifiques** en anglais, de date récente, pris dans le domaine des Sciences de la terre et des planètes (extraits de 500 mots).

Corpus **ER-TRAD-SP**

annoté selon la typologie d'erreurs **MeLLANGE**

- Les traductions sont annotées selon la typologie d'erreurs développée dans le cadre du projet MeLLANGE (Castagnoli *et al.* 2011).
- Distinction entre les erreurs liées au **transfert de contenu** et les **erreurs de langue**. Ces deux catégories sont divisées en sous-catégories du type « Intrusion de la langue source » ou « Terminologie et lexique » qui, à leur tour, regroupent des erreurs plus spécifiques du type « Trop littéral » ou « Collocation incorrecte ».
- L'hierarchie inclut également, à tous les niveaux, des erreurs de type « Défini par l'utilisateur ».

Typologie d'erreurs MeLLANGE

MeLLANGE WP4 Translation Error Typology (version 01/08/2006)

Content Transfer

- Omission (TR-OM)
- Addition (TR-AD)
- Distortion (TR-DI)
- Indecision (TR-IN)
- User-Defined (TR-UD)

SL Intrusion

- Untranslated Translatable (TR-SI-UT)
- Too Literal (TR-SI-TL)
- Units of Weight/Measurement, Dates and Numbers (TR-SI-UN)
- User-Defined (TR-SI-UD)

TL Intrusion

- Translated DNT (TR-TI-TD)
- Too Free (TR-TI-TF)
- User-Defined (TR-TI-UD)

Typologie d'erreurs MeLLANGE

- Language
 - Syntax (LA-SY)
 - Wrong preposition (LA-PR)
 - User-Defined (LA-UD)
- Inflection and Agreement
 - Tense/Aspect (LA-IA-TA)
 - Gender (LA-IA-GE)
 - Number (LA-IA-NU)
 - User-Defined (LA-IA-UD)
- Terminology and Lexis
 - Incorrect (LA-TL-IN)
 - False Cognate (LA-TL-FC)
 - Term Translated by Non-Term (LA-TL-NT)
 - Inconsistent with Glossary (LA-TL-IG)
 - Inconsistent within TT (LA-TL-IT)
 - Inappropriate collocation (LA-TL-IC)
 - User-Defined (LA-TL-UD)
- Hygiene
 - Spelling (LA-HY-SP)
 - Accents or Diacritics (LA-HY-AC)
 - Incorrect Case (Upper/Lower) (LA-HY-CA)
 - Punctuation (LA-HY-PU)
 - User-Defined (LA-HY-UD)
- Register
 - Inconsistent with ST (E.g. Form of Address) (LA-RE-IS)
 - Inappropriate for TT Text Type (LA-RE-IN)
 - Inconsistent within TT (LA-RE-IT)
 - User-Defined (LA-RE-UD)
- Style
 - Awkward (LA-ST-AW)
 - Tautology (LA-ST-TA)
 - User-Defined (LA-ST-UD)

Base textométrique alignée ER-TRAD-SP (M. Zimina)

Le Trameur - Le Métier Lexicométrique @CLA2T-P3 V. 12.108

Cadre Ventilation **Section** Forme-Lemme Catégorie-Tag Segment Cocc Stat Concordance Patron Graphe Relation Sélection Rapport Param

Chargement de la Carte des sections :

Délimiteur de sections : \$ Partie

La carte des sections peut être construite soit en choisissant un délimiteur soit en choisissant une partie du cadre (nom de partie)

Parties
corpus

(Ctrl-Clic : désélection partie)

Recherche Forme sur la carte :

[^\\W]Erreur[\\W] RegExp

(Ecrire motif supra puis Entrée)

Spécificités sur Sections

BI-TEXT

V1 1 V2 2

TMX

Sélection Annotation :

Forme Lemme Catégorie

Annotation sélectionnée : Forme 1

Shift-clic sur carré : affichage | clic-droit sur carré : spécificités | Control-clic sur carré : sélection | Shift-Control-clic sur sélection : désélection

Seuillage : 1 5 10 ++ | Modifier le seuillage : [Alt]

(+/-) : masquage/affichage des sections en cliquant sur la partie

- corpus EN

- corpus FR

Control-clic sur marqueur de page : sélection 5 sections | Shift-control-clic sur marqueur de page : sélection 25 sections (1 ligne)

Nb L. Sections sélectionnées : 0 N° Sect. : 419:(24486,24531) Annotation : 1 Aperçu : 50 Nb Voles 2 BiText

Going beyond density functional theory with a quantum Monte Carlo simulation, melting was obtained at 6900 K (15) at 330 GPa\$

En allant au-delà de la théorie de la fonctionnelle de la densité avec une simulation du quantum de Monte Carlo de 6900

Position:<56877>
Forme:<Monte>|Freq:2
Lemme:<Monte>|Freq:2
Cat:<NAM>|Freq:948
a-00004:<Terme-traduit-par-non-terme_LA-IL-NT>|Freq:167
a-00005:<simulation Monte Carlo quantique ? >|Freq:5
a-00006:<Erreur>|Freq:2243

Shift-Clic : sélection | Clic-droit : édition | Ctrl-Clic : noeud | 2-Clic : graphe | Shift-Clic-droit : relation | Control-Clic-droit : recherche relation

Annotations : 1 2 3 4 5 6

Lecture **T**extométrique **D**ifférentielle (**LTD**) d'erreurs de traduction

- Les textes source et cible sont affichés simultanément à l'écran.
- La *éléments caractéristiques* sont rendus « visibles » au fil des textes par un *système* de surlignage.
- L'affichage tient compte des *répétitions segmentales* sur annotations multiples (formes, lemmes, catégories, etc.).
- Prise en main « globale » de la différentiation à *plusieurs niveaux d'analyse linguistique* (Zimina & Fleury, 2015).

LTD

Spécificités (éléments caractéristiques)

2

Forme	Ind-Sp	FQ	fq
NN Noun, singular or mass	3	2482	197
MD Modal verb	2	87	11
VVP Verb, present, non-3rd p.	2	80	10
FW Foreign Word	2	1	1
NNS Noun plural	2	904	72

1

EN

FR

Erreurs type Syntaxe LA-SY

Control-clic sur marqueur de page : sélection 5 sections | Shift-control-clic sur marqueur de page : sélection 25 sections (1 ligne)

Nb L. : 0 N° Sect. : 179: (11252, 11310) Annotation : 1 Aperçu : 50 ✓ Nb Volet 2 ✓ BiText

The **mechanisms** that trigger these **supereruptions** are elusive because the processes occurring in conventional volcanic systems cannot simply be scaled up to the much larger **magma chambers** beneath **supervolcanoes**§

Les mécanismes à l'origine de ces superéruptions restent élusifs car les processus en œuvre dans les systèmes volcaniques conventionnels **ne peuvent simplement pas être mis** à l'échelle des chambres magmatiques beaucoup plus larges des supervolcans§

Détection des difficultés spécifiques des apprentis traducteurs

ER-TRAD-SP

VOLET : EN	VOLET : FR
However, this would have only a small effect on the calculated ages (Fig 3)§	Cependant, cela aura peu de conséquences sur les âges calculés (Figure 3)§
For instance, using a 15% higher Hf/U ratio changes the ages by only 2 Myr§	Par exemple, utiliser un ratio plus riche en Hf/U de $\approx 15\%$ ne change les âges que d'environ 2 millions d'années§
A first-order estimate for core formation in the Moon (or the impactor) is obtained using a two-stage model§	Une première façon d'estimer la formation du noyau de la Lune (ou de l'impacteur) est d'utiliser une méthode en deux étapes§
This model implies that the bulk silicate Earth and Moon started off with slightly different 182W/184W ratios (,0 5 e units difference at most; see Supplementary Fig 1) and fortuitously evolved to identical present-day 182W/184W ratios§	Cette méthode implique que les parties silicatées de la Terre et de la Lune sont apparues avec des ratios de 182W/184W légèrement différents (maximum de $\approx 0,5$ e unités de différence ; voir Figure 1 en Annexe) et ont évolué par hasard jusqu'à devenir aujourd'hui des ratios de 182W/184W identiques§
However, as the Moon consists predominantly of mantle material with high Hf/W and hence most probably radiogenic 182W/184W, this two-stage model is not valid§	Cependant, étant donné que le manteau de la Lune est constitué majoritairement d'importants ratios Hf/W et donc très probablement de 182W/184W radiogénique, cette méthode à deux étapes ne fonctionne pas§
The initial 182W/184W of the Moon was most probably higher than chondritic, resulting in an age younger than 3,7 Myr and implying that the bulk silicate Moon and Earth must have had indistinguishable initial 182W/184W ratios (Supplementary Fig 1)§	Les ratios initiaux de 182W/184W de la Lune étaient probablement plus élevés que ceux présents dans la chondrite, ce qui implique un âge plus jeune que $\approx 3,7$ millions d'années et implique que les parties silicatées de la Lune et de la Terre doivent avoir des ratios initiaux de 182W/184W identiques (Figure 1 en Annexe)§
A more reliable age constraint for core formation is obtained from the identical 182W/184W ratios of the bulk silicate Moon and Earth in conjunction with their distinct Hf/W ratios and by assuming identical initial 182W/184W ratios for the bulk silicate Moon and Earth§	Une définition plus exacte de l'âge de la formation du noyau est obtenue grâce aux ratios identiques de 182W/184W des parties silicatées de la Lune et de la Terre, mis en relation avec leurs ratios différents de Hf/W ; à condition d'assumer des ratios initiaux identiques de 182W/184W dans les parties silicatées de la Lune et de la Terre§
As shown in Fig 3, core formation in the Moon must have occurred later than at 50 Myr, otherwise the lunar mantle would have a 182W excess relative to the terrestrial mantle§	Comme le montre la Figure 3, la formation du noyau lunaire aurait eu lieu après 50 millions d'années, sinon la manteau lunaire contiendrait davantage de 182W que le manteau terrestre§
We did not include minor and trace elements, because their effect on iron partitioning can be neglected to a first approximation§	Nous n'avons pas inclus d'éléments mineurs et traces, car leur effet sur le partage du fer est négligeable dans une première estimation§
We did the experiments in a diamond-anvil cell (DAC) at pressures from 40 GPa to 120 GPa, using very thin (510 μ m) samples that were melted throughout their thickness using infrared lasers (Fig 1)§	Nous avons fait ces expériences dans une cellule à enclumes de diamant (CED), à des pressions allant de 40 Gpa à 120 Gpa et en utilisant de très petits échantillons (510 μ m) qui ont été fondus à travers leur épaisseur à l'aide de lasers infrarouges (Figure 1)§
Melting criteria are based on the use of in situ X-ray diffraction,§	Les critères de fusion se fondent sur l'utilisation in situ de la diffraction des rayons X§

Conclusions (deuxième partie)

- L'annotation des productions des étudiants et l'analyse quantitative des erreurs de traduction permet un retour efficace sur l'enseignement de la traduction spécialisée.
- La LTD permet d'envisager une amélioration de la méthodologie d'enseignement (identification des difficultés spécifiques dans les textes sources).
- **Perspectives** : prise en compte de l'analyse des *dépendances syntaxiques* (nouvelle annotation).

Démonstrations sur les *treebanks* :

ParTUT

ParTUT (*Parallel Turin University TreeBank*)

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

Le projet a permis la création et l'alignement phrastique des *treebanks* parallèles en italien, anglais et français à l'aide d'un seul schéma unifié.

Seules les ressources en **français** et en **anglais** ont été utilisées dans notre exploration (environ 97 000 occ.).

Démonstrations sur les *treebanks* :

ParTUT

ParTUT (*Parallel Turin University TreeBank*)

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

Le projet a permis la création et l'alignement phrastique des *treebanks* parallèles en italien, anglais et français à l'aide d'un seul schéma unifié.

Seules les ressources en **français** et en **anglais** ont été utilisées dans notre exploration (environ 97 000 occ.).

Les données *ParTUT* utilisent le format *CoNLL* (Conference on Computational Natural Language Learning)

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>1'</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	-	-	

S. Buchholz and E. Marsi,

“CoNLL-X shared task on multilingual dependency parsing”, dans *CoNLL-X'06*, PA, USA, 2006.

Base *ParTUT2Trameur* (1/2)

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

Resumption of the session . \$ I declare resumed

of the session . \$ I declare resumed
the session . \$ I declare resumed t
I would like once again t
would like once again to w
year in the hope that you e
festive period . \$ You have rec
I should like to observe a
The House rose and observed
Mr Kumar Ponnambalam , who had vis
you , Madam President , to write
Madam President , to write a letter t
to the Sri Lankan President *expressing*

Position:<3>
Forme:<of>|Freq:793
Lemme:<OF>|Freq:783
Cat:<PREP>|Freq:6911
a-00004:<1>|Freq:1420
a-00005:<PREP>|Freq:6911
a-00006:<MONO>|Freq:6426
a-00007:<OBJ(1)>|Freq:1
a-00008:<OBJ_EN(1)>|Freq:1
a-00009:<_>|Freq:45869
a-00010:<_>|Freq:45869

Parliament adjourned
liament adjourned on
n the hope
e hope that
od . \$ You
ct in the course
r of Members
am President
ust a few months ago
esident *expressing*
pressing Parliament 's
and the other violent

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

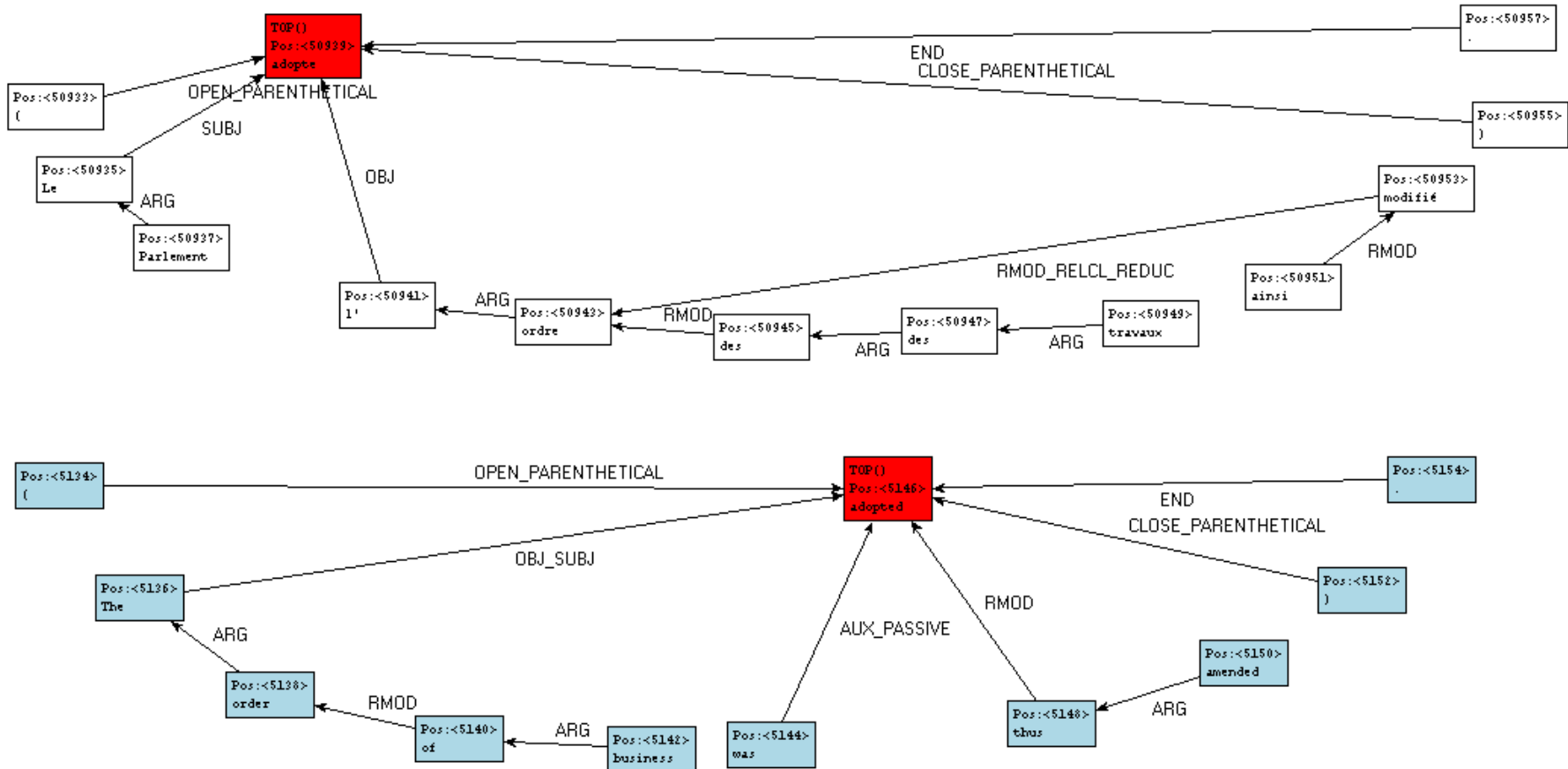
Reprise de la session . \$ Je déclare reprise

de la session . \$ Je déclare reprise
la session . \$ Je déclare repri
le vendredi 17 décembre dernier
décembre dernier et je vous *renou*
vous *renouvelle* tous mes vux en
vux en *espérant* que vous a
bonnes vacances . \$ Vous avez
, comme un certain nombre de c
comme un certain nombre de coll
de collègues me l' ont *demandé*
l' ont *demandé* , que nous *observi*
qui ont été touchés . \$ Je
(Le Parlement , debout , *observe* une minute de silence) . \$

Position:<45142>
Forme:<de>|Freq:1123
Lemme:<DE>|Freq:2074
Cat:<PREP>|Freq:6911
a-00004:<1>|Freq:1420
a-00005:<PREP>|Freq:6911
a-00006:<MONO>|Freq:6426
a-00007:<OBJ(45140)>|Freq:1
a-00008:<OBJ_FR(45140)>|Freq:1
a-00009:<_>|Freq:45869
a-00010:<_>|Freq:45869

européen qui
opéen qui avait
en *espérant* que
vous avez *passé*
bonnes vacances .
Vous avez
les prochains
ous *observians*
observians une
le silence pour
tes les victimes
te minute

Base *ParTUT2Trameur* (2/2)



Références bibliographiques

- **S. Castagnoli, D. Ciobanu, N. Kübler, K. Kunz, A. Volanschi** «Designing a Learner Translator Corpus for Training Purposes». In N. Kübler. (ed) (2011) *Corpora, Language, Teaching, and Resources : From Theory to Practice*. Bern: Peter Lang. 221-248.
- **J-H Cho 2009**. « Traductions franco-coréennes ». *Lexicometrica*, n° spécial *Corpus multilingues*.
- **S. Fleury 2013**. *Le Trameur. Propositions de description et d'implémentation des objets textométriques*. Sorbonne nouvelle – Paris 3. En ligne : <http://www.tal.univ-paris3.fr/trameur/trameur-propositions-definitions-objets-textometriques.pdf>
- **S. Fleury and M. Zimina 2014**. «Trameur: A Framework for Annotated Text Corpora Exploration», *Proceedings of COLING 2014: System Demonstrations*, Dublin, 2014, pages 57-61.
- **M.A.K. Halliday 1992**. « Language theory and translation practice ». *Rivista internazionale di tecnica della traduzione* n°0, pages 15-25.
- **B. Harris 1988**. « Bi-text, a new concept in translation theory ». *Language Monthly* n°54, pages 8-10.
- **C. Lamalle, S. Fleury, A. Salem 2006**. « Vers une description formelle des traitements textométriques ». *Journées Internationales d'Analyse Statistiques des Données Textuelles*, Besançon, 2006.
- **N. Kübler, A. Mestivier, M. Pecman 2015**. « Etude sur l'utilisation des corpus dans l'enseignement de la terminologie et de la traduction spécialisée ». *Terrains de recherche en linguistique appliquée (TRELA)*. Conférence internationale, CLILLAC-ARP, Université Paris Diderot, Paris 8-10 juillet 2015.
- **C. Lamalle, S. Fleury, A. Salem 2006**. « Vers une description formelle des traitements textométriques ». *Journées Internationales d'Analyse Statistiques des Données Textuelles*, Besançon, 2006.
- **M. Zimina and S. Fleury 2014**. « *Approche systémique de la résonance textuelle multilingue* ». *Journées internationales d'Analyse statistique des Données Textuelles*, Paris, juin 2014.