COTRALIS corpus traductologie linguistique société
EA 4028 Textes et Cultures

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

SORBONNE PARIS CITÉ
RECHERCHE ENSEIGNEMENT SUPÉRIEUR

université PARIS DIDEROT PARIS 7

arp CLILLAC

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…

Multiples 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…

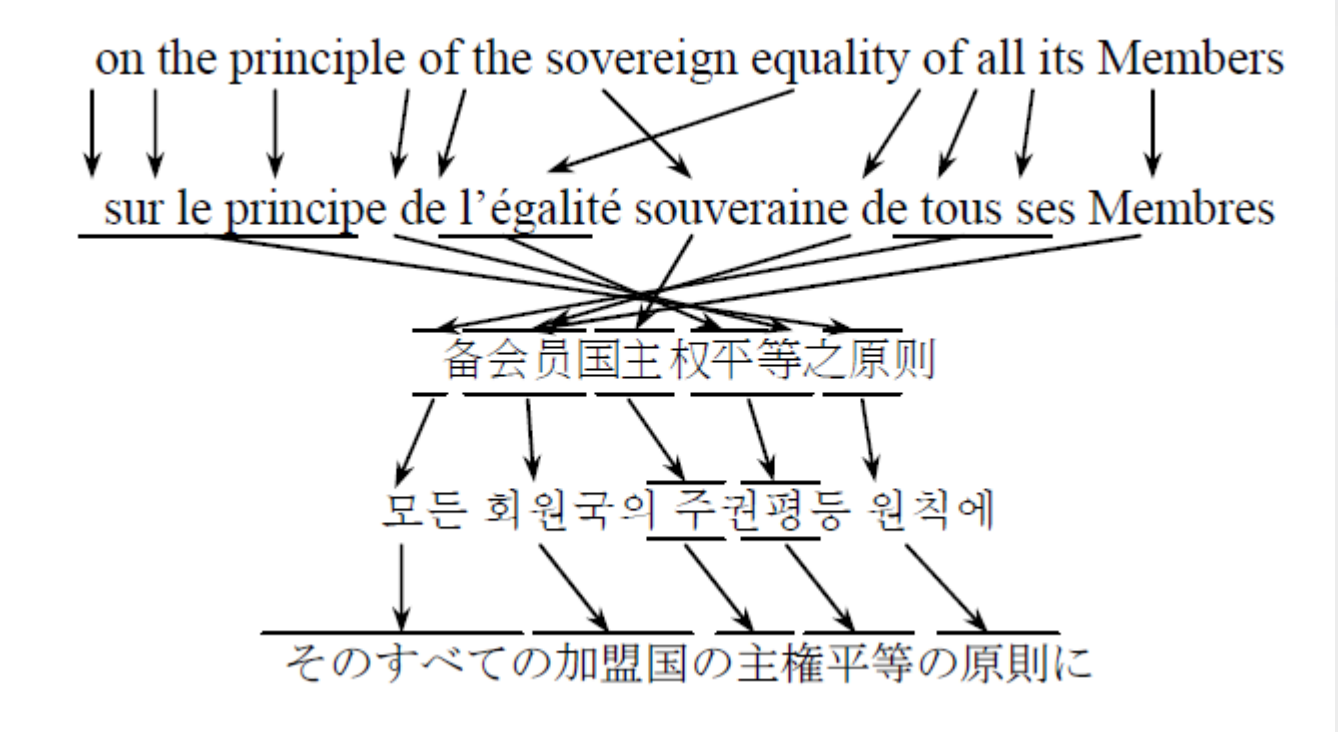
Corpus multilingues

- 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*)

AGENCE NATIONALE DE LA RECHERCHE ANR TransRead ANR-12-CORD-0015

Approche traditionnelle : TMX

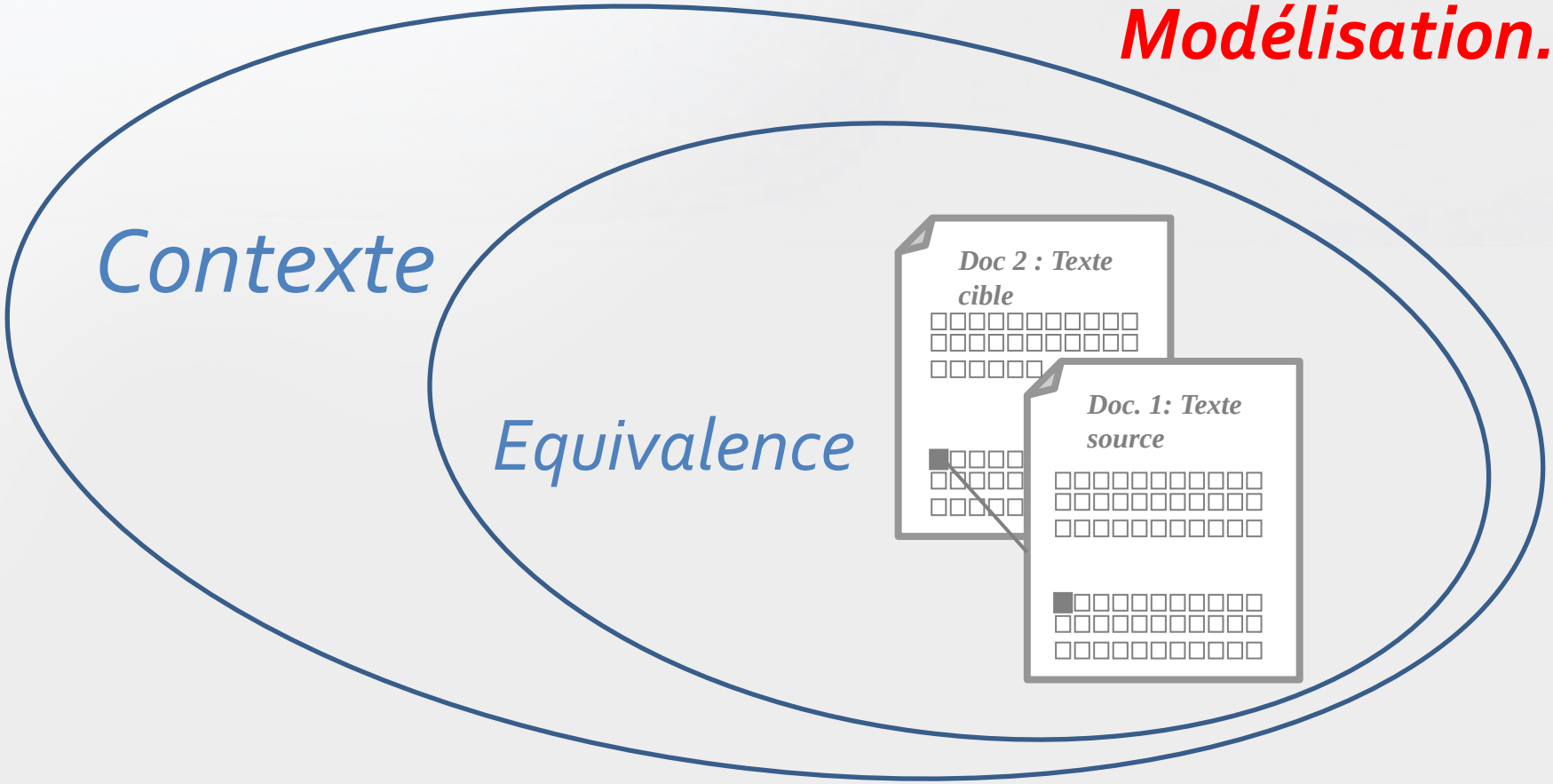
```
<?xml version="1.0" encoding="UTF-8"?>
<tmx version="1.4">
  <header
      creationtool="Stingray"
      creationtoolversion="1.0-1"
      srclang="en"
      adminlang="en"
      datatype="xml"
      o-tmf="XLIFF 1.2"
      segtype="block"/>
  <body>
    <tu tuid="818265400-0-48">
      <tuv xml:lang="en"><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)

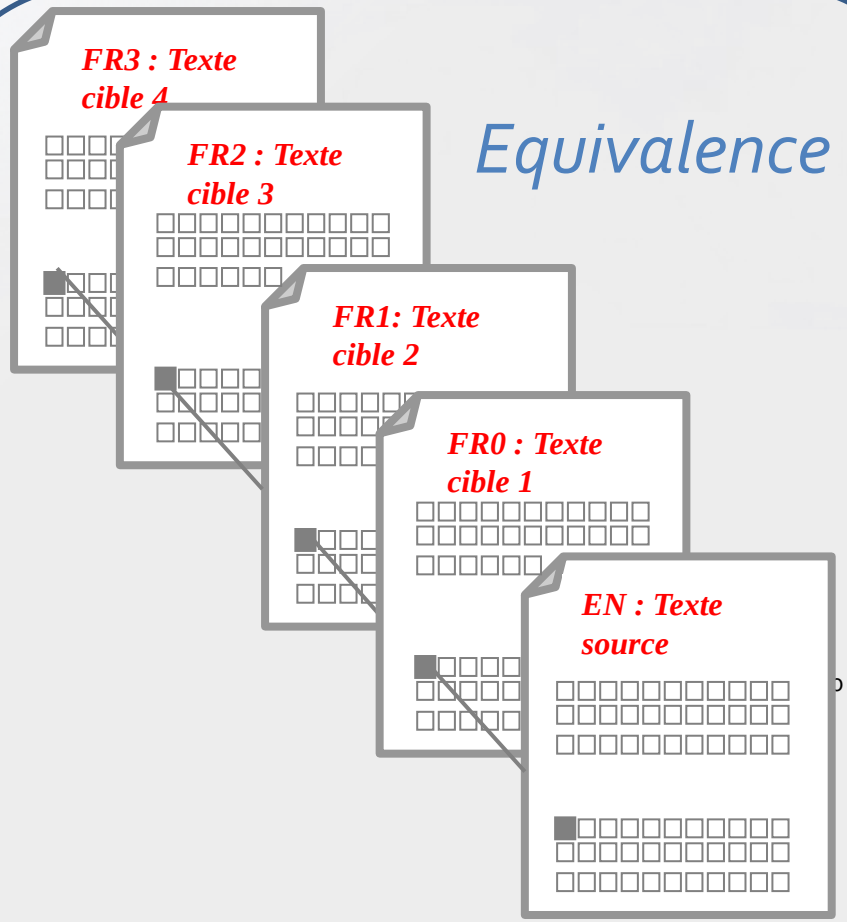


Bi-texte *Obama (EN / FRoFR1FR2FR 3)*

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*

FRo-1-2-3: différentes traductions françaises de ce discours (*Maison Blanche*, journaux et sites français…)

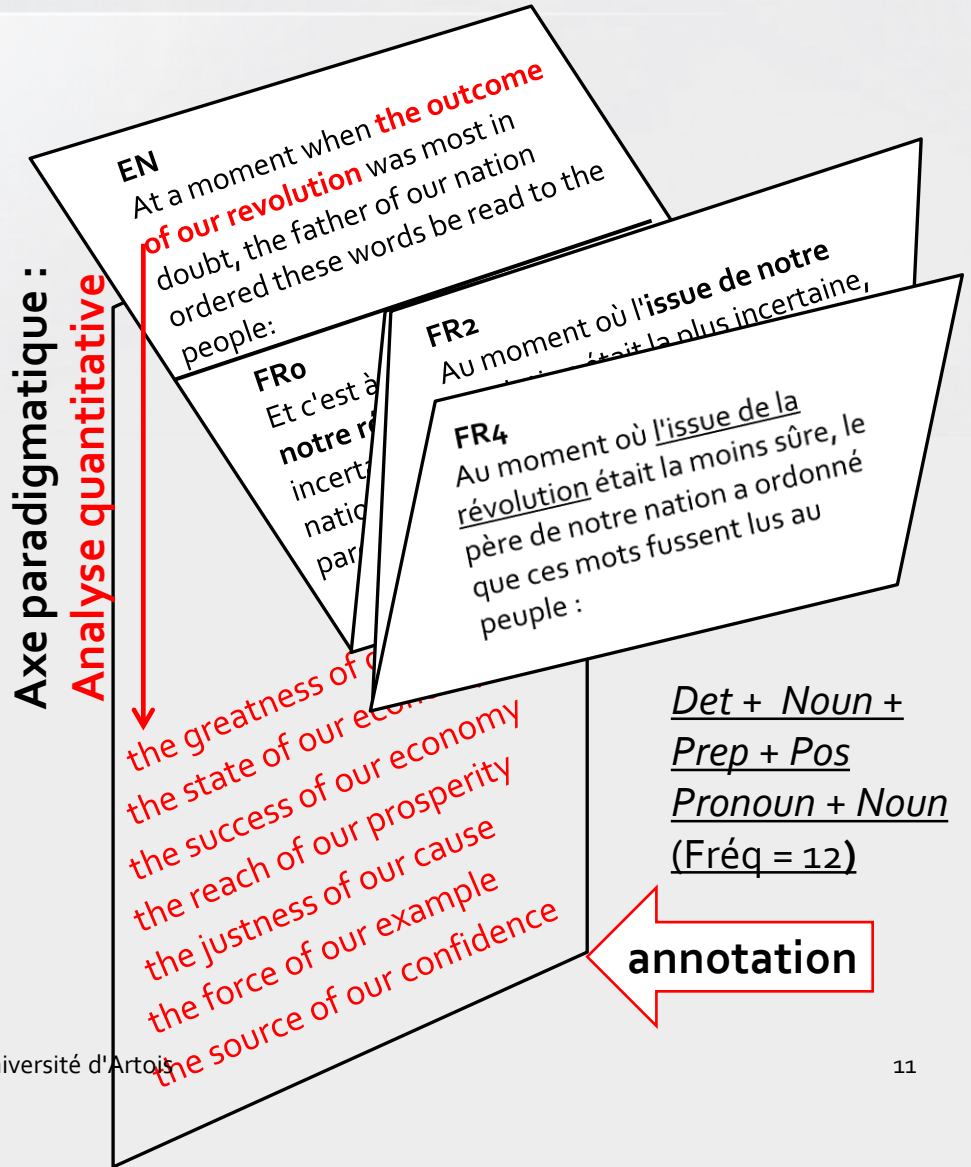


Equivalence

Limites de l'approche centrée sur le TMX

Translation Memory eXchange

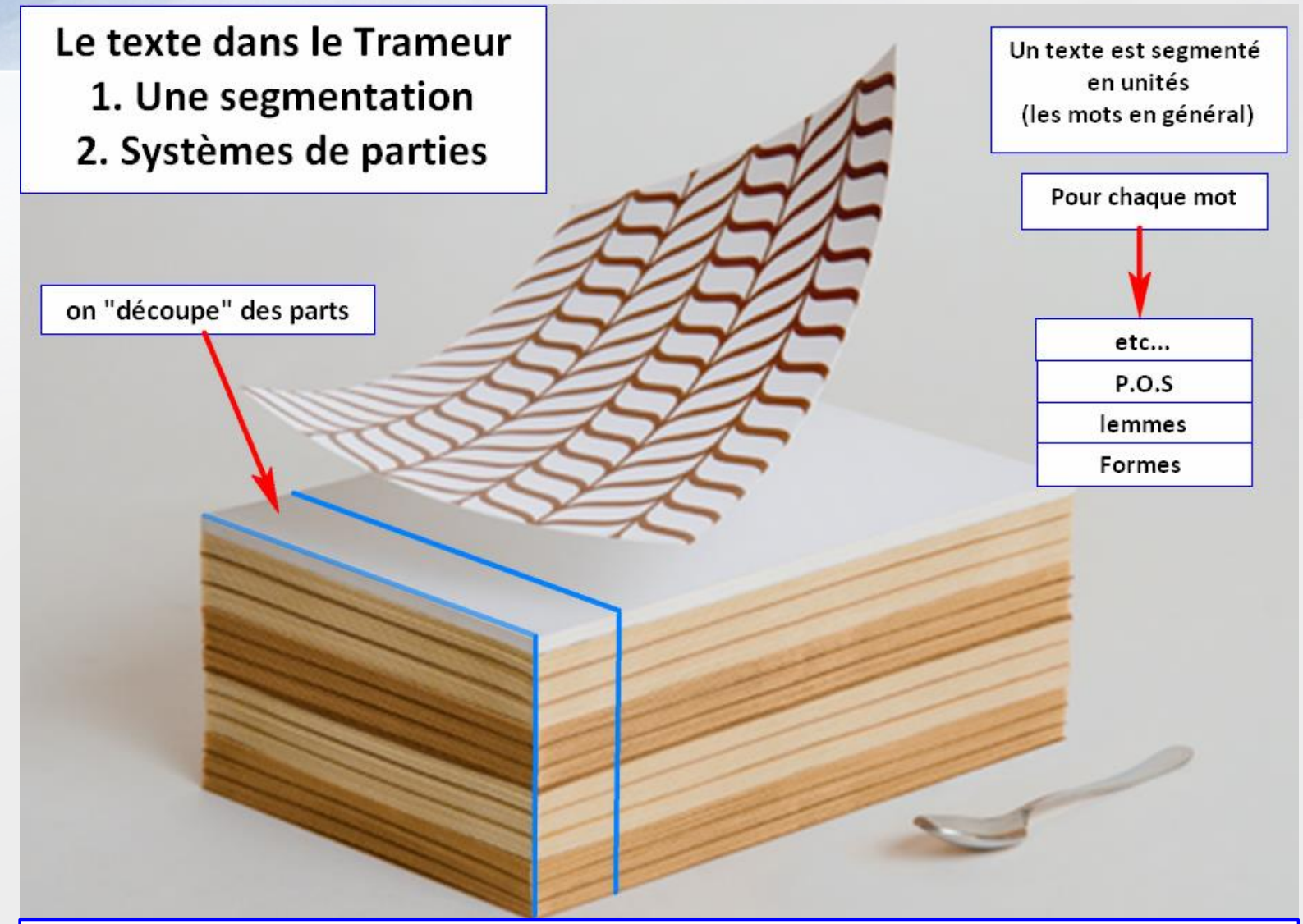
EN	FRo
At a moment when **the outcome of our revolution** was most in doubt, the father of our nation ordered these words be read to the people:	Et c'est à l'heure où **l'issue de notre révolution** était le plus incertaine que le père de notre nation ordonna que les paroles suivantes fussent lues à la population :
"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. "	" 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. "



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***



On compte des « contenus » dans des « contenants »

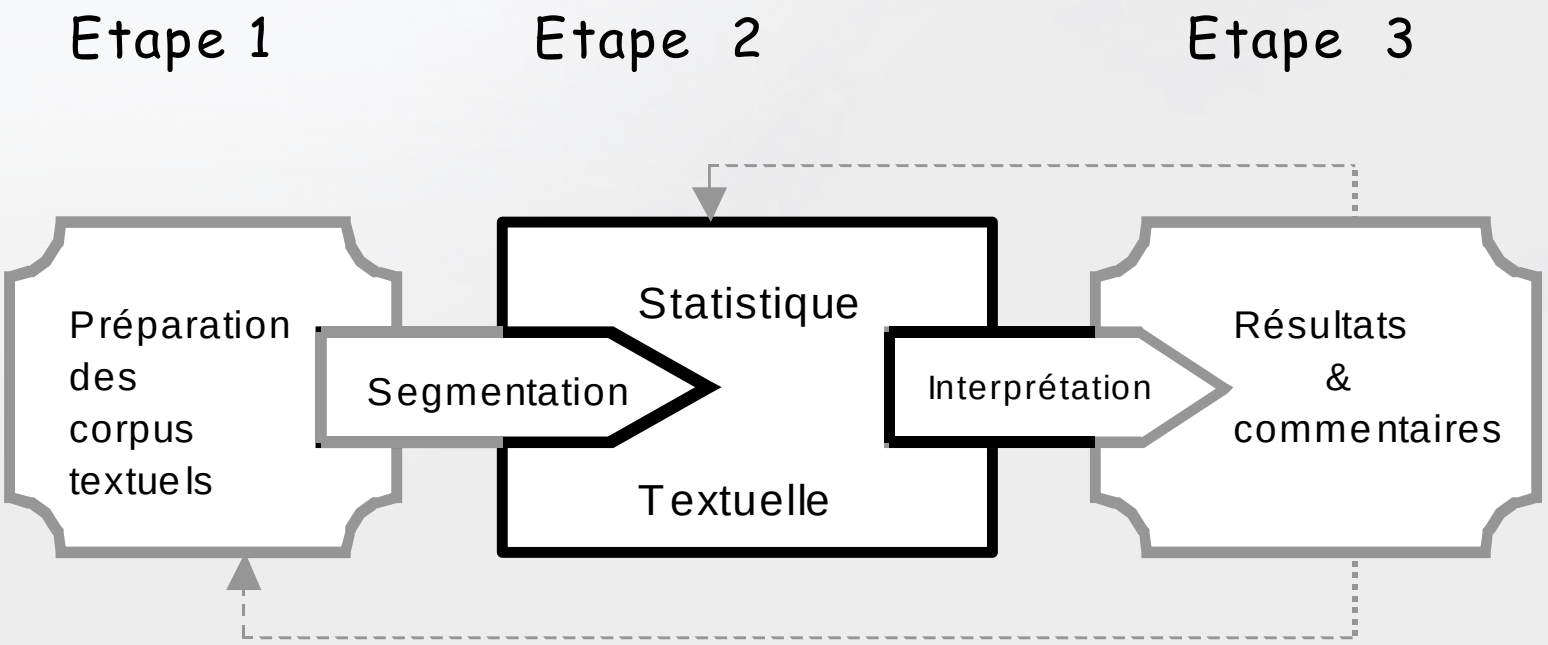
Le Trameur

```
<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>
<acces>
<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>
</acces>
</cadre>
</baselexicometrique>
```

Élément de la ***Trame***

Cadre

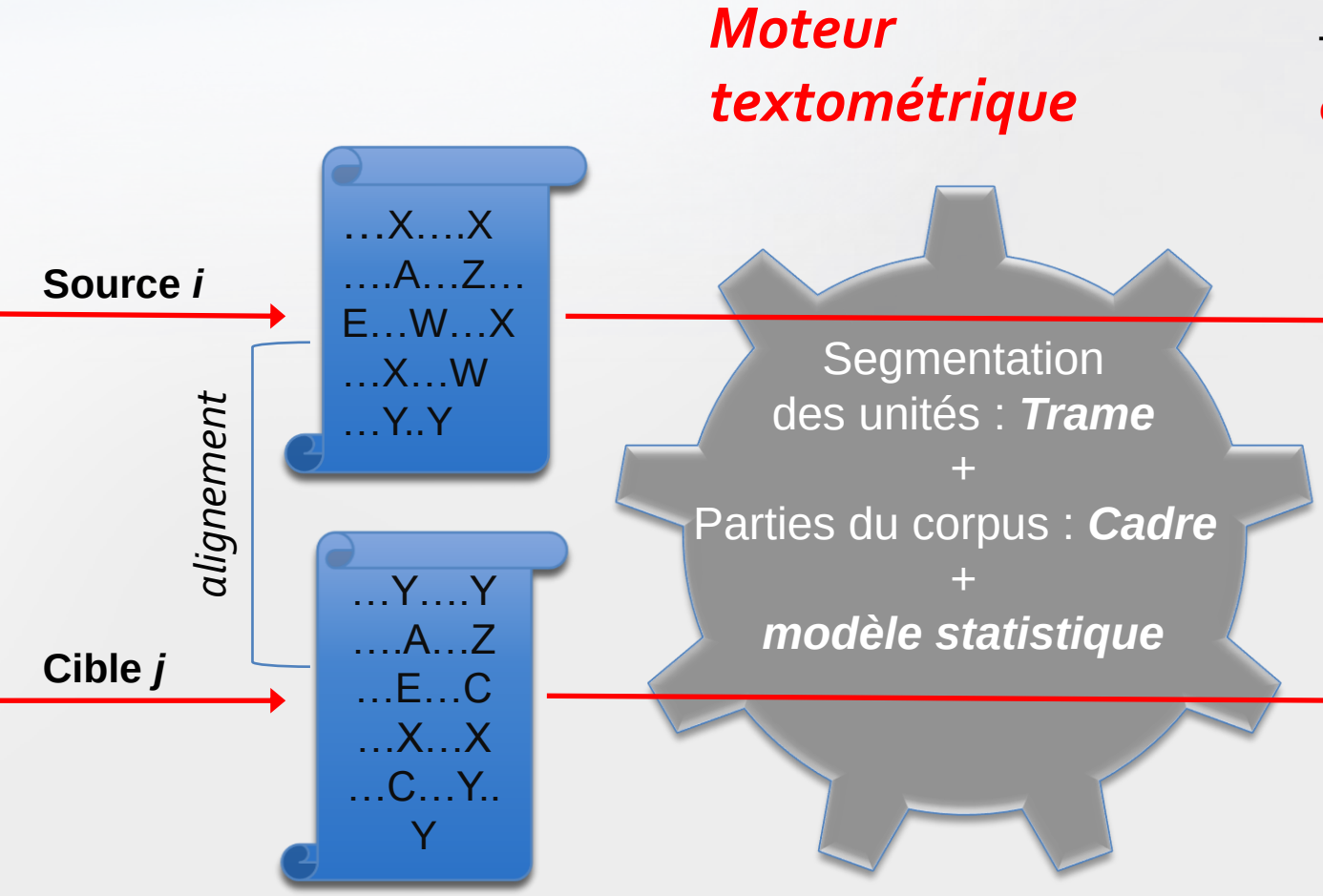
Textométrie : procédures



données modes d'observation résultats

Lamalle *et al.*, 2006

Analyse quantitative du bi-texte



Tableaux statistiques
calculés dynamiquement

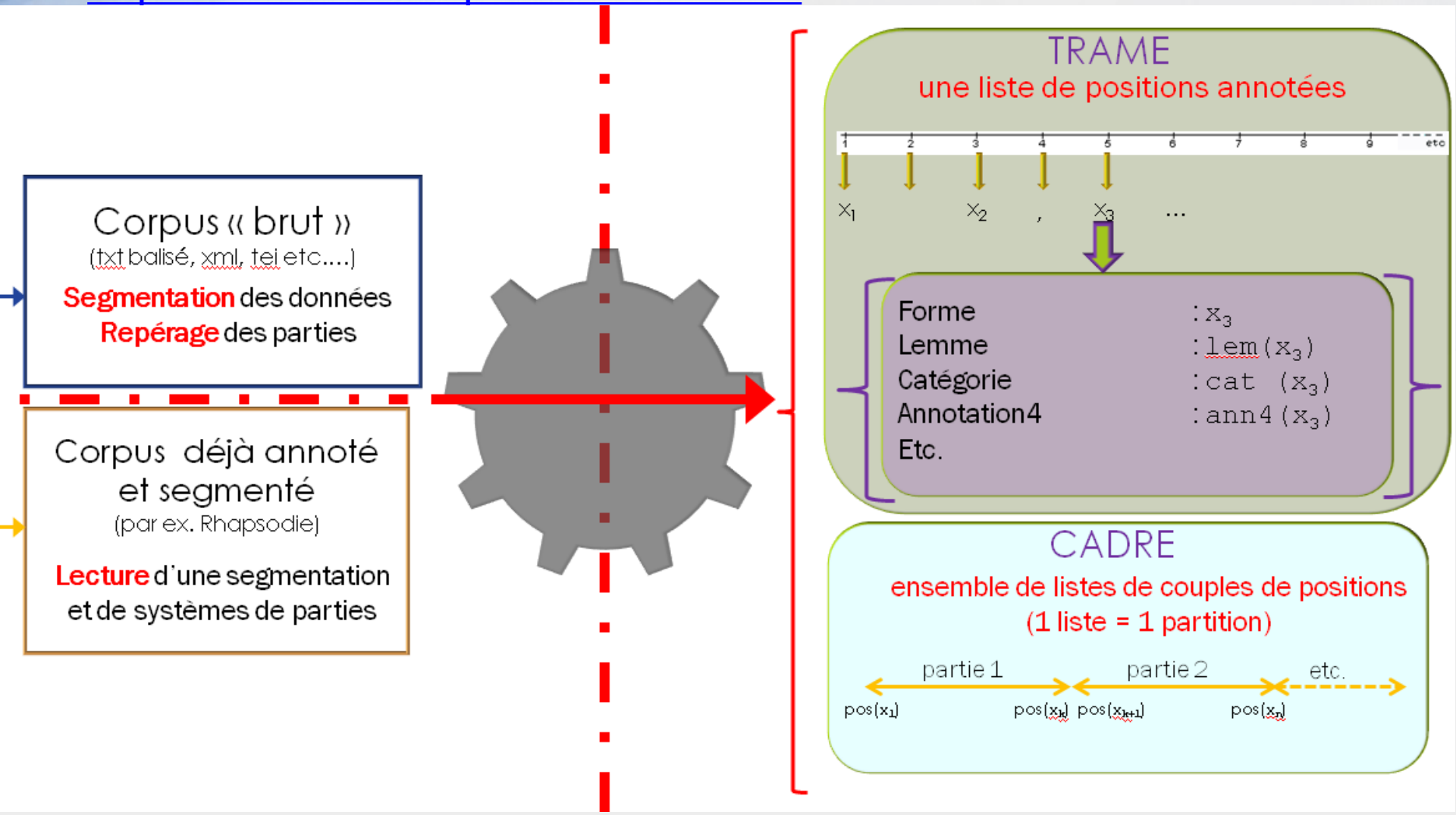
Forme	Spécif.
X	23.43
Y	12.68
Z	5.57
W	5.66

Forme	Spécif.
Y	13.73
X	21.86
A	7.75
C	6.55

Le Trameur

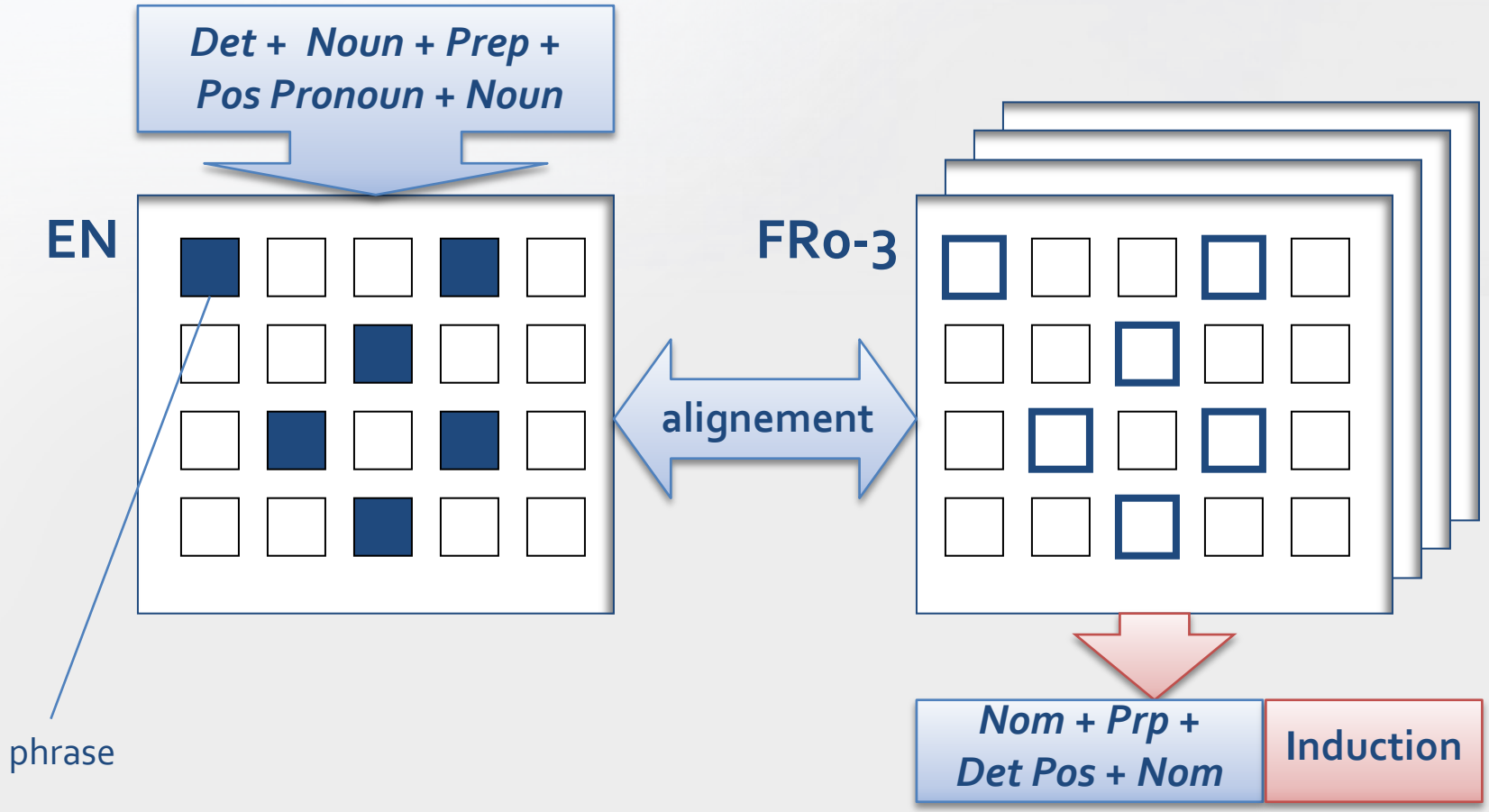
Fleury, Zimina, COLING 2014

http://www.tal.univ-paris3.fr/trameur/



Correspondances induites dynamiquement

Corpus ***Obama***

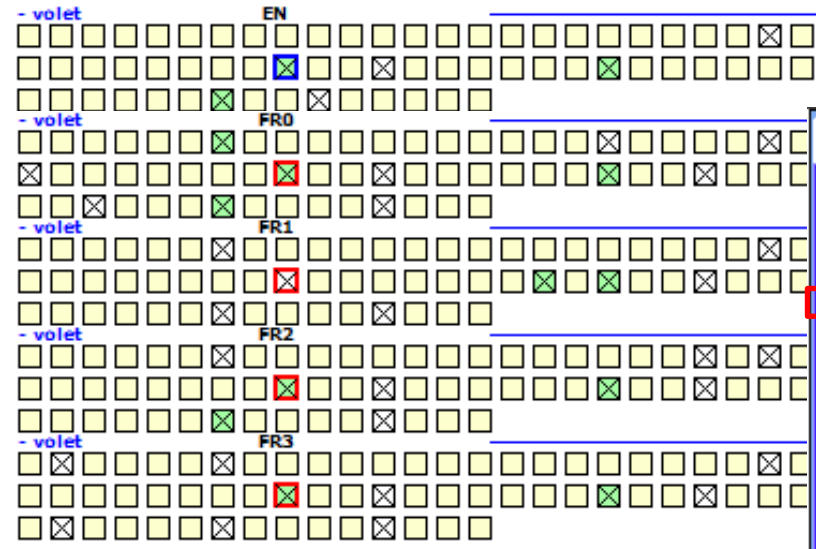


Alignements multiples

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



Spécif | Spécif-tri | Tfq | AFC | Facteurs | PCLC | R

Actions sur Résultats spécifs | co-présence

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

N° Sect. : 62:(2519,2574) Annotation : 1 Aperçu : 50 Nb Volet 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;
§

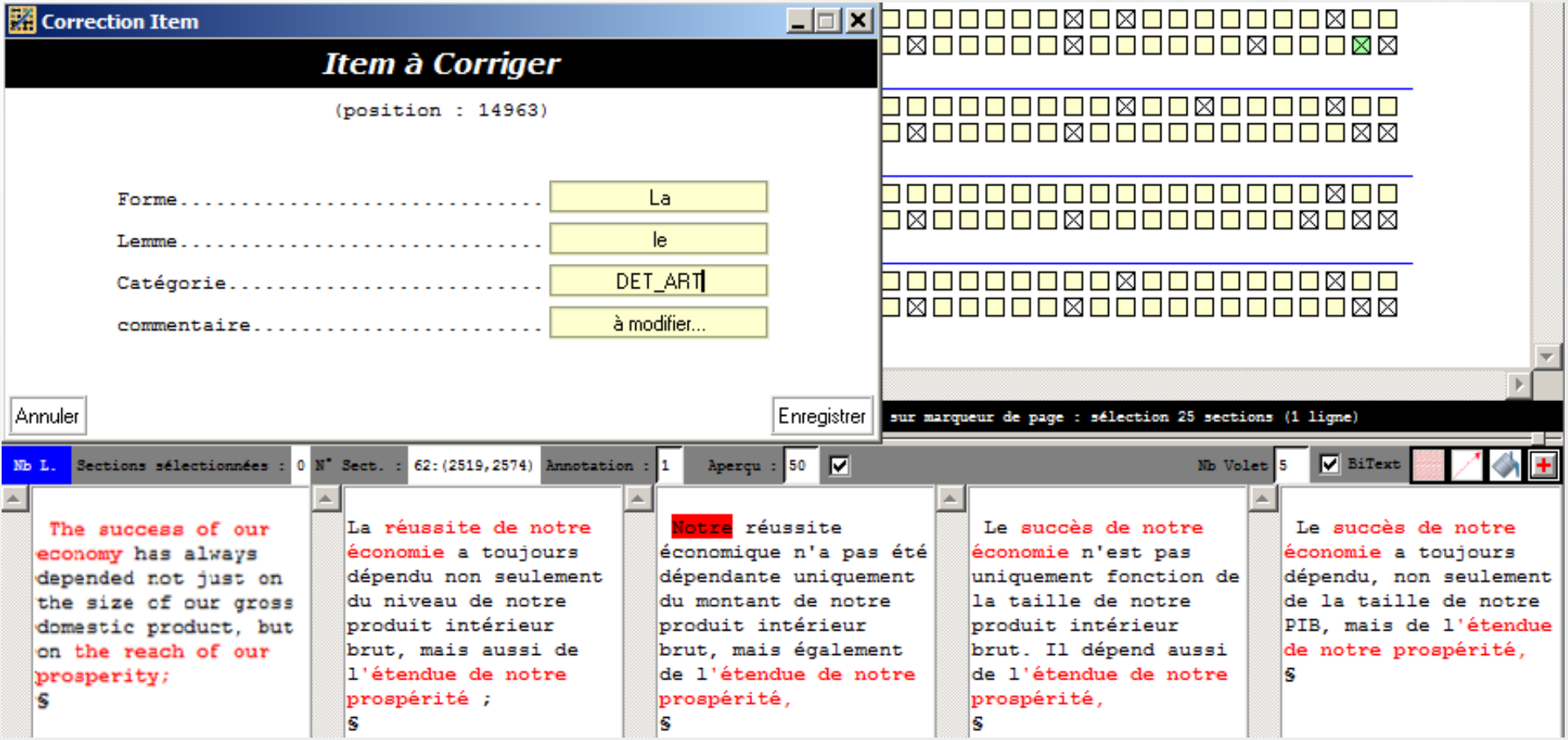
La réussite de notre économie a toujours dépendu non seulement du niveau de notre produit intérieur brut, mais aussi de l'étendue de notre prospérité ;
§

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

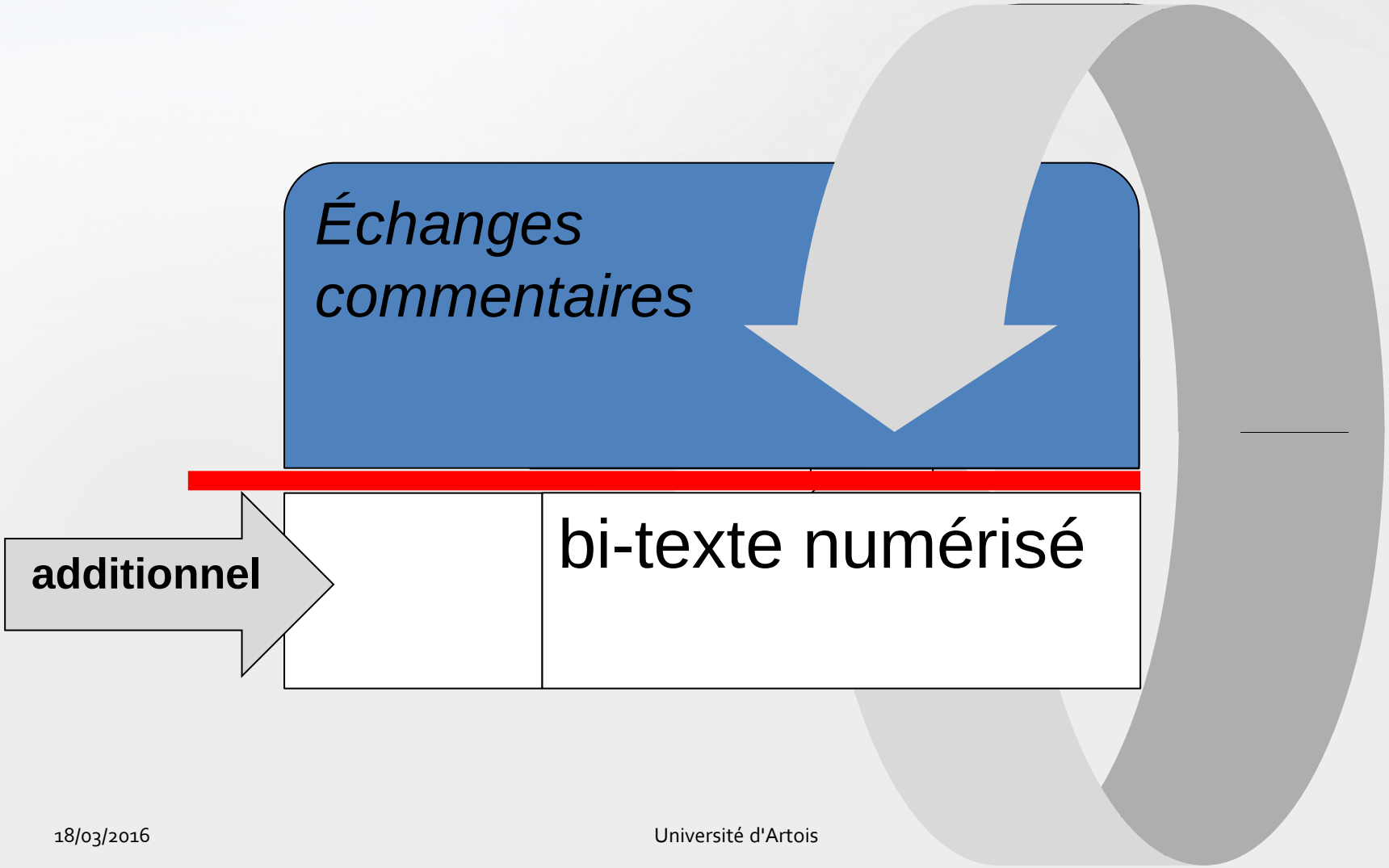
Le succès de notre économie n'est pas uniquement fonction de la taille de notre produit intérieur brut. Il dépend aussi de l'étendue de notre prospérité,
§

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

Système d'annotations dynamiques



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 différentiation
- Lecture bilingue à l'écran
- Détection des **difficultés spécifiques** des apprentis traducteurs : *démonstrations*

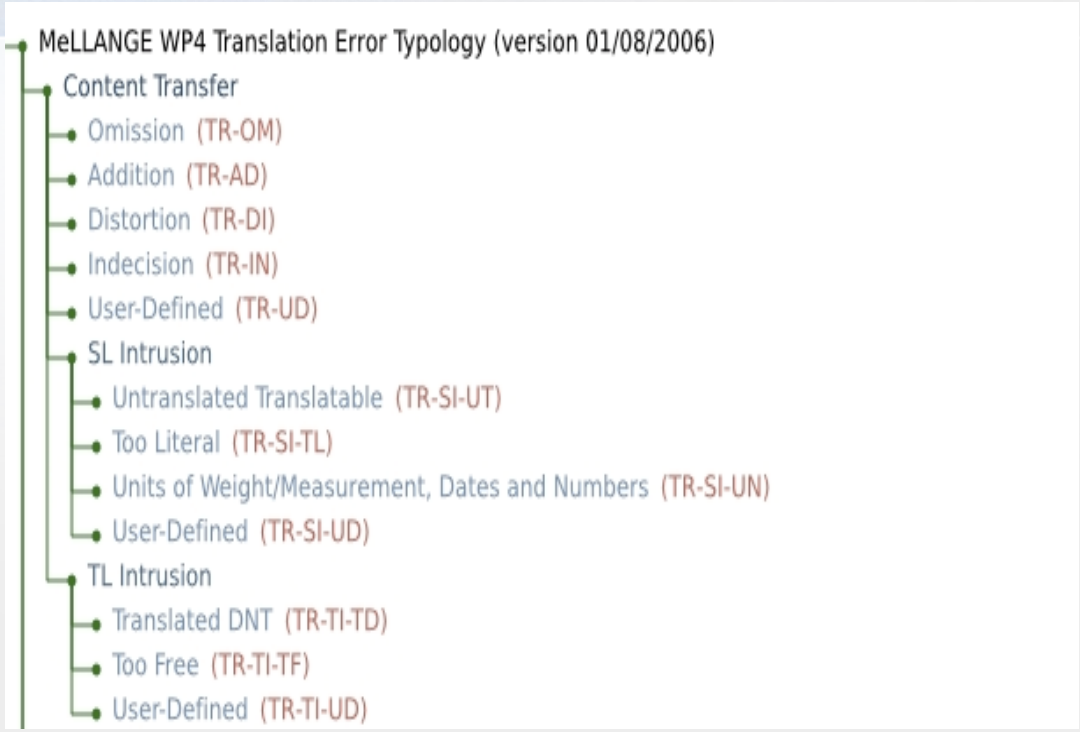
Corpus d'erreurs de traduction ER-TRAD-SP (N. Kübler, M. Pecman, A. Mestivier-Volanschi)

- 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

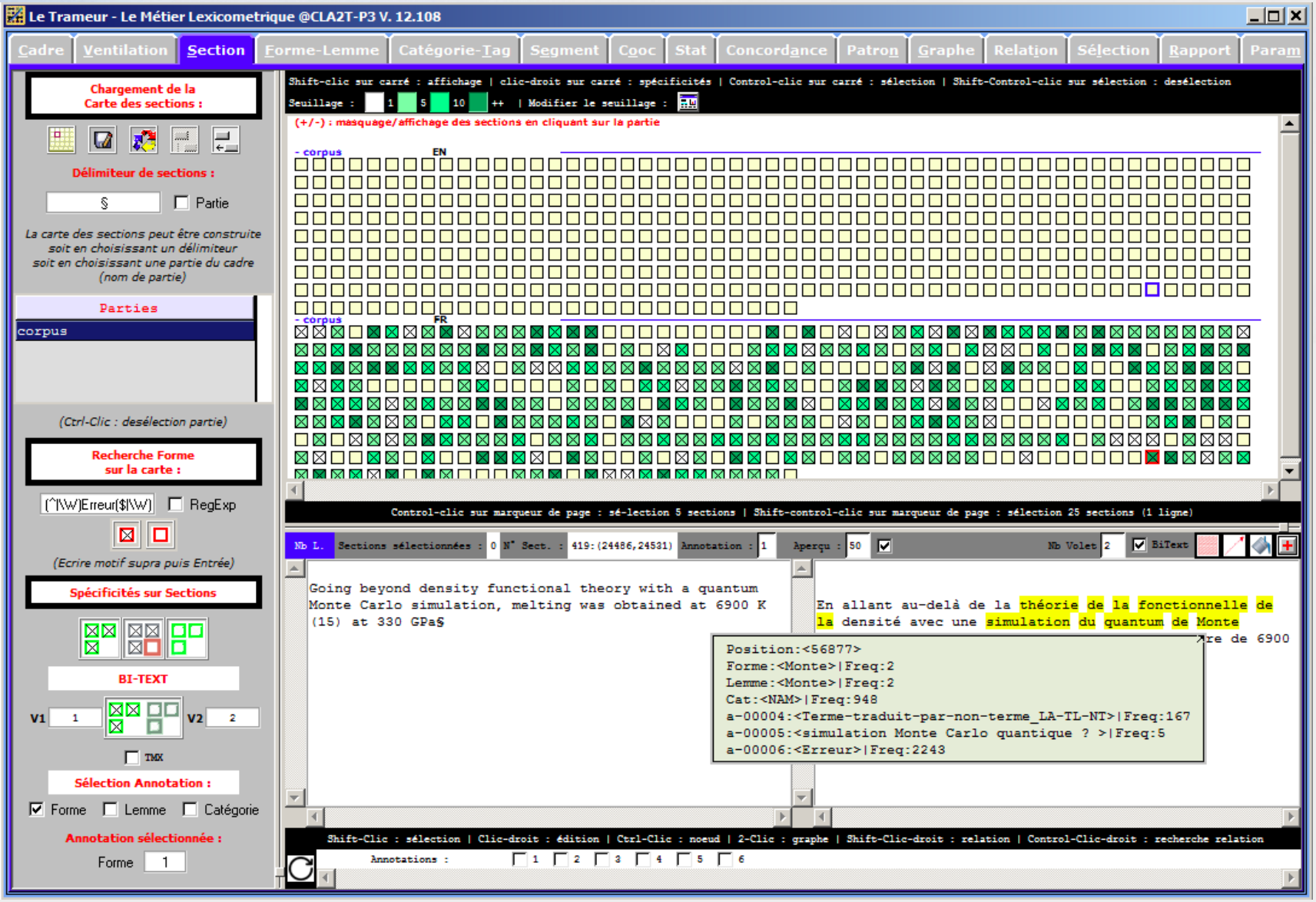
- Les traductions sont annotées selon la typologie d'erreurs développée dans le cadre du projet MeLLANGE (Castagnoli *et al*. 2011).
- Distinction entre les erreurs liées au transfert de contenu et les erreurs de langue. Ces deux catégories sont divisées en sous-catégories du type « Intrusion de la langue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

Typologie d'erreurs MeLLANGE



Typologie d'erreurs MeLLANGE

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



Lecture Textométrique Différentielle (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

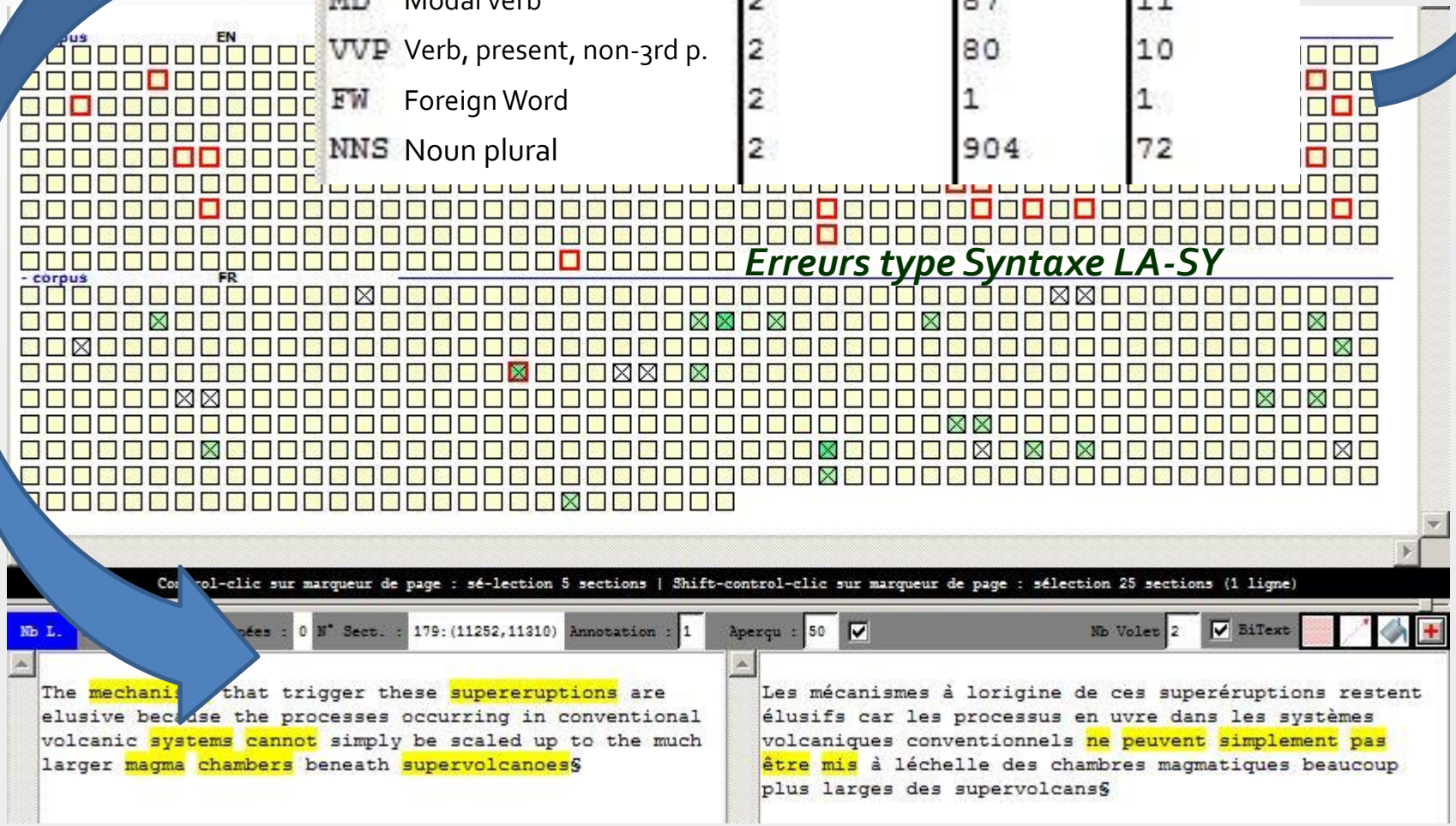
Spécificités (éléments caractéristiques)

2

1

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

Erreurs type Syntaxe LA-SY



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 ≈ 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 unitsdifference 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 ≈ 0,5ε 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 theMoon consists predominantly of mantle material with high Hf/Wand hence most probably radiogenic 182W/184W, this two-stagemodel 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,37 Myr and implying that the bulk silicate Moon and Earth musthave had indistinguishable initial 182W/184W ratios (SupplementaryFig 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 ≈ 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 fromthe identical 182W/184Wratios of the bulk silicate Moon and Earth inconjunction with their distinct Hf/W ratios and by assuming identicalinitial 182W/184Wratios 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 182Wexcess 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 µm) 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 µm) 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 comte de l'analyse des *dépendances syntaxiques* (nouvelle annotation).

Démonstrations sur les ***treebanks :***
ParTUT

ParTUT (*Par*allel *T*urin *U*niversity *T*reeBank)
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 (*Par*allel *T*urin *U*niversity *T*reeBank)
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.

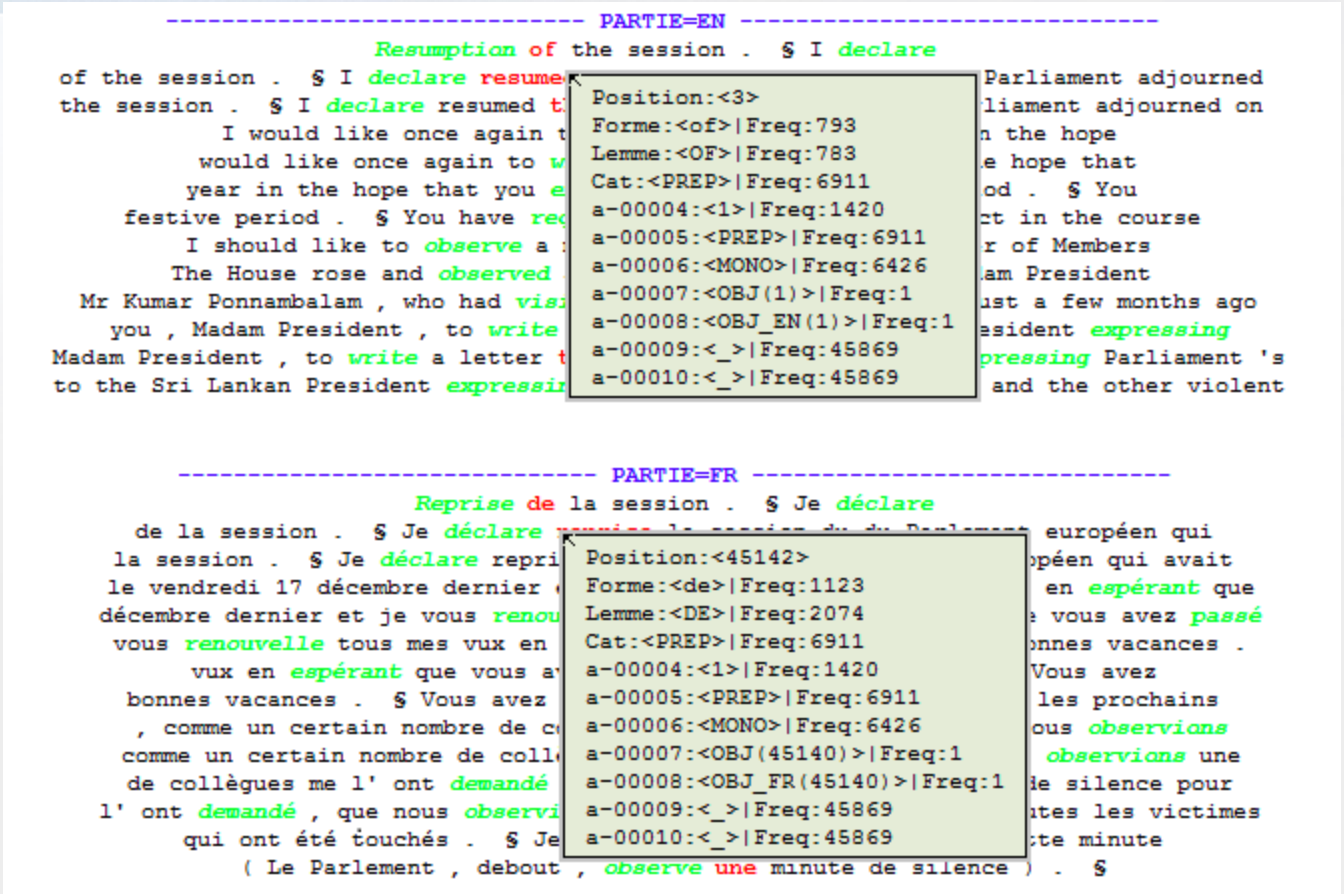
Seules les ressources en français et en anglais ont été utilisées dans notre exploration (environ 97 000 occ.).

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

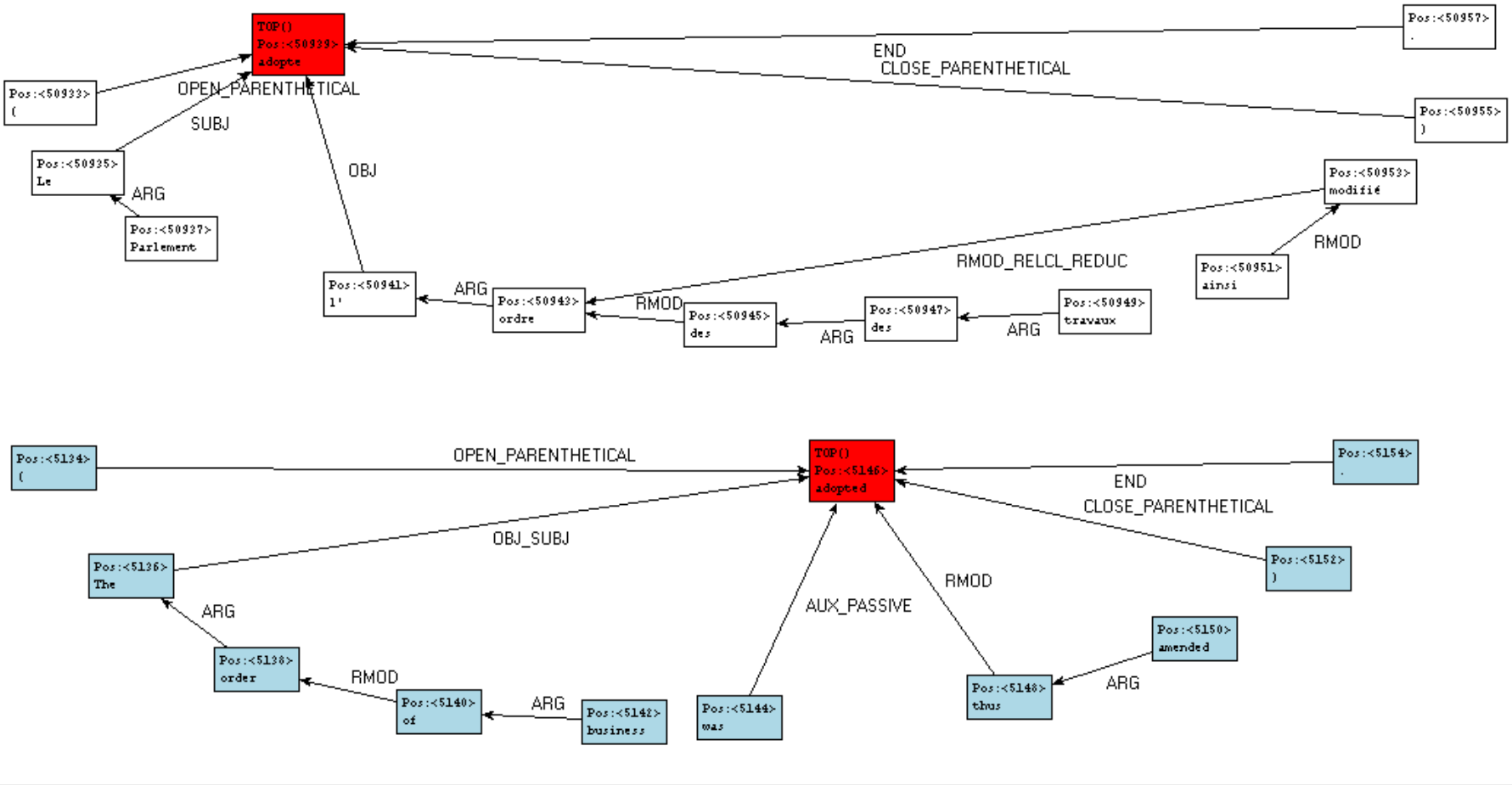
```
1   Paternité   PATERNITÉ   NOUN    NOUN    COMMON|F|SING   0   TOP _   _
2   -   #\- PUNCT   PUNCT   _   1   SEPARATOR   _   _
3   Partage PARTAGE NOUN    NOUN    COMMON|M|SING|PARTAGER|TRANS    1   APPOSITION  _   _
4   Des DE  PREP    PREP    MONO    1   RMOD    _   _
5   Des LE  ART ART DEF|ALLVAL|PL   4   ARG _   _
6   Conditions  CONDITION   NOUN    NOUN    COMMON|F|SING   5   ARG _   _
7   Initiales   INITIALE    ADJ ADJ QUALIF|ALLVAL|PL    6   CONTIN+DENOM    _   _
8   À   À   PREP    PREP    MONO    6   RMOD    _   _
9   l'  LE  ART ART DEF|ALLVAL|SING 8   ARG _   _
10  Identique   IDENTIQUE   ADJ ADJ QUALIF|ALLVAL|SING  9   ARG _   _
11  2.0 >2.0>   NUM NUM _   1   APPOSITION  _   _
12  .   #\. PUNCT   PUNCT   _   1   END _   _

1   Creative    CREATIVE    NOUN    NOUN    PROPER  4   SUBJ    _   _
2   Commons COMMONS NOUN    NOUN    PROPER  1   CONTIN+DENOM    _   _
3   n'  NE  ADV ADV NEG 4   RMOD    _   _
4   est ÊTRE    VERB    VERB    MAIN|IND|PRES|INTRANS|3|SING    0   TOP _   _
5   pas PAS ADV ADV NEG 4   RMOD    _   _
6   un  UN  ART ART INDEF|M|SING    4   PREDCOMPL+SUBJ  _   _
7   cabinet CABINET NOUN    NOUN    COMMON|M|SING   6   ARG _   _
8   d'  DE  PREP    PREP    MONO    7   RMOD    _   _
9   avocats AVOCAT  NOUN    NOUN    COMMON|M|PL 8   ARG _   _
10  et  ET  CONJ    CONJ    COORD|COORD 4   COORD+BASE  _   _
11  ne  NE  ADV ADV NEG 12  RMOD    _   _
12  fournit FOURNIR VERB    VERB    MAIN|IND|REMPAST|TRANS|3|SING   10  COORD2ND+BASE   _   _
13  pas PAS ADV ADV NEG 12  RMOD    _   _
14  de  DE  ART ART INDEF|ALLVAL|ALLVAL 12  OBJ _   _
15  services    SERVICE NOUN    NOUN    COMMON|M|PL 14  ARG _   _
16  de  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)



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.