

# Conceptual Annotations Preserve Structure Across Translations: A French-English Case Study

**Elior Sulem**<sup>1</sup>, Omri Abend<sup>2</sup>, and Ari Rappoport<sup>1</sup>

<sup>1</sup>Institute of Computer Science, The Hebrew University of Jerusalem

<sup>2</sup>School of Informatics, University of Edinburgh

ACL 2015 Workshop on Semantics-Driven Statistical Machine Translation (S2MT)

July 30<sup>th</sup> 2015, Beijing



# Integration of Structural Information in Machine Translation

- Usually required before the development of statistical models for translation.
- Syntax-based models for Statistical Machine Translation (Chiang, 2005; Liu et al., 2006; Mi et al., 2008)
- **Advantages:**
  - Take into account the hierarchical structure of the languages.
  - Better reordering at the global level.
- **Challenge:**
  - Cross-linguistic divergences** (Dorr, 1994)

# Syntactic Cross-Linguistic Divergences

Example:

John took a shower

John se duchó

- Translations (same meaning)

BUT

- Different syntactic structure

# Syntactic Cross-Linguistic Divergences

Example :

John **kicked** the ball

John **donna un coup de pied** dans le ballon

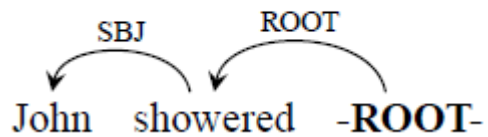
*(John gave a kick in the ball)*

- Translations (same meaning)

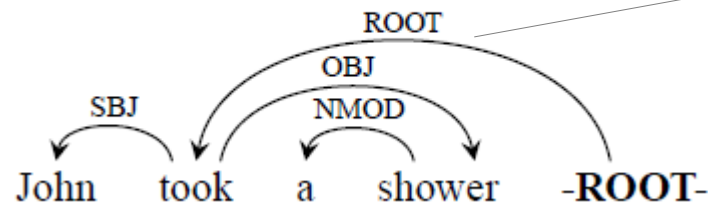
BUT

- Different syntactic structure

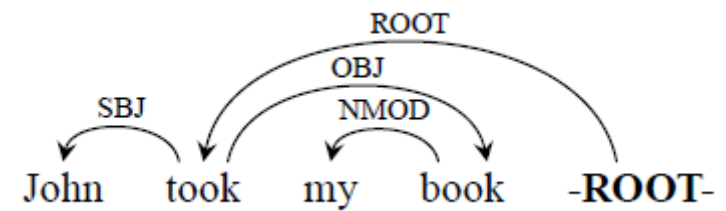
# Syntactic Cross-Linguistic Divergences



John showered  
John se duchó



John took a shower



John took my book

John tomó mi libro

Different meaning

Same structure

Same meaning

Different structure

Adapted from Abend and Rappoport (2013)  
ACL presentation

# Semantic Annotation in Machine Translation

- Semantics - Promising candidate for providing structures that are stable across languages.
- A main goal of translation is to conserve the meaning of a sentence in a source language when translating it to the target language.

## Recent work:

- Using SRL (Wu and Fung, 2009 ; Liu and Gildea, 2010; Liu and Gildea, 2013)
- Using Intermediary Representation (Jones et al., 2012).

## However:

- The **stability of semantic annotation across translations** is seldom addressed and has yet to be adequately supported .



We present here a detailed corpus analysis.

# Outline

- UCCA (Universal Conceptual Cognitive Annotation)
- Portability
- Construction of a French-English Parallel UCCA Corpus
- Stability: Quantitative Study
- Translation Divergences
- Divergence Analysis
- Conclusion
- Future Work

# Universal Cognitive Conceptual Annotation (UCCA)

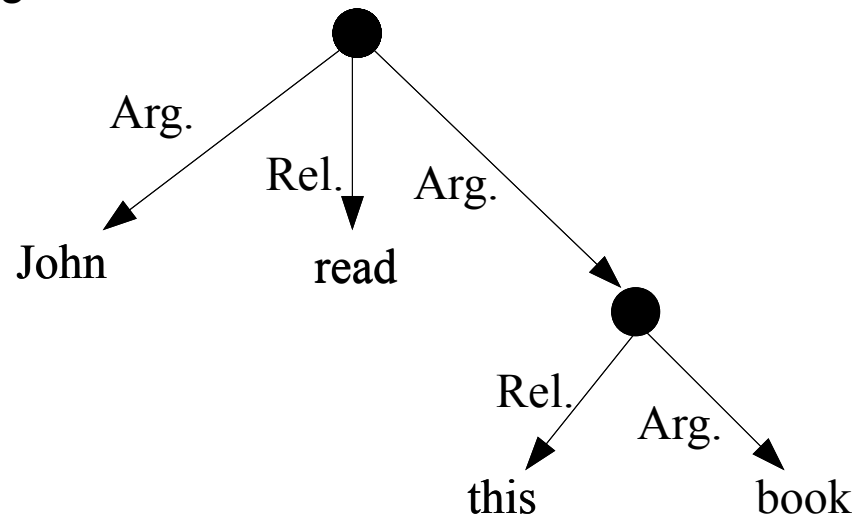
(Abend and Rappoport ACL 2013)

- UCCA is an annotation scheme for encoding **semantic information**.
- Primarily based on **Basic Linguistic Theory** (BLT) (Dixon 2005, 2010,2012) and on **cognitive theories** (Langacker 2008)
- Abstracts away from specific syntactic forms.
- Represents **semantic distinctions** explicitly.



# UCCA's Formalism

- The foundational layer focuses on predicate-argument relations and linkage between them.
  - Terminals
  - Units
  - Relations and arguments



John read this book.

# UCCA's Formalism

- The foundational layer focuses on predicate-argument relations and linkage between them.

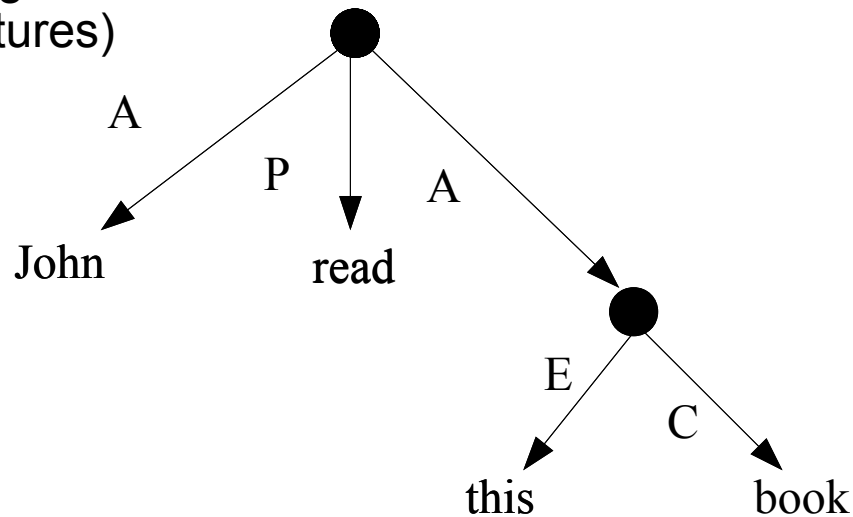
- Terminals
- Units
- Relations and arguments
- Categories (Features)

Process (P)

Participant (A)

Center (C)

Elaborator (E)



John read this book.

# UCCA's Scenes

- The most basic notion foundational layer/ layer is the **Scene**, describing some movement, action or state.

Examples:

[John]<sub>A</sub> [read]<sub>P</sub> [this book]<sub>A</sub> [yesterday]<sub>D</sub>

[John]<sub>A</sub> [is tall]<sub>S</sub>

Participant (A)

Process (P)

State (S)

Adverbial (D)

# Syntactic Cross-Linguistic Divergences

[John]<sub>A</sub> [showered]<sub>P</sub>  
[John]<sub>A</sub> [se duchó]<sub>P</sub>

[John]<sub>A</sub> [took a shower]<sub>P</sub>

[John]<sub>A</sub> [took]<sub>P</sub> [my book]<sub>A</sub>

[John]<sub>A</sub> [tomó]<sub>P</sub> [mi libro]<sub>A</sub>

Same meaning

Same structure

Same meaning

Same structure

Adapted from Abend and Rappoport (2013)  
ACL presentation

# Structural Annotation in MT: Requirements

- Two kinds of universality:

1. **Portability**

Same category set, same guidelines

2. **Stability**

Similar annotations for translations

- We show these two properties for UCCA, focusing on English and French.

# Portability

## Procedure:

- **Tool:** "French Grammar and Usage" (Hawkins and Towell, 2001)
- Check that UCCA categories can be applied to major grammatical phenomena in French .

## Findings:

- Even for French-specific phenomena: current UCCA categories permit their annotation in the foundational layer
  - without requiring changes in the definitions
  - without additional categories.

# Portability

## Example: Reflexive pronouns in French

- (a) The reflexive pronoun refers to the same Participant as the subject.

**Annotation: Participant (A).**

[Jean]<sub>A</sub> [s']<sub>A</sub> [est<sub>F</sub> acheté<sub>C</sub>]<sub>P</sub> [une<sub>E</sub> voiture<sub>C</sub>]<sub>P</sub>. (Jean bought a car for himself)

- (b) The pronoun changes in an unpredictable way the original of the verb or alternatively, the verb appears only at a pronominal form. No semantic reflexivity.

**Annotation: Forms unanalyzable unit with the verb.**

[II]<sub>A</sub> [[s']<sub>C-</sub> [est]<sub>F</sub> [aperçu]<sub>C(CONT.)</sub>]<sub>P</sub> [qu'<sub>F</sub> il<sub>F</sub> [était<sub>F</sub> tard]<sub>C</sub>]<sub>S</sub><sub>A</sub> (He realized that it was late)

- Full analysis in Sulem (2014)

[www.cs.huji.ac.il/~eliors](http://www.cs.huji.ac.il/~eliors)

# Stability: Type-Level Analysis

## Dorr's Translation Divergences (Dorr 1994, 2002, Dorr et al.2004)

- **Categorical divergence:** The translation of words in one language into words that have different parts of speech in another language.  
**Examples:** [to<sub>F</sub> be<sub>F</sub> cold<sub>C</sub>]<sub>S</sub> – [avoir<sub>F</sub> froid<sub>C</sub>]<sub>S</sub> (*to have cold*)
- **Conflational divergence:** The translation of two or more words in one language into one word in another language.  
**Example:** [to<sub>F</sub> kick<sub>C</sub>]<sub>P</sub> – [[donner]<sub>F</sub> [un<sub>E</sub> [coup de pied]<sub>C</sub>]<sub>A</sub>]<sub>P</sub> (*to give a kick*)
- **Structural divergence:** The realization of verb arguments in different syntactic configurations in different languages.  
**Example:** [to<sub>F</sub> enter<sub>C</sub>]<sub>P</sub> [the<sub>E</sub> house<sub>C</sub>]<sub>A</sub> – [entrer]<sub>P</sub> [dans<sub>R</sub> la maison<sub>C</sub>]<sub>A</sub> (*to enter in the house*)
- **Thematic divergence:** The realization of verb arguments in syntactic configurations that reflect different thematic to syntactic mapping orders.  
**Example:** [I]<sub>A</sub> [like]<sub>P</sub> [this<sub>E</sub> house<sub>C</sub>]<sub>A</sub> – [cette maison]<sub>A</sub> [me]<sub>A</sub> [plaît]<sub>P</sub> (*this house pleases to me*)

⇒ UCCA permits **structure conservation**.



## Dorr's Translation Divergences (Dorr 1994, 2002, Dorr et al.2004) [Continuation]

- **Demotional divergence:** "Promotion" of a modifier in the source language to a main verb in the target language.

**Example:** [to<sub>F</sub> run<sub>C</sub>]<sub>P</sub> [in<sub>R</sub>]<sub>A,IMPLICIT-C</sub> - [entrer<sub>C</sub> [en<sub>F</sub> courant<sub>C</sub>]]<sub>E</sub><sub>P</sub>

- **Promotional divergence:** "Demotion" of a main verb in the source language to a modifier in the target language

**Example:** [John]<sub>A</sub> [usually]<sub>D</sub> [goes]<sub>P</sub> [home]<sub>A</sub> - [John]<sub>A</sub> [a l'habitude<sub>E</sub> de rentrer<sub>C</sub>]<sub>P</sub> [à<sub>R</sub> la maison<sub>C</sub>]<sub>A</sub>

These are two subcases of **head-swapping divergences:** The inversion of a structural dominance relation between two semantically equivalent words when translating from one language to another.

⇒ Updated guidelines permit addressing promotional divergence:

[John]<sub>A</sub> [usually]<sub>D</sub> [goes]<sub>P</sub> [home]<sub>A</sub> - [John]<sub>A</sub> [**a l'habitude de**]<sub>D</sub> [rentre]<sub>P</sub> [à<sub>R</sub> la maison<sub>C</sub>]<sub>A</sub>

In both cases – **Elaboration of the main relation.**

# Parallel French-English UCCA Corpus

Vingt-Mille Lieues Sous les Mers (Twenty Thousand Leagues Under the Sea)  
Jules Verne (1870)  
English translation by J.P. Walter

[http://fr.wikisource.org/wiki/Vingt\\_mille\\_lieues\\_sous\\_les\\_mers](http://fr.wikisource.org/wiki/Vingt_mille_lieues_sous_les_mers)

<http://jv.gilead.org.il/fpwalter>

- First five chapters of the book  
583 sentences in each of the languages  
12.5 K tokens in English, 13.1K tokens in French
- 154 **parallel passages**
- **Manual annotation** in English and in French using UCCA web application  
[vm-05.cs.huji.ac.il](http://vm-05.cs.huji.ac.il)



Publically available: [www.cs.huji.ac.il/~elior](http://www.cs.huji.ac.il/~elior)

# Stability: Quantitative Analysis

## Experimental Setup:

We compare the numbers of UCCA's Scenes, Participants and Adverbials in a parallel passage to numbers obtained by syntactic annotation.

### Tools:

- Stanford POS tagger package (Toutanova et al. 2003)
- English Stanford parser (Klein and Manning 2003)
- French Stanford parser (Green et al. 2011)

### Comparisons:

- **Scenes / non-auxiliary verbs** (number of clauses)
- **Participants and Adverbials / NPs, PPs, ADVPs**

# Stability: Quantitative Analysis

## Similarity computation:

- For each unit/constituent type we compute **the number of instances** of that type in each passage.
- We compare the obtained vector to its corresponding vector in the other language.
- We compute **distance between the vectors** using  $l_1$  and  $l_2$  metrics, and similarity using an F-score, with the precision and recall of the French vector against the English one.

# Stability: Quantitative Analysis

## Results:

	$l_1$	$l_2$	$F$	Fr. Avg.	En. Avg.
Scenes	<b>124</b>	<b>14.97</b>	<b>0.96</b>	9.25	9.49
Verbs	157	18.79	0.94	9.30	9.10
Participants (As)	<b>273</b>	<b>31.13</b>	<b>0.95</b>	17.68	18.27
NPs and PPs	952	102.74	0.89	26.64	32.33
NPs	847	88.89	0.87	18.78	24.20
PPs	299	32.05	0.87	7.86	8.13
Adverbials (Ds)	<b>133</b>	<b>17.18</b>	<b>0.86</b>	3.3	3.07
Adverb Phrases	342	40.0	0.15	0.24	2.49
As + Ds	<b>334</b>	<b>37.18</b>	<b>0.95</b>	20.99	21.34
NPs+PPs+ADVPs	1226	127.40	0.87	26.88	34.82

- In all cases **the UCCA annotation is more stable** than its syntactic constituent counterparts.
- The distances for NPs and for the union of syntactic annotations are at least **three times bigger** than their UCCA counterparts.
- Manual annotation of a sample of the corpus: the stability observed is **not a result of the parser's biases**.

# UCCA Divergences

- Given a parallel corpus, a unit in one language corresponds to a unit in the other language if they have the **same category** and if the units have the **same meaning**, disregarding modifiers.
- Given a UCCA category, some of the units of that category are left unaligned between the two sides of the parallel corpus, creating a **UCCA divergence**.
- We classify UCCA divergences according to the category and the language of the **unaligned units** .

# UCCA Divergences

## Formally:

- A **sufficient subset** of a unit  $u$  is a subset of  $u$  which contains its heads (the main relation in the case of a Scene, the Centers in the case of a non-Scene).  
Example: “He **ran**” is a sufficient subset of “He slowly **ran**”.
- A unit  $e$  in English and a unit  $f$  in French correspond to each other if they have the same category and any of the 3 following conditions hold:
  - (1)  $e$  is a translation of  $f$
  - (2) a sufficient subset of  $e$  is a translation of  $f$
  - (3) a sufficient subset of  $f$  is a translation of  $e$

## Example:

The English Scene “He slowly ran” corresponds to the French Scene “Il a couru” (He ran) since condition (2) holds.

# UCCA Divergences

## Examples:

eng. [of]<sub>R</sub> [the]<sub>E</sub> [ship]<sub>C</sub> [[victimized]<sub>P</sub> [by this new ramming]<sub>A</sub>]<sub>E,REMOTE-A("ship")</sub>

fr. [du]<sub>R+E</sub> [navire]<sub>C</sub> [[victime]<sub>C</sub> [de ce nouvel abordage]<sub>E</sub>]<sub>E</sub>

“ship victimized by this new ramming” is **an unaligned English Scene**, creating a Scene English divergence.

eng. [He]<sub>A</sub> [slowly]<sub>D</sub> [ran]<sub>P</sub>.

fr. [Il]<sub>A</sub> [a couru]<sub>P</sub>.

- No Scene divergence.
- “slowly” is **an unaligned English Adverbial**, creating an English Adverbial divergence.



# Divergence Analysis

## Number of UCCA Divergences :

### Scene divergences:

- **92.3%** of the English Scenes have a French correspondent.
- **94.9%** of the French Scenes have an English correspondent.
- Only 25% of the sentences (148 out of 583) contains any Scene divergences.

# Divergence Analysis

## Number of UCCA Divergences :

### Participant and Adverbial divergences:

- Studied only on parallel passages without Scene divergences.
- **94.0%** of the English Participants have a correspondent in French.  
**95.3%** of the French Participants have a correspondent in English.
- **80.6%** of the English Adverbials have a correspondent in French.  
**79.4%** of the French Adverbials have correspondent in English.

# Divergence Analysis

## Properties of UCCA Divergences:

	Property	Scene Div.		Participant Div.		Adverbial Div.	
		Eng.	Fre.	Eng.	Fre.	Eng.	Fre.
<b>Translation Study</b>							
1	Similar Translation Possible	65.18	58.33	50	35.29	70.83	50.0
2	Similar Source Possible	73.21	63.89	54.55	47.06	75.0	46.15
-	None	18.75	31.94	38.64	47.06	16.67	42.31
<b>Annotation Study</b>							
3	Conforming Analysis	41.96	54.16	72.72	73.53	25.0	53.85
4	Different Interpretation	10.71	1.39	25	23.53	8.33	7.69
-	None	55.36	44.44	25	20.59	70.83	46.15
<b>Semantic Effect of the Unaligned Unit</b>							
5	Additional Information	38.39	18.06	25.0	20.59	37.50	0.0
6	Tense Information	8.04	5.56	-	-	-	-
7	Emphasis	19.64	8.33	31.82	26.47	41.67	3.85
-	None	50.89	80.56	61.36	64.71	58.33	96.15

We analyze the divergences according to 3 groups of properties.

- Translation Study

- Many of the divergences can be ascribed to the specific translation selected. For example, more than 65% of the English Scene divergences can be avoided through a **different translation**.

# Divergence Analysis

## Properties of UCCA Divergences:

Property	Scene Div.		Participant Div.		Adverbial Div.		
	Eng.	Fre.	Eng.	Fre.	Eng.	Fre.	
<b>Translation Study</b>							
1	Similar Translation Possible	65.18	58.33	50	35.29	70.83	50.0
2	Similar Source Possible	73.21	63.89	54.55	47.06	75.0	46.15
-	None	18.75	31.94	38.64	47.06	16.67	42.31
<b>Annotation Study</b>							
3	Conforming Analysis	41.96	54.16	72.72	73.53	25.0	53.85
4	Different Interpretation	10.71	1.39	25	23.53	8.33	7.69
-	None	55.36	44.44	25	20.59	70.83	46.15
<b>Semantic Effect of the Unaligned Unit</b>							
5	Additional Information	38.39	18.06	25.0	20.59	37.50	0.0
6	Tense Information	8.04	5.56	-	-	-	-
7	Emphasis	19.64	8.33	31.82	26.47	41.67	3.85
-	None	50.89	80.56	61.36	64.71	58.33	96.15

We analyze the divergences according to 3 groups of properties.

- Translation Study
- Annotation Study

- Most of the Scene and Adverbial divergences could have been avoided had a **different annotation** been selected.

# Divergence Analysis

## Properties of UCCA Divergences:

Property		Scene Div.		Participant Div.		Adverbial Div.	
		Eng.	Fre.	Eng.	Fre.	Eng.	Fre.
<b>Translation Study</b>							
1	Similar Translation Possible	65.18	58.33	50	35.29	70.83	50.0
2	Similar Source Possible	73.21	63.89	54.55	47.06	75.0	46.15
-	None	18.75	31.94	38.64	47.06	16.67	42.31
<b>Annotation Study</b>							
3	Conforming Analysis	41.96	54.16	72.72	73.53	25.0	53.85
4	Different Interpretation	10.71	1.39	25	23.53	8.33	7.69
-	None	55.36	44.44	25	20.59	70.83	46.15
<b>Semantic Effect of the Unaligned Unit</b>							
5	Additional Information	38.39	18.06	25.0	20.59	37.50	0.0
6	Tense Information	8.04	5.56	-	-	-	-
7	Emphasis	19.64	8.33	31.82	26.47	41.67	3.85
-	None	50.89	80.56	61.36	64.71	58.33	96.15

We analyze the divergences according to 3 groups of properties.

- Translation Study
- Annotation Study
- Semantic effect of the unaligned unit

- Many divergences can be explained a **true semantic difference** between the source and the translation.

# Divergence Analysis

## Properties of UCCA Divergences: Category Replacement

Replaced by	Scene Div.		Participant Div.		Adverbial Div.	
	Eng.	Fre.	Eng.	Fre.	Eng.	Fre.
Linker	6.25	1.39	–	–	8.33	7.69
Ground	1.79	1.39	–	–	4.17	3.85
Elaborator of Participant	–	–	0	2.94	4.17	19.23
Main Relation	–	–	20.45*	20.59*	25.0*	26.92*
Parallel Scene	–	–	13.64	2.94	–	–
Participant	–	–	–	–	4.17	11.54
Adverbial	–	–	6.82	2.94	–	–
2 Participants	–	–	11.36	2.94	–	–
2 Adverbials	–	–	–	–	4.17	0.0
None	91.96	98.21	47.73	67.65	50.0	30.77

\* In these cases an Adverbial/ a Participant in one of the languages is included in the meaning of the main relation in the other language.

- In some cases UCCA divergences can be formulated in terms of **category replacement**.
- These cases mainly concern Participant and Adverbial divergences.
- New UCCA guidelines, annotating secondary verbs (“try”, “begin”) as D can **reduce many of the Adverbial divergences**.

# Divergence Analysis

## Other semantic annotations: Preliminary Study:

- We annotate 10 sentence pairs with AMR (Abstract Meaning Representation) (Banarescu et al., 2013) from our corpus.
- Our analysis shows that AMR conserve the main structures in most sentences (7 out of 10).
- This suggests that other semantic annotations may also be structurally stable.
- However, semantic roles, used in PropBank (Palmer et al., 2005) and AMR, are often a source of divergences across languages.

# Conclusion (1)

- We showed that basic semantic structures can be stably preserved across French-English translations.
  - Applying UCCA on an inventory of structural divergences
  - Corpus Analysis: Quantitative comparison with syntax

Divergence analysis

Corpus publically available: [www.cs.huji.ac.il/~eliors](http://www.cs.huji.ac.il/~eliors)



## Conclusion (2)

- Semantic structures may be more suitable to SMT systems than syntactic ones, which exhibit well-known divergence phenomena.
- We show stability for UCCA and expect these advantages to generalize to other structured semantic schemes.

# Future Work

- Integration of UCCA into structure-based statistical machine translation.
  - Adding UCCA as **features** to phrase-based and syntax-based models
  - Replacement of existing syntactic **structures** by UCCA annotation
- Related tasks that would benefit from UCCA's stability:
  - Bilingual alignment
  - MT evaluation

# Thank you

Elior Sulem

[eliors@cs.huji.ac.il](mailto:eliors@cs.huji.ac.il)

- **English Scene Divergence:**

eng: [of]<sub>R</sub> [the]<sub>E</sub> [ship]<sub>C</sub> [[victimized]<sub>P</sub> [by this new ramming]<sub>A</sub>]<sub>E,REMOTE-A("ship")</sub>

fr: [du]<sub>R+E</sub> [navire]<sub>C</sub> [[victime]<sub>C</sub> [de ce nouvel abordage]<sub>E</sub>]<sub>E</sub>

“ship victimized by this new ramming” is **an unaligned English Scene**.

- **French Scene Divergence:**

eng: [Officers]<sub>A</sub> [were probing]<sub>P</sub> [the increasing gloom]<sub>A</sub> [with their night glasses]<sub>A</sub>.

fr: [[Les]<sub>E</sub> [officiers]<sub>C</sub> , [[armés]<sub>S</sub> [de leur lorgnette de nuit]<sub>A</sub> ]<sub>E,REMOTE-A("officiers")</sub> ]<sub>A</sub> , [fouillaient]<sub>P</sub> [l'obscurité croissante]<sub>A</sub>.

“officiers armés de leur lorgnette de nuit” is **an unaligned French Scene**.

- **English Participant Divergences:**

eng: [[No]<sub>D</sub> [middle of the road]<sub>A</sub> [for these two]<sub>A</sub> ]<sub>H,IMPLICIT-S</sub>

fr: [[Pas]<sub>D</sub> [de milieu]<sub>A</sub> ]<sub>H,IMPLICIT-S</sub> .

“for these two” is **an unaligned English Participant**.

- **French Participant Divergence:**

eng: [this]<sub>A</sub> [is]<sub>S</sub> [an excerpt from the well-padded article I published in the Issue  
of April 30]<sub>A</sub> .

fr: [je]<sub>A</sub> [donne]<sub>P</sub> [ici]<sub>A</sub> [un extrait d’un article très-nourri que je publiai dans le numéro  
du 30 avril]<sub>A</sub> .

“je” is **an unaligned French Participant**.

- **English Adverbial Divergence:**

eng: [But]<sub>L</sub> [[now]<sub>D</sub> [nothing]<sub>A</sub> [could hold]<sub>P-</sub> [me]<sub>A</sub> [back]<sub>-P(CONT.)</sub>]<sub>H</sub>·

fr: [Mais]<sub>L</sub> [[rien]<sub>A</sub> [ne]<sub>D</sub> [put]<sub>P-</sub> [me]<sub>A</sub> [retenir]<sub>-P(CONT.)</sub>]<sub>H</sub>

“now” is **an unaligned English Adverbial**.

- **French Adverbial Divergence:**

eng: [[we]<sub>A</sub> [are<sub>F</sub> forced<sub>E</sub> ]<sub>P-</sub>]<sub>H-</sub>, [every other supposition having been refuted]<sub>H</sub>,  
[[to<sub>F</sub> accept<sub>C</sub> ]<sub>P-(CONT.)</sub>] [the existence of an extremely powerful marine animal]<sub>A-H(CONT.)</sub>·

fr: [toute autre supposition étant rejetée]<sub>H</sub>, [[il]<sub>F</sub> [faut<sub>E</sub> ]<sub>P-</sub> [nécessairement]<sub>D</sub>  
[admettre<sub>C</sub> ]<sub>P-(CONT.)</sub>] [l'existence d'un animal marin d'une puissance excessive]<sub>A-H,IMPLICIT-A</sub>]

“nécessairement” is **an unaligned French Adverbial**.

## Another English Scene Divergence:

eng: in<sub>R</sub> *Cosmos*<sub>C</sub> [[published]<sub>P</sub> [by Father Moigno]<sub>A</sub>]<sub>E,REMOTE-</sub>

A("Cosmos")

fr: du<sub>R+E</sub> *Cosmos*<sub>C</sub> [de l'abbé Moigno]<sub>E</sub>

“Cosmos published by Father Moigno” is an **unaligned English Scene**.

# Divergence Analysis and Discussion

## Number of UCCA Divergences :

### Scene divergences:

- 112 English Scene divergences  
72 French Scene divergences.
- **92.3%** of the English Scenes have a French correspondent (1352 out of 1424)  
**94.9%** of the French Scenes have an English correspondent (1350 out of 1462)