

Procédés de traduction sous-phrastiques : ressources, reconnaissance, validations

Yuming Zhai

Directrice de thèse : Anne Vilnat

Co-encadrant de thèse : Gabriel Illouz

19 décembre 2019

LIMSI, CNRS, Université Paris-Sud, Université Paris-Saclay



Une bonne traduction : plusieurs possibilités

Traduction mot à mot

the temperature continues to rise

la température continue à augmenter

Une bonne traduction : plusieurs possibilités

Traduction mot à mot

the temperature continues to rise
la température continue à augmenter

Changement de catégorie grammaticale

after you've burned it
après la combustion

Une bonne traduction : plusieurs possibilités

Traduction mot à mot

the temperature continues to rise
la température continue à augmenter

Changement de catégorie grammaticale

after you've burned it
après la combustion

Glissement syntaxique

take captured CO₂, which can easily be [...]
récupérer du CO₂ capté, qu'on peut facilement [...]

Une bonne traduction : plusieurs possibilités

Glissement syntaxique et sémantique

EN

I **like** the dreams of the future better than the history of the past.

ZH

我 不 缅怀 过去 的 历史, 而 致力于 未来 的 梦想。

FR trad

Je ne chéris pas la mémoire de l'histoire du passé,
et je **me dévoue plutôt aux** rêves du futur.

Une bonne traduction : plusieurs possibilités

Glissement syntaxique et sémantique

EN	I like the dreams of the future better than the history of the past.
ZH	我 不 缅怀 过去 的 历史, 而 致力于 未来 的 梦想。 Je ne chéris pas la mémoire de l'histoire du passé, et je me dévoue plutôt aux rêves du futur.
FR trad	

Une bonne traduction : plusieurs possibilités

Glissement syntaxique et sémantique

EN	I like the dreams of the future better than the history of the past.
----	--

ZH	我 不 缅怀 过去 的 历史, 而 致力于 未来 的 梦想。 Je ne chéris pas la mémoire de l'histoire du passé, et je me dévoue plutôt aux rêves du futur.
FR trad	

Traduction figurative

EN	He gave the required information, in words as suitable as he could find.
----	---

ZH	他 字斟句酌 地 作 了 回答。
FR trad	Il a répondu en choisissant soigneusement ses mots.

Difficulté posée pour le traitement automatique

Différences entre les cultures et les langues

卤素鸡 → Google Translate → poulet **halogène**

Difficulté posée pour le traitement automatique

Différences entre les cultures et les langues

卤素鸡 → Google Translate → poulet **halogène**

卤 素 鸡 : poulet **végétarien** mariné ~ steak de soja mariné

Difficulté posée pour le traitement automatique

Différences entre les cultures et les langues

卤素鸡 → Google Translate → poulet halogène

卤素鸡 : poulet végétarien mariné ~ steak de soja mariné

Alignement automatique de mots

Though a global increase in transport prices may be on the cards, the biggest change will nonetheless be in price structure.

⇒ [westmos.eu](#)

Si une augmentation globale des prix du transport est prévisible, c'est toutefois surtout la structure des prix qui devrait changer le plus.

⇒ [westmos.eu](#)

With a new original album and a tour on the horizon, retirement is still not on the cards for Michel Delpech.

⇒ [rfimusique.com](#)

Avec un nouvel opus original et une tournée à venir, la retraite n'a pas encore sonné pour Michel Delpech.

⇒ [rfimusique.com](#)

New connections with Europe also appear to be on the cards.

⇒ [news.aivp.org](#)

De nouvelles liaisons vers l'Europe semblent aussi se dessiner.

⇒ [news.aivp.org](#)

Fortunately it seems to us that this is no longer on the cards, since it would in fact go against the principles of economic, [...]

⇒ [cpmr.org](#)

Heureusement, celui-ci ne nous semble plus aujourd'hui d'actualité, car il serait en effet contraire aux principes de cohésion économique, [...]

⇒ [cpmr.org](#)

Procédé de traduction

Sujet important dans la traductologie

Solutions particulières pour contourner les difficultés de traduction : ex. *équivalence, généralisation, idiome, métaphore*

Procédé de traduction

Sujet important dans la traductologie

Solutions particulières pour contourner les difficultés de traduction : ex. *équivalence, généralisation, idiome, métaphore*

Problématique de recherche

Reconnaître automatiquement les procédés de traduction

Bénéfices de la reconnaissance des procédés de traduction en TAL

Contexte de travail

Proposition de l'approche

Bannard et Callison-Burch (2005) : méthode par pivot

Extension en NMT par Mallison et al. (2017)

Construction de la ressource de paraphrases PPDB

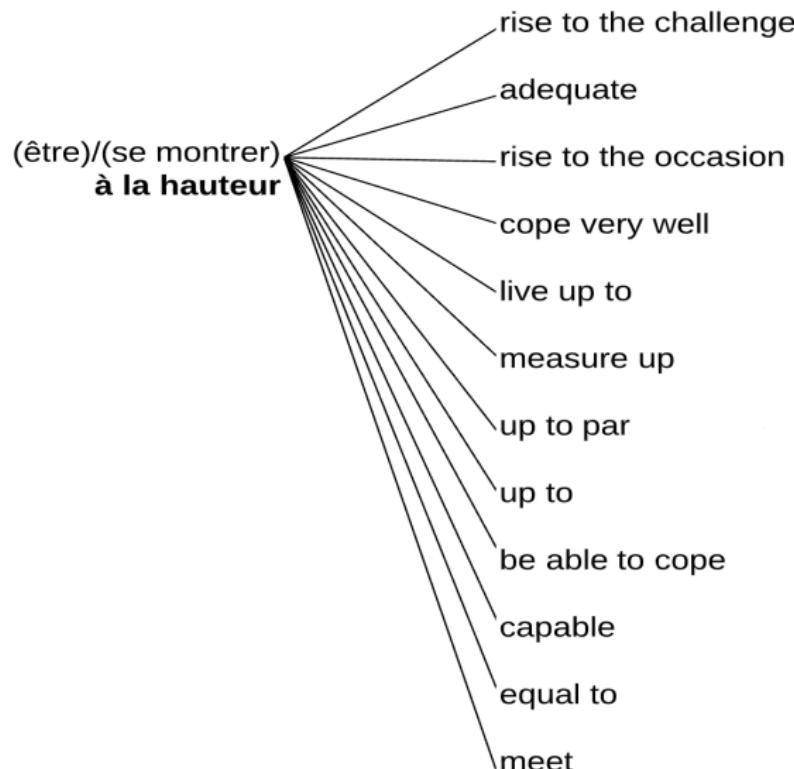
Premières approches : Callison-Burch (2008), Ganitkevitch et al. (2011, 2012)

Résultats : PPDB 1.0 et 2.0 : Ganitkevitch et al. (2013), Ganitkevitch et Callison-Burch (2014), Pavlick et al. (2015a, 2015b)

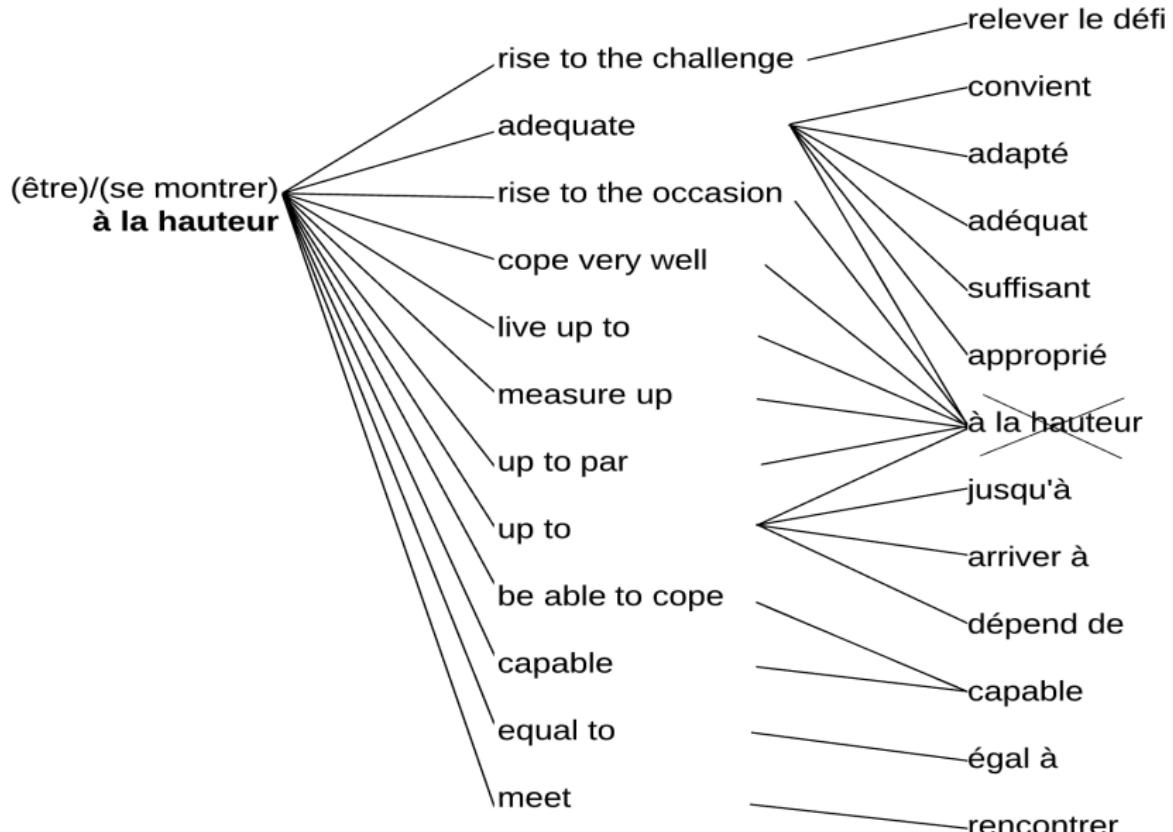
Paraphraser par équivalence de traduction

(être)/(se montrer)
à la hauteur

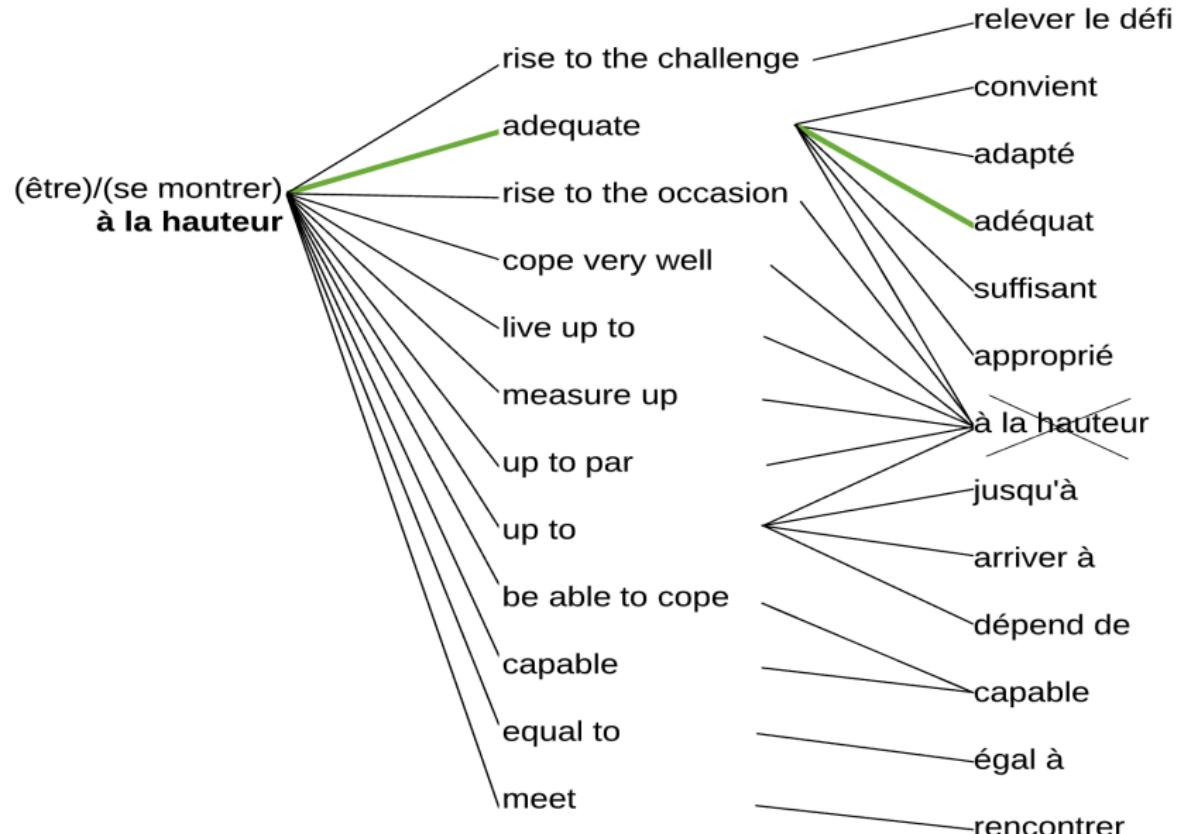
Paraphraser par équivalence de traduction



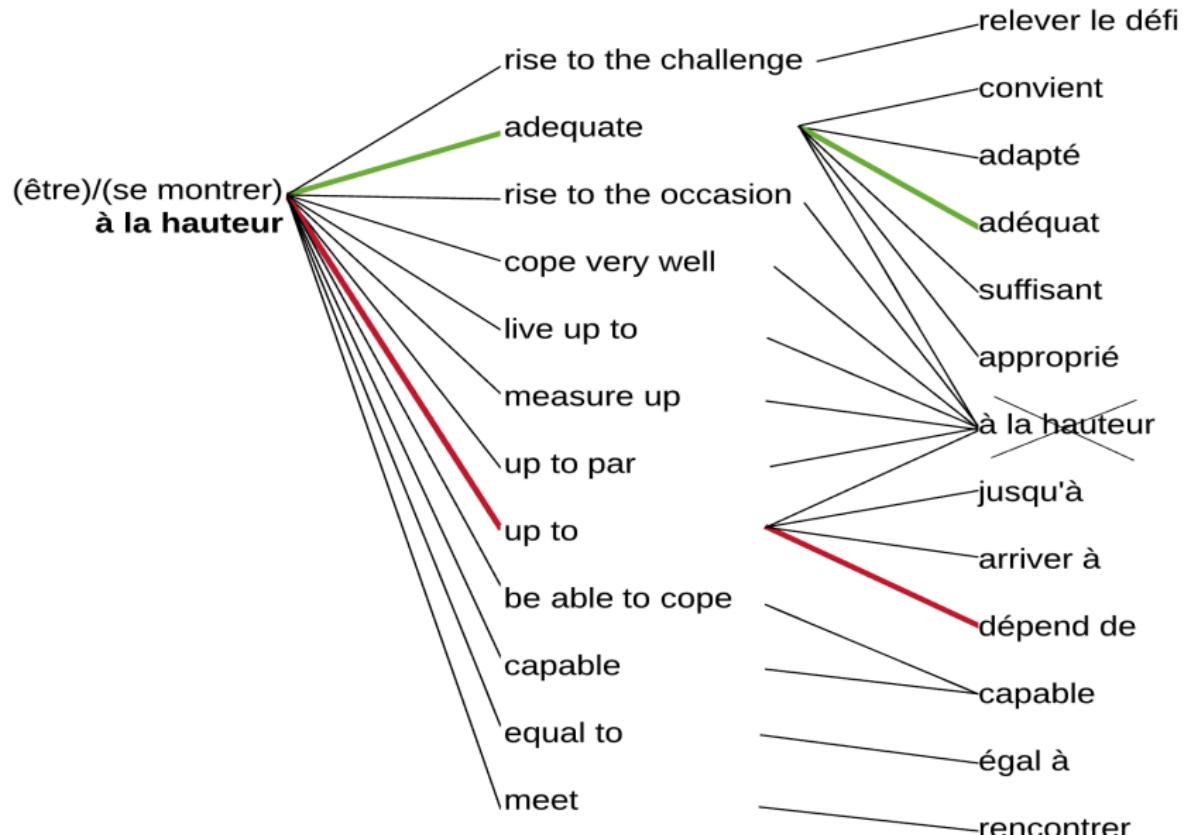
Paraphraser par équivalence de traduction



Paraphraser par équivalence de traduction

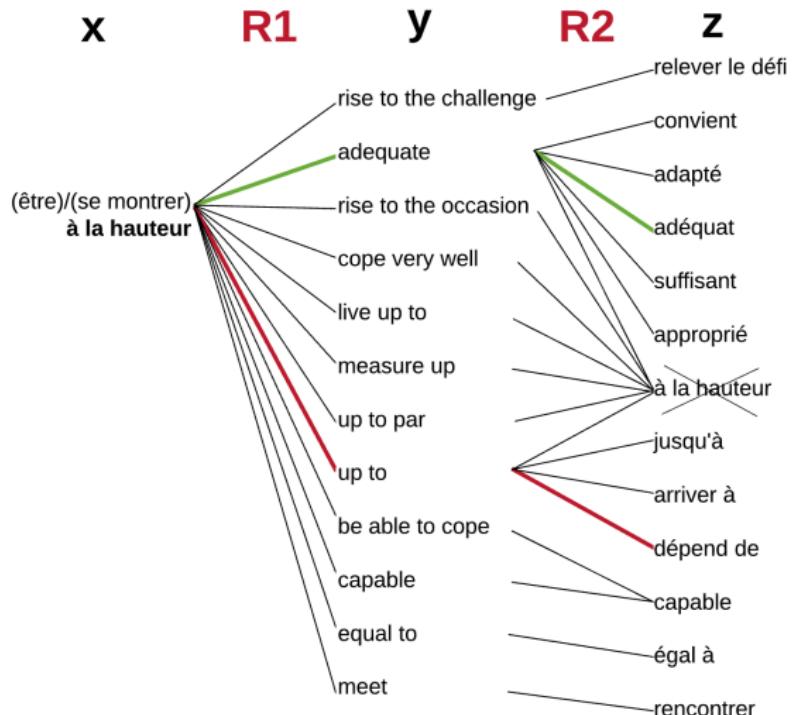


Paraphraser par équivalence de traduction



Paraphraser par équivalence de traduction

Nécessaire de qualifier R1 et R2 en procédé de traduction :

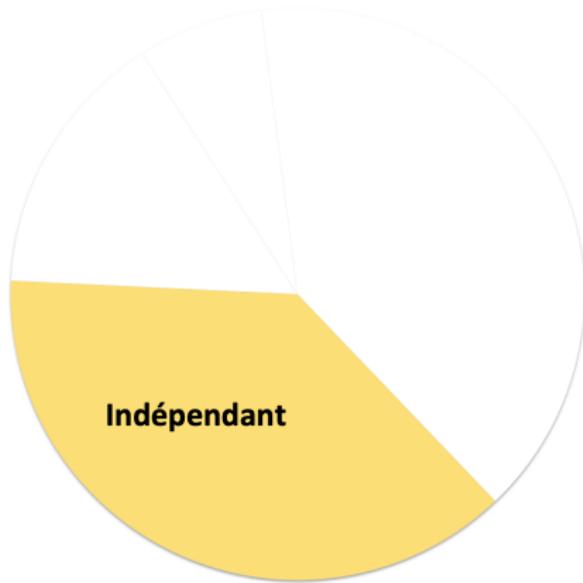


Diverses relations sémantiques dans PPDB 2.0

Manque de contrôle sémantique

(Pavlick et al., 2015)

Indépendant : *found # party*



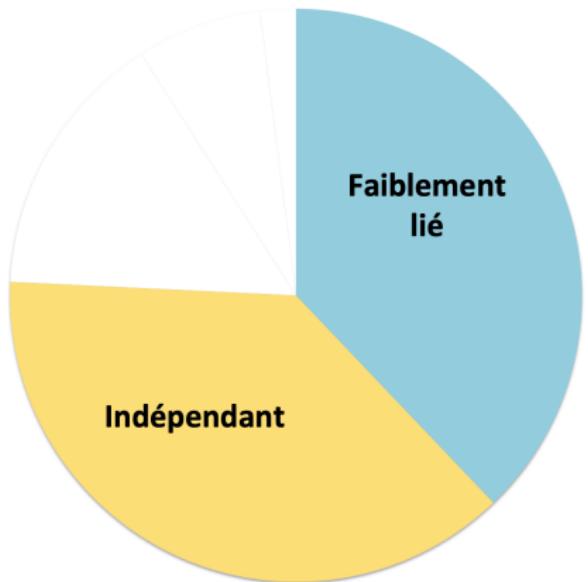
Diverses relations sémantiques dans PPDB 2.0

Manque de contrôle sémantique

(Pavlick et al., 2015)

Indépendant : *found # party*

Faiblement lié : *husband ~ marry to*



Diverses relations sémantiques dans PPDB 2.0

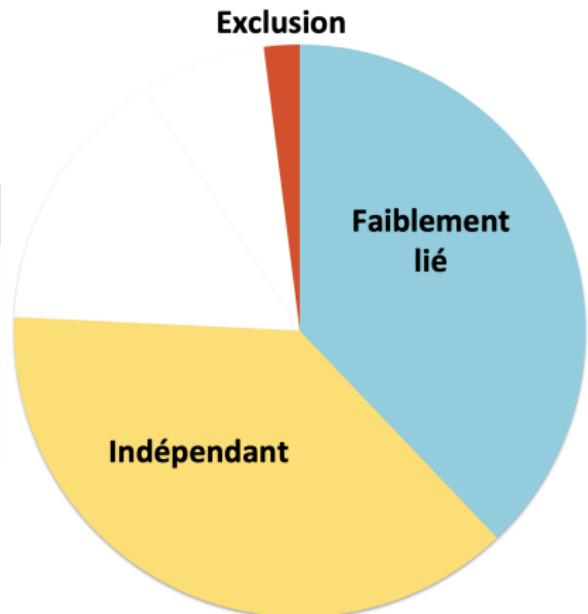
Manque de contrôle sémantique

(Pavlick et al., 2015)

Indépendant : *found # party*

Faiblement lié : *husband ~ marry to*

Exclusion : *close | open*



Diverses relations sémantiques dans PPDB 2.0

Manque de contrôle sémantique

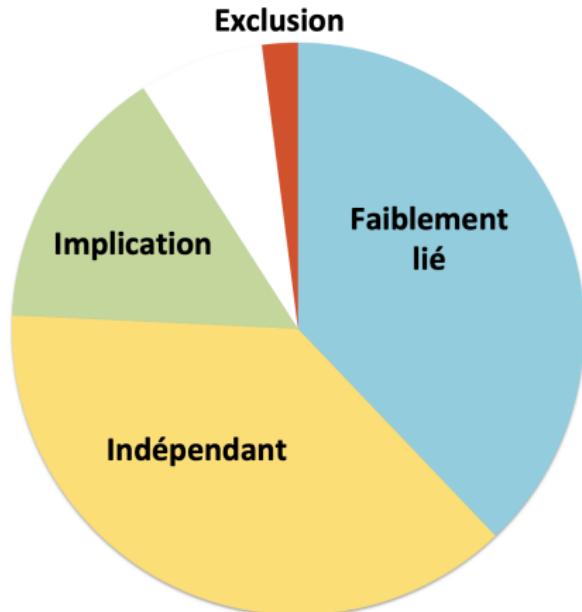
(Pavlick et al., 2015)

Indépendant : *found # party*

Faiblement lié : *husband ~ marry to*

Exclusion : *close | open*

Implication : *tower → building*



Diverses relations sémantiques dans PPDB 2.0

Manque de contrôle sémantique

(Pavlick et al., 2015)

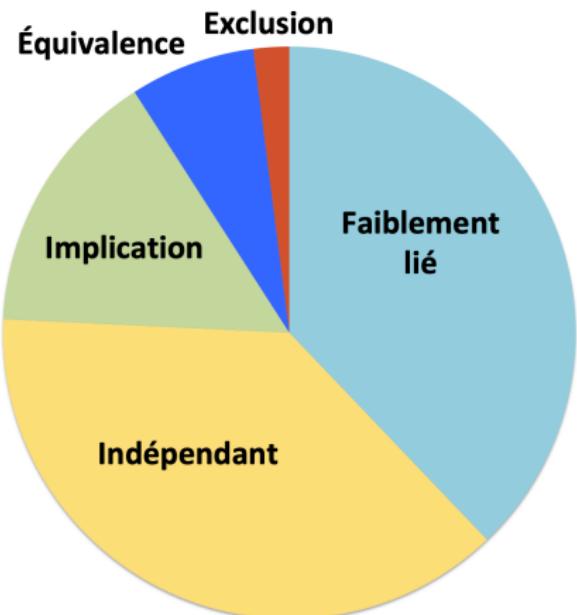
Indépendant : *found # party*

Faiblement lié : *husband ~ marry to*

Exclusion : *close | open*

Implication : *tower → building*

Équivalence : *distant ↔ remote*



Bilan sur le contexte

Difficultés posées par la traduction non littérale

Rôle important de la paraphrase dans diverses tâches en TAL

Manque de contrôle sémantique de la méthode par pivot

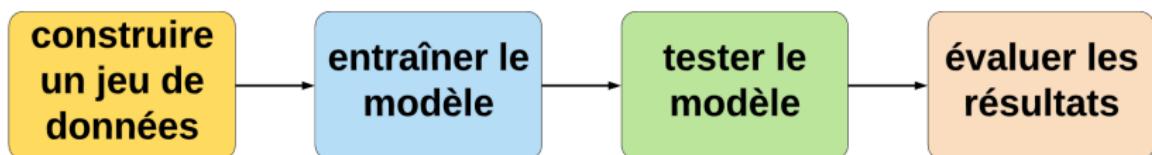
Bilan sur le contexte

Difficultés posées par la traduction non littérale

Rôle important de la paraphrase dans diverses tâches en TAL

Manque de contrôle sémantique de la méthode par pivot

Schéma général pour étudier l'hypothèse :

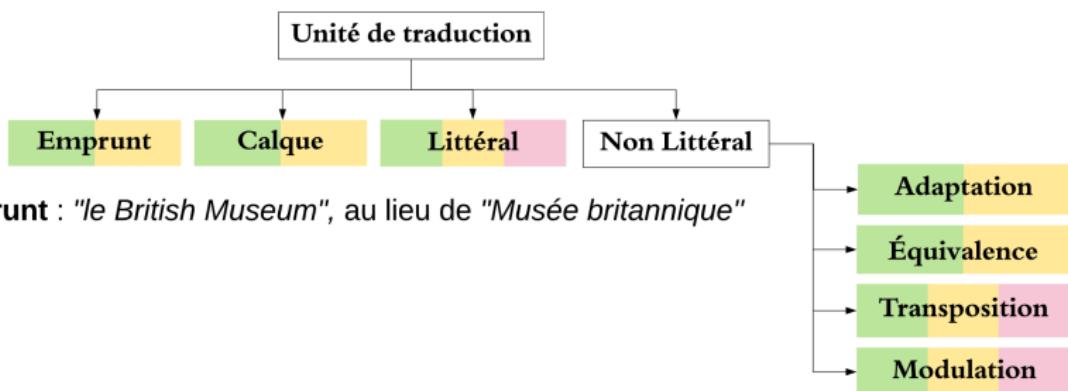


Annotation de corpus parallèle en procédés de traduction

Typologie des procédés : les sept procédés basiques

Principaux travaux précédents

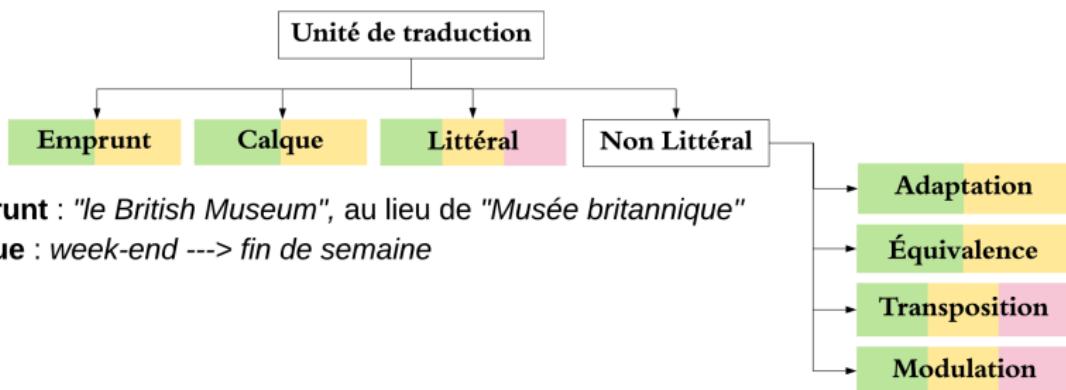
Vinay et Darbelnet (1958), Chuquet et Paillard (1989)
Molina et Hurtado Albir (2002)



Typologie des procédés : les sept procédés basiques

Principaux travaux précédents

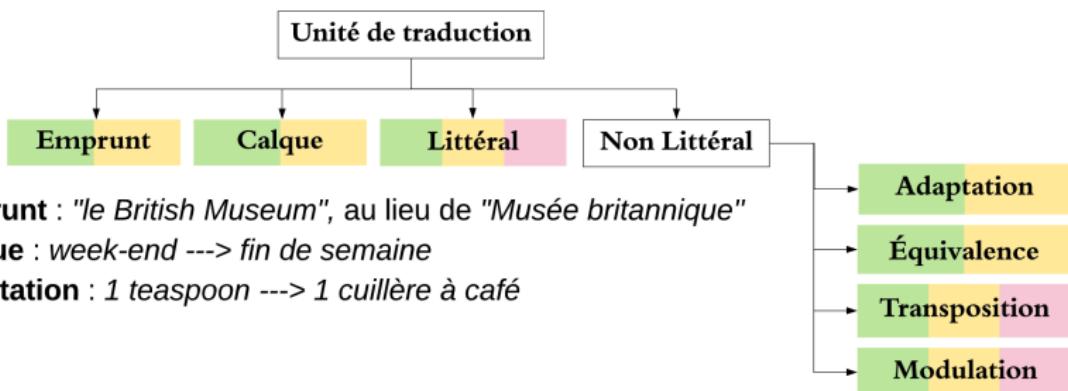
Vinay et Darbelnet (1958), Chuquet et Paillard (1989)
Molina et Hurtado Albir (2002)



Typologie des procédés : les sept procédés basiques

Principaux travaux précédents

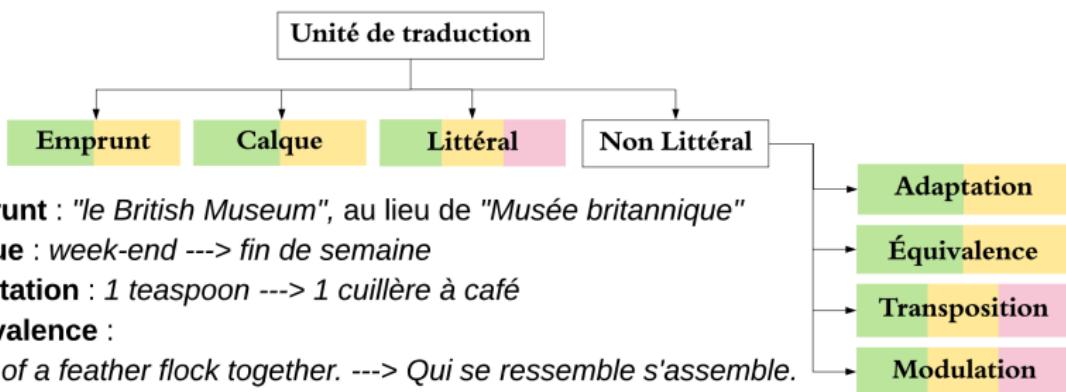
Vinay et Darbelnet (1958), Chuquet et Paillard (1989)
Molina et Hurtado Albir (2002)



Typologie des procédés : les sept procédés basiques

Principaux travaux précédents

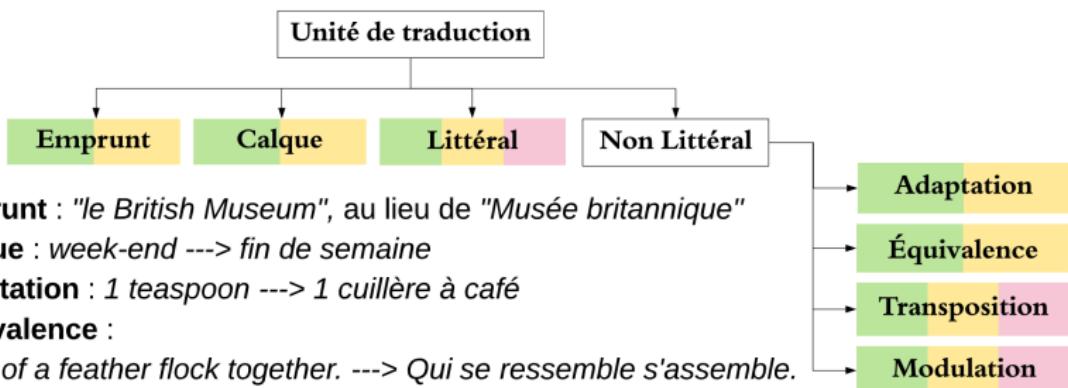
Vinay et Darbelnet (1958), Chuquet et Paillard (1989)
Molina et Hurtado Albir (2002)



Typologie des procédés : les sept procédés basiques

Principaux travaux précédents

Vinay et Darbelnet (1958), Chuquet et Paillard (1989)
Molina et Hurtado Albir (2002)



Emprunt : "le British Museum", au lieu de "Musée britannique"

Calque : week-end ---> fin de semaine

Adaptation : 1 teaspoon ---> 1 cuillère à café

Équivalence :

Birds of a feather flock together. ---> Qui se ressemble s'assemble.

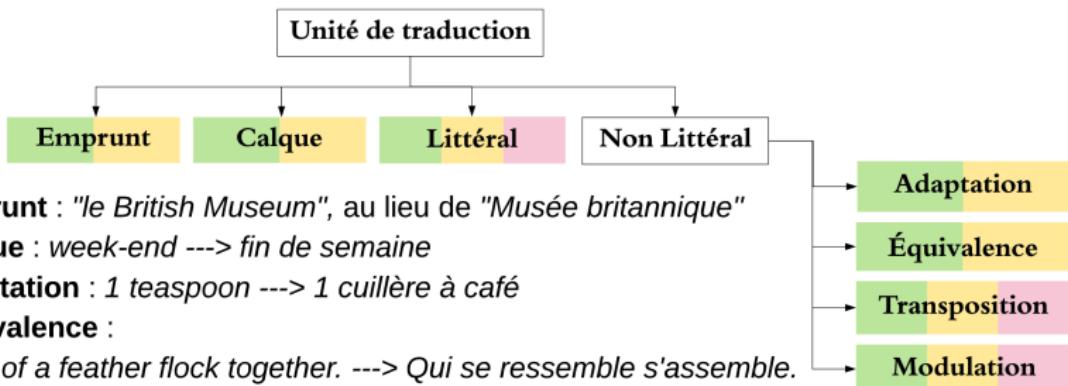
Transposition : after you've burned it ---> après la combustion

Typologie des procédés : les sept procédés basiques

Principaux travaux précédents

Vinay et Darbelnet (1958), Chuquet et Paillard (1989)

Molina et Hurtado Albir (2002)



Emprunt : "le British Museum", au lieu de "Musée britannique"

Calque : week-end ---> fin de semaine

Adaptation : 1 teaspoon ---> 1 cuillère à café

Équivalence :

Birds of a feather flock together. ---> Qui se ressemble s'assemble.

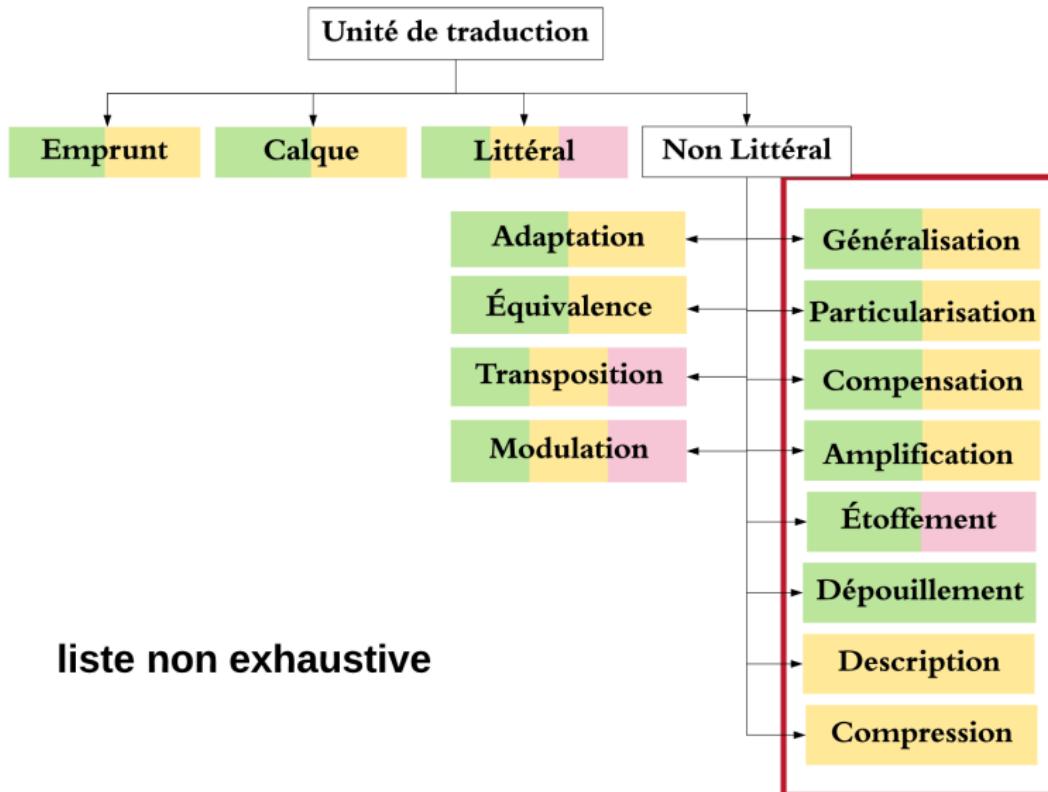
Transposition : after you've burned it ---> après la combustion

Modulation :

and **that scar has stayed with him** for his entire life --->

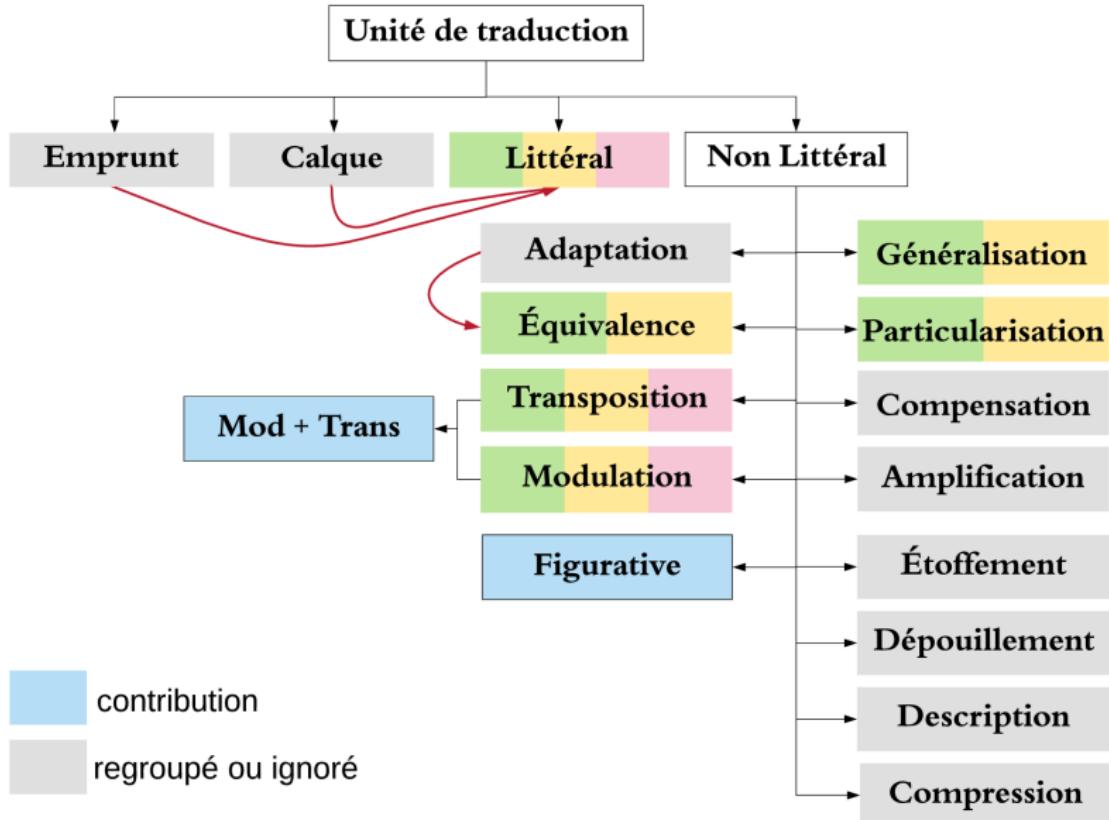
et que, toute sa vie, **il a souffert de ce traumatisme**

Typologie des procédés : les procédés supplémentaires

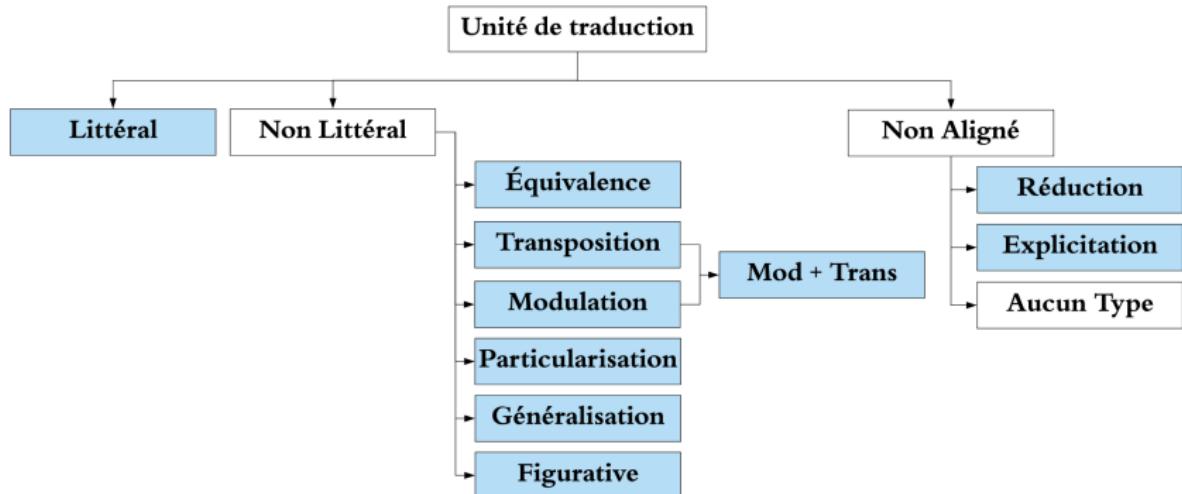


liste non exhaustive

Typologie des procédés : adaptation



Typologie proposée



Exemple pour Modulation + Transposition

there is a huge wave of interest in happiness

le thème du bonheur devient à la mode

Exemples

Traduction littérale : *around the world* → *autour du monde*

Équivalence

around the world → *de par le monde*

Généralisation

the warming heats up the frozen ground

→ *le réchauffement affecte le pergélisol*

Explication

if you drill into that, it's especially the case for men

→ *si on creuse encore un peu, c'est particulièrement vrai pour les hommes*

Présentation de corpus (EN → FR)



Ideas worth
spreading

- TED Talks : conférences
- grand éventail de sujets abordés
- divers phénomènes de traduction
- segmentation de phrase fournie par *WIT*³

Présentation de corpus (EN → FR)



- TED Talks : conférences
- grand éventail de sujets abordés
- divers phénomènes de traduction
- segmentation de phrase fournie par *WIT*³

Nombre de conférences	19
Durée totale	4h30min
Nombre de lignes	2 436
Nombre de tokens	51 930 (EN)
	53 749 (FR)

Un extrait du guide d'annotation

6.1.2 Equivalence

Percentage: 17.79% (of non-literal translations)

Definition There is no change of point of view like in *Modulation* (subsection 6.1.4). A word-for-word translation makes sense but the translator has expressed differently. However, if there exist changes of grammatical categories, the pair should be annotated with *Transposition* (see examples in subsection 6.1.3).

if you 'll pardon the pun → *si vous me passez ce calembour*

sense each other → *se reconnaître entre eux*

Un extrait du guide d'annotation

6.1.2 Equivalence

Percentage: 17.79% (of non-literal translations)

Definition There is no change of point of view like in *Modulation* (subsection 6.1.4). A word-for-word translation makes sense but the translator has expressed differently. However, if there exist changes of grammatical categories, the pair should be annotated with *Transposition* (see examples in subsection 6.1.3).

if you 'll pardon the pun → *si vous me passez ce calembour*

sense each other → *se reconnaître entre eux*

Counterexamples:

magic trick → *tour de magie*

(Word-for-word translation, the category is *Literal*.)

at no time → *à aucun moment*

(We can not say *à aucun temps* in French. This is a literal translation (see Rule 2 of *Literal*).)

Borderline examples:

which is → *soit*

(Borderline with *Literal*. Since it is not the most literal translation, it is annotated as *Equivalence*.)

that 's something the world needs right now → *c' est quelque chose dont le monde a besoin maintenant*

(Borderline with *Literal*, annotated as *Equivalence*.)

Construction itérative du guide d'annotation

Méthodes classiques (Pustejovsky et Stubbs, 2012; Fort 2016)

Catégorie difficile : modulation

Changement de point de vue, glissement sémantique :

and that scar has stayed with him for his entire life

→ *et que, toute sa vie, il a souffert de ce traumatisme*

Construction itérative du guide d'annotation

Méthodes classiques (Pustejovsky et Stubbs, 2012; Fort 2016)

Catégorie difficile : modulation

Changement de point de vue, glissement sémantique :

and that scar has stayed with him for his entire life

→ *et que, toute sa vie, il a souffert de ce traumatisme*

Règle précise : Modulation figée est annotée comme *Littéral* :

a life jacket → **un gilet de sauvetage**

Outil d'annotation : Application Web Yawat

Alignment bilingue :

it 's only possible to see that insight when you step back and look

il n'est possible de faire cette observation qu'en prenant du recul

Outil d'annotation : Application Web Yawat

Alignment bilingue :

it 's only possible to see that insight when you step back and look

il n'est possible de faire cette observation qu'en prenant du recul

Attribution de catégorie et modification de frontière :

remove 'when' from this group
dissolve this group

-
- label group as ...
 - literal
 - equivalence
 - transposition
 - modulation
 - modulation_transposition
 - generalization
 - particularization
 - figurative
 - lexical_shift
 - uncertain
 - translation_error

Outil d'annotation : Application Web Yawat

Impossible d'aligner : *Réduction, Explication, Aucun Type*

and the official dogma runs like this : if we are interested in maximizing the welfare of our citizens , the way to do that is to maximize individual freedom .

Réduction

il dit ceci : pour maximiser le bien-être des citoyens il faut maximiser leur liberté individuelle .

Explication

Outil d'annotation : Application Web Yawat

Impossible d'aligner : *Réduction, Explication, Aucun Type*

and the official dogma runs like this : if we are interested in maximizing the welfare of our citizens , the way to do that is to maximize individual freedom .

Réduction

il dit ceci : pour maximiser le bien-être des citoyens il faut maximiser leur liberté individuelle .

Explication

Trois catégories de plus pour l'annotation

changement lexical mineur

erreur de traduction

incertain

Étude de contrôle : deux annotations indépendantes

100 paires de phrases (3 055 tokens EN et 3 238 tokens FR)

Mêmes frontières :

κ	%EN tokens
0,67	72,60%

Étude de contrôle : deux annotations indépendantes

100 paires de phrases (3 055 tokens EN et 3 238 tokens FR)

Mêmes frontières :

κ	%EN tokens
0,67	72,60%

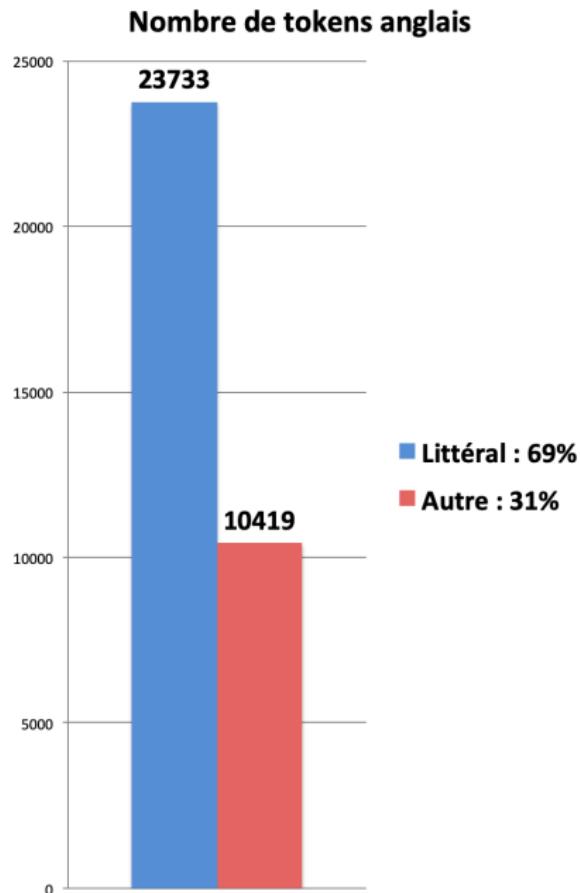
Frontières différentes :

κ	%EN tokens
0,62	85,56%

Processus d'annotation en trois passes

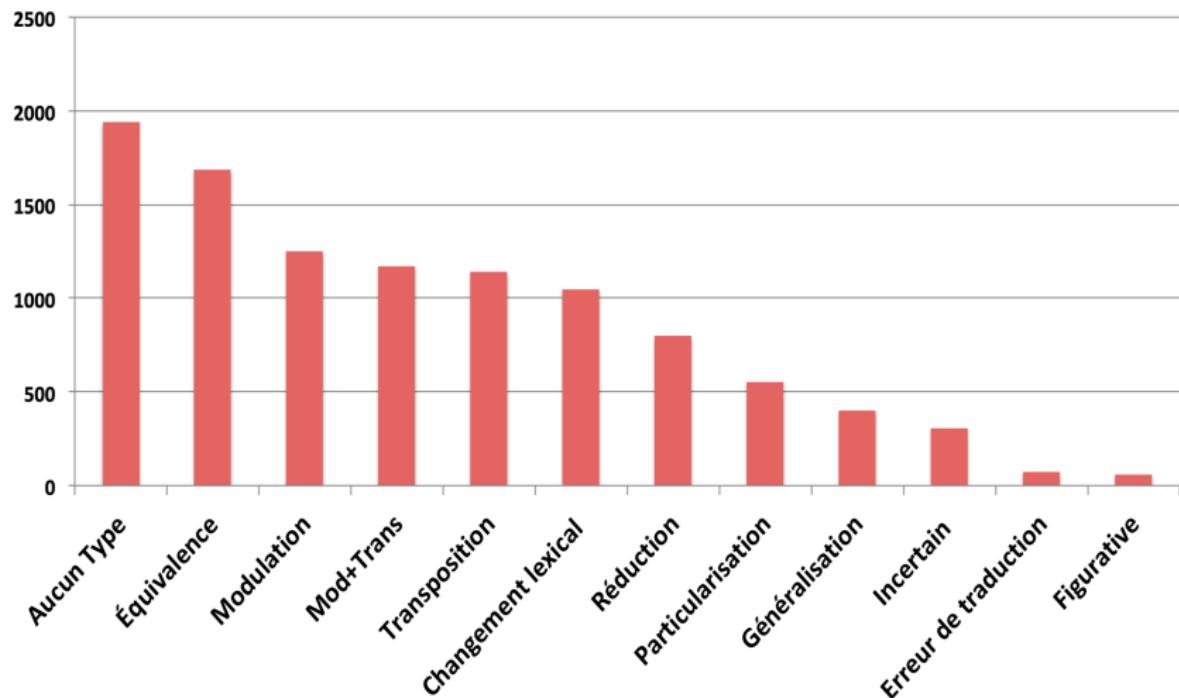
	Annotateur 1	Annotateur 2
1	Aligner des unités de traduction, attribuer des catégories	-
2	-	Vérifier, modifier s'il existe un désaccord
3	Discuter des différences, faire le consensus	

Statistiques d'annotation (corpus EN-FR de TED Talks)

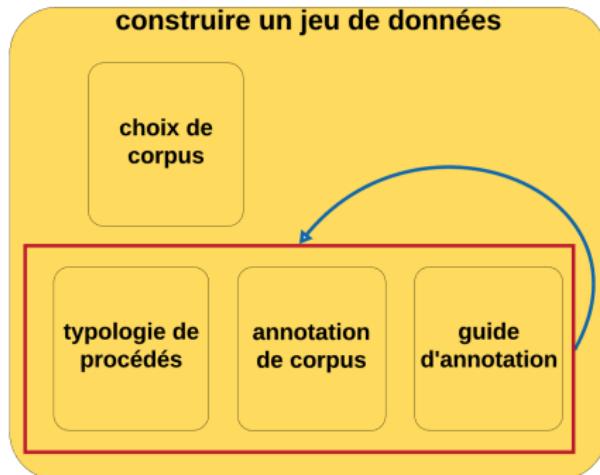


Statistiques d'annotation (corpus EN-FR de TED Talks)

Tokens anglais annotés par des catégories autres que “Littéral” :

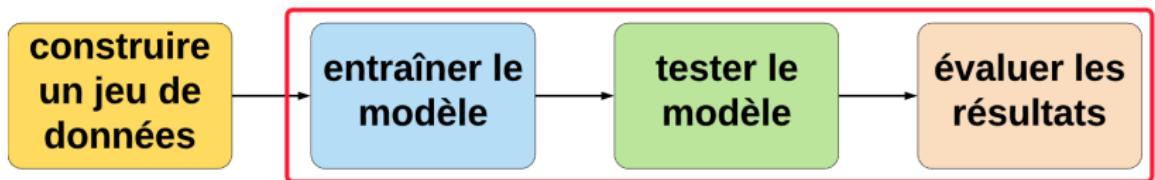


Bilan sur l'annotation manuelle



- Typologie de procédés
- Choix de corpus parallèle
- Construction itérative du guide d'annotation
- Étude de contrôle
- Annotation de corpus

Reconnaissance automatique



Étude sur la traduction non littérale

Étudier l'impact sur l'alignement automatique de mots

(Dorr et al., 2002; Deng and Xue, 2017)

Développer une mesure de traduction littérale

(Carl and Schaeffer, 2017)

Comparer la traduction automatique et humaine (Ahrenberg, 2017)

Travaux liés

Étude sur la traduction non littérale

Étudier l'impact sur l'alignement automatique de mots

(Dorr et al., 2002; Deng and Xue, 2017)

Développer une mesure de traduction littérale

(Carl and Schaeffer, 2017)

Comparer la traduction automatique et humaine (Ahrenberg, 2017)

Détection automatique des divergences de traduction

Filtrer automatiquement des couples de phrases divergentes

(Carpuat et al., 2017, Vyas et al., 2018, Pham et al., 2018)

Nombre d'instances par catégorie

Littéral	3 771
Équivalence	289
Contient_transposition <i>(Transposition (289) +Mod_Trans (53))</i>	342
Modulation	195
Particularisation	215
Généralisation	86

Catégorie *Figurative* : ignorée (seulement 13 instances)

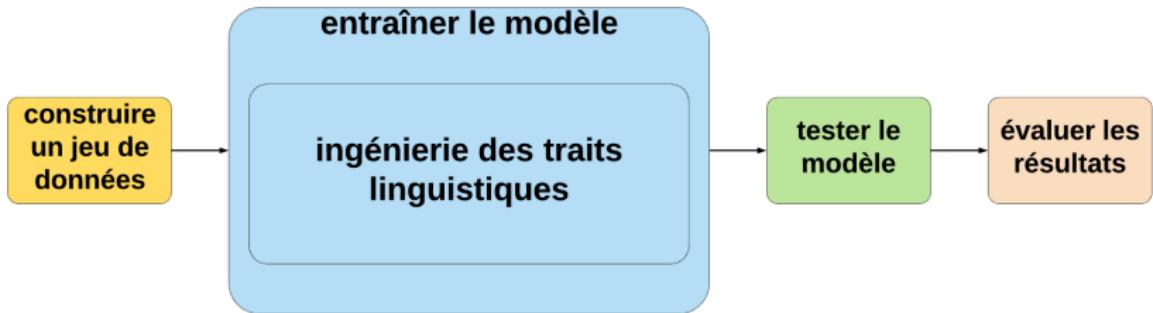
Objectif de classification

Scénario simplifié : frontière des segments fournie

Entrée : *are a better match to* → *correspondent mieux à*

Sortie : la catégorie **contient_transposition**

- sans changement de sens
- changement de catégories grammaticales



Traits : surface

- nombre de tokens anglais (len_e), français (len_f)
- ratio de ces nombres (len_e/len_f , len_f/len_e)
- distance Levenshtein entre deux segments

Traits : analyse morpho-syntaxique

- comptage du nombre d'occurrence des étiquettes de PoS :

	ADJ	DET	NOUN	...	INTJ
anglais	1	0	0	...	0
français	0	1	2	...	0

- similarité cosinus entre ces deux vecteurs
- vérification du patron de changement de séquence de PoS
vision impairment → l'altération visuelle
NOUN NOUN → DET NOUN ADJ

Traits syntaxiques en constituants

trait binaire (0 ou 1) :

étiquettes PoS :

adapt (verbe) → *adaptation (nom)*

étiquettes du nœud non terminal :

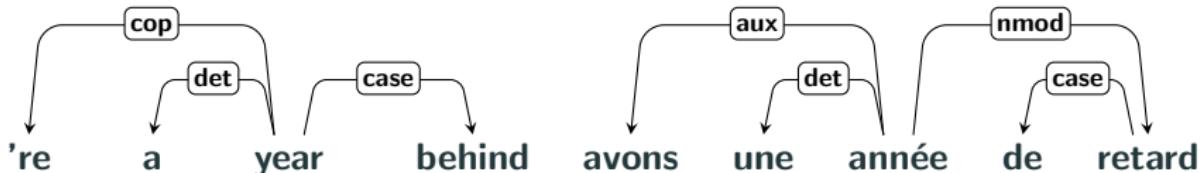
stamp a letter into it (groupe verbal)

→ avec une lettre en creux (*groupe nominal prépositionnel*)

catégorie des étiquettes :

adjacent (adjectif) → *qui n'était pas loin (groupe verbal)*

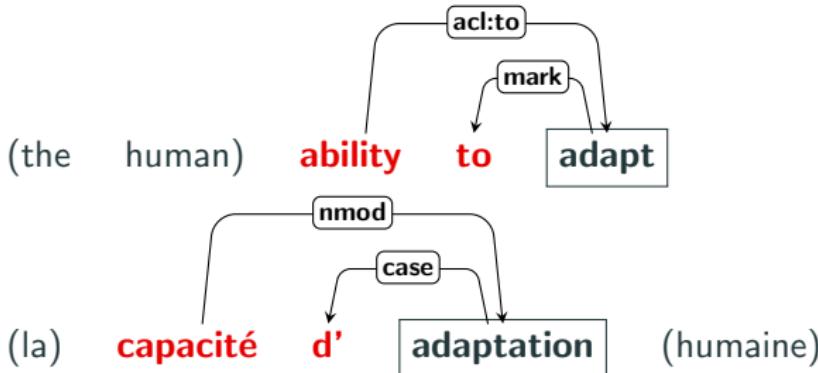
Traits syntaxiques en dépendance : à l'intérieur



	amod	det	nmod	case	...	nsubj
anglais	0	1	0	1	...	0
français	0	1	1	1	...	0

Comptage des étiquettes de relation de dépendance

Traits syntaxiques en dépendance : à l'extérieur



	acl	mark	nmod	case	...	nsubj
anglais	1	1	0	0	...	0
français	0	0	1	1	...	0

Comptage des étiquettes de relation de dépendance

ConceptNet

- similarité cosinus entre les plongements

(*ConceptNet Numberbatch* (Speer et al., 2017))

expressions multi-mots :

climate_change → **changement_climatique**

moyenne des plongements lexicaux sur les mots pleins :

're less burdened by → **est moins un fardeau**

ConceptNet

- similarité cosinus entre les plongements
(*ConceptNet Numberbatch* (Speer et al., 2017))
expressions multi-mots :
climate_change → **changement_climatique**
moyenne des plongements lexicaux sur les mots pleins :
're less burdened by → **est moins un fardeau**
- comment la paire est liée dans la ressource
0 : lié directement
1 : lié indirectement :
complete ← **complet / entier** → **total**
2 : pas lié

ConceptNet

- similarité cosinus entre les plongements
(*ConceptNet Numberbatch* (Speer et al., 2017))
expressions multi-mots :
climate_change → **changement_climatique**
moyenne des plongements lexicaux sur les mots pleins :
're less burdened by → **est moins un fardeau**
- comment la paire est liée dans la ressource
0 : lié directement
1 : lié indirectement :
complete ← **complet / entier** → **total**
2 : pas lié
- **pourcentage** des tokens liés indirectement
deceptive ← **illusoire** → **une illusion**

Traits : alignement automatique de mots

Table de traduction lexicale (Berkeley Word Aligner)

- moyenne des entropies de traduction lexicale (mots pleins)

Table de traduction lexicale (Berkeley Word Aligner)

- moyenne des entropies de traduction lexicale (mots pleins)
- pondération lexicale bi-directionnelle sur les mots pleins
(Koehn et al., 2003)

$$\text{lex}(e|f, A) = \prod_{i=1}^{\text{length}(e)} \frac{1}{|\{j|(i,j) \in A\}|} \sum_{\forall(i,j) \in A} w(e_i|f_j) \quad (1)$$

Table de traduction lexicale (Berkeley Word Aligner)

- moyenne des entropies de traduction lexicale (mots pleins)
- pondération lexicale bi-directionnelle sur les mots pleins
(Koehn et al., 2003)

$$\text{lex}(e|f, A) = \prod_{i=1}^{\text{length}(e)} \frac{1}{|\{j|(i,j) \in A\}|} \sum_{\forall(i,j) \in A} w(e_i|f_j) \quad (1)$$

- somme de différence de probabilités de traduction lexicale
 alternatives → **alternatives** $P = 0,4$
alternatives → solutions de remplacement $P = 0,07$

Plan expérimental

Vecteur de traits linguistiques :

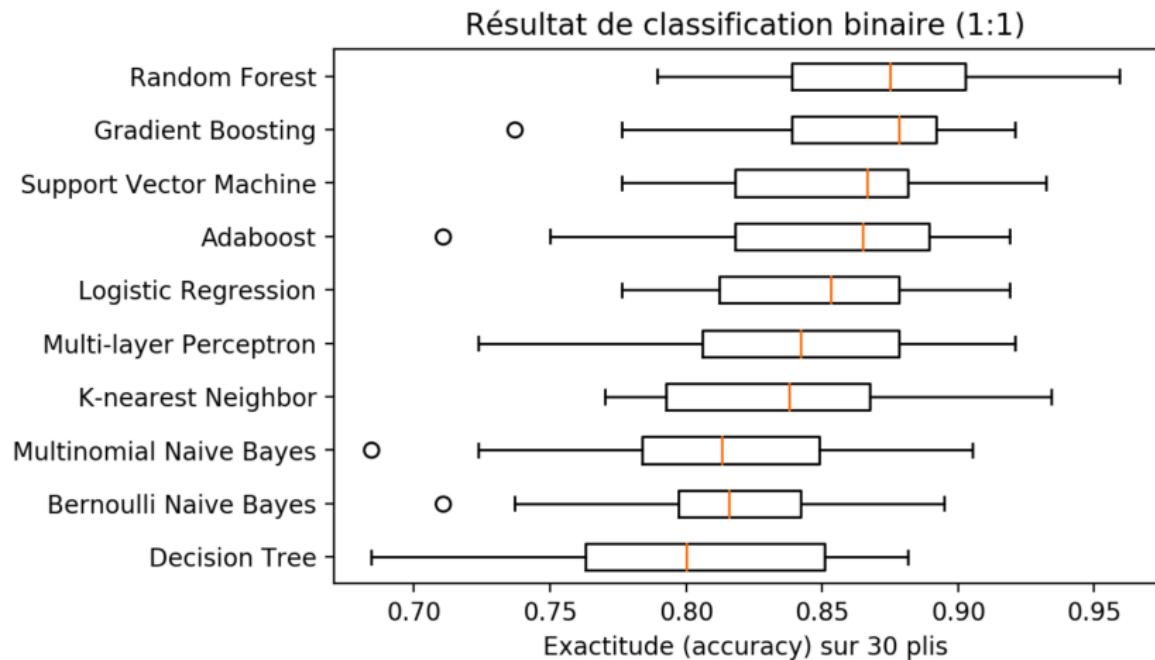
Traits	Surface	POS	Constituant	Dépendance	ConceptNet	Alignement	Total
Nb colonnes	5	37	1	137	4	14	198

- Littéral (3771) vs Non_littéral (5 classes, 1127)
- Entraîner différents classifieurs avec *Scikit-learn*
- Les hyperparamètres sont optimisés
- Validation croisée à 30 plis
- Baseline : prédire toujours la classe majoritaire

Résultat

Nombre d'instances : Littéral 1127, Non-Littéral 1127

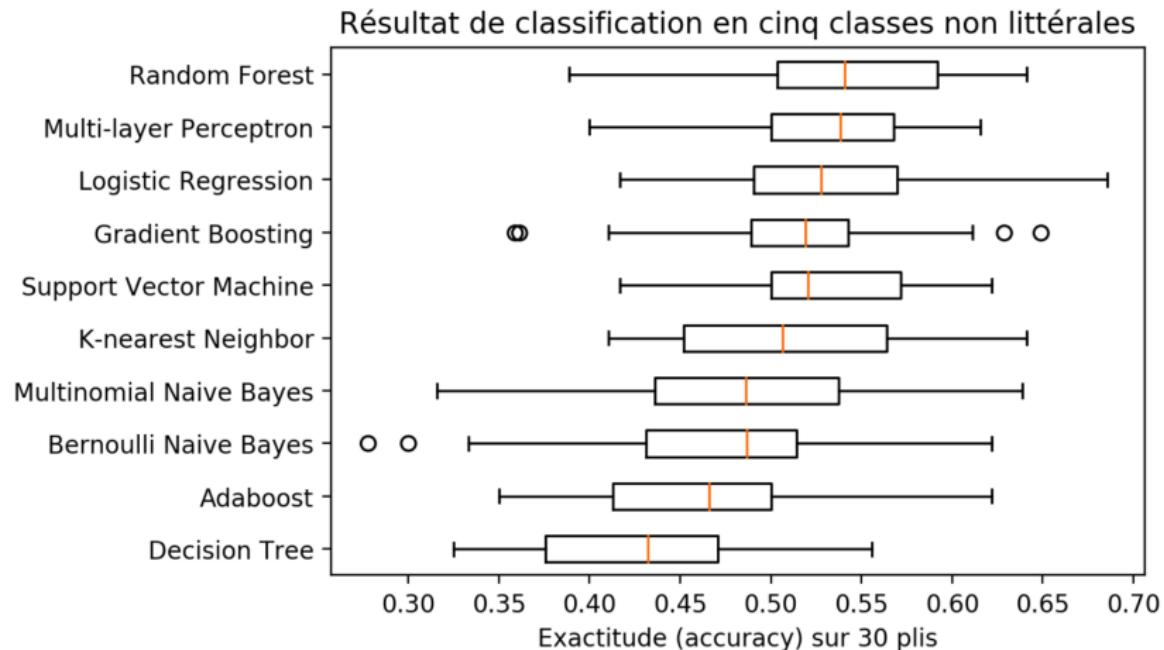
Baseline (classe majoritaire): 0.50

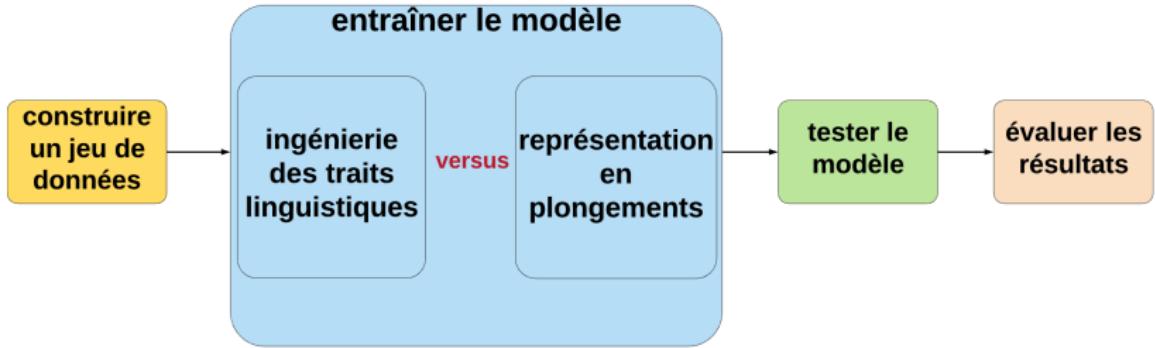


Résultat

Nombre d'instances : en total 1127, distribution déséquilibrée

Baseline (classe majoritaire): 0.30





Comparer différents traits¹

Vecteur de traits linguistiques

Traits	Surface	POS	Constituant	Dépendance	ConceptNet	Alignement	Total
Nb colonnes	5	37	1	137	4	14	198

¹Travail en collaboration avec Pooyan Safari, doctorant du LIMSI

Comparer différents traits¹

Vecteur de traits linguistiques

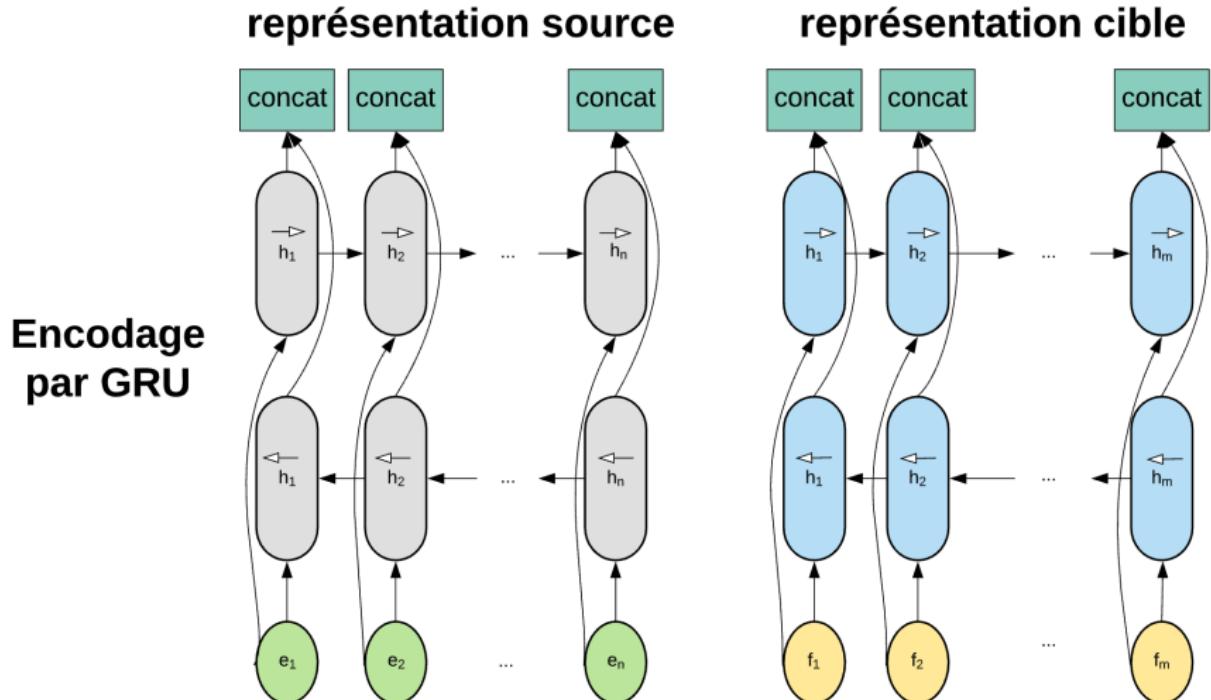
Traits	Surface	POS	Constituant	Dépendance	ConceptNet	Alignement	Total
Nb colonnes	5	37	1	137	4	14	198

Seulement en plongements

- Plongements de caractère initialisés aléatoirement (dimension 10)
- Plongements lexicaux pré-entraînés avec FastText (Bojanowski et al., 2017) (dimension 100)

¹Travail en collaboration avec Pooyan Safari, doctorant du LIMSI

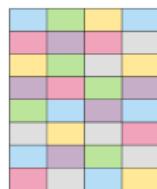
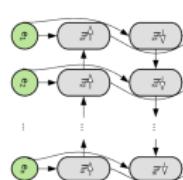
Encodage par des unités de GRU bidirectionnelles



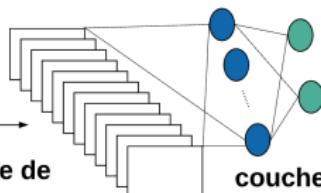
Classifieur CNN

Matrice d'alignement de mots (produit scalaire) + classifieur CNN

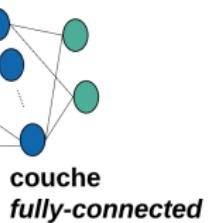
représentation
source 



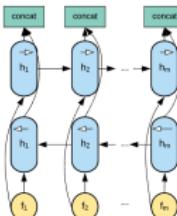
couche de convolution



couche de pooling

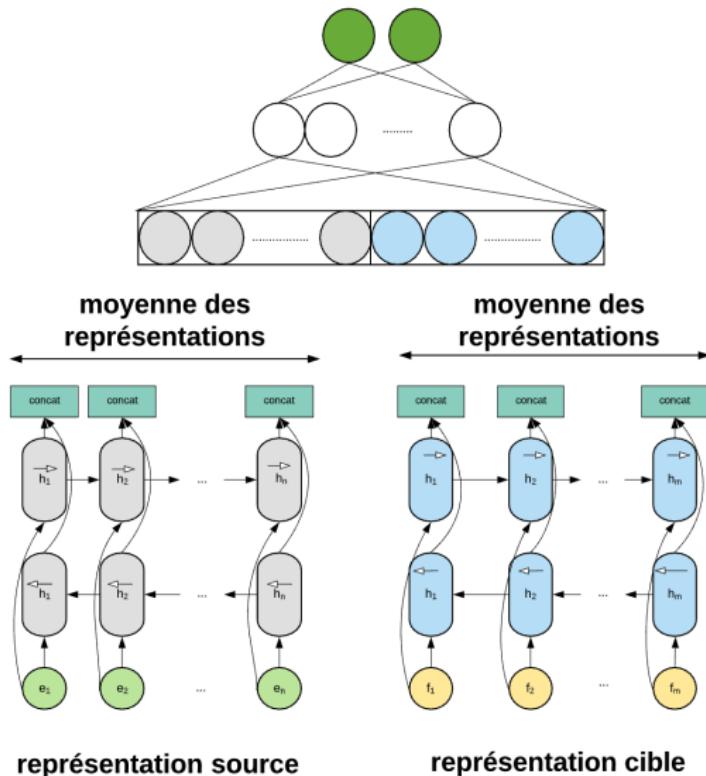


représentation
cible 



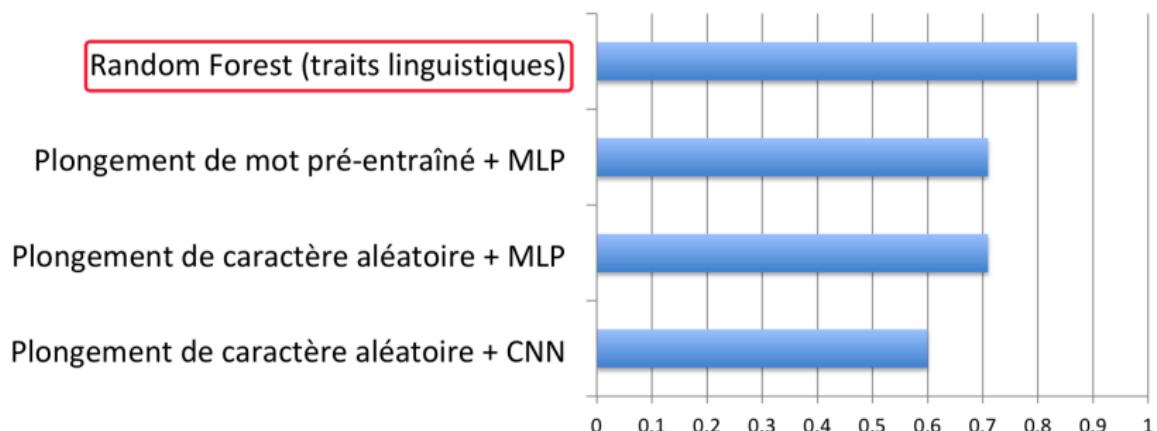
Classifieur MLP

Représentation en moyenne + classifieur MLP



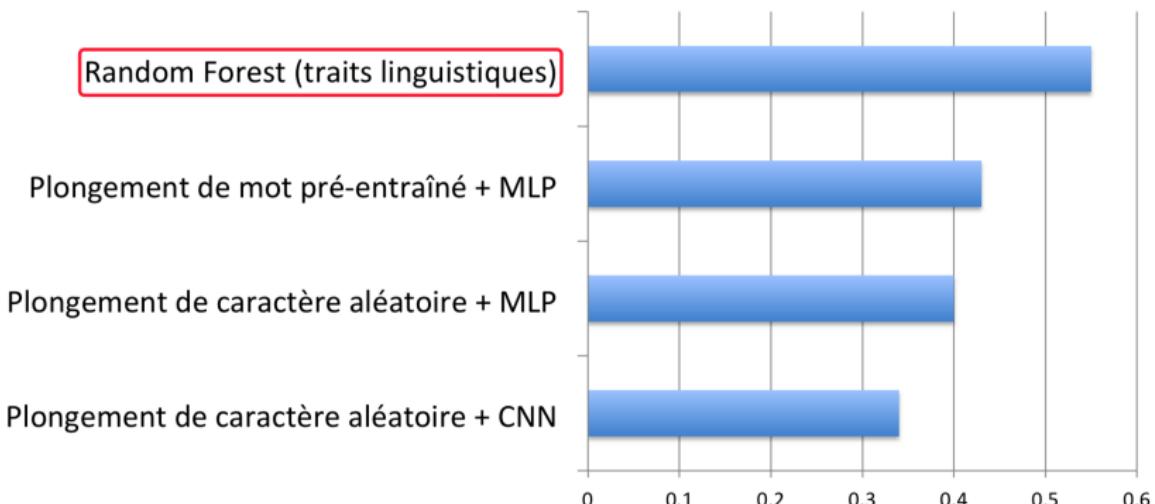
Résultats : exactitude moyenne sur 5 plis

Classification binaire (1:1)

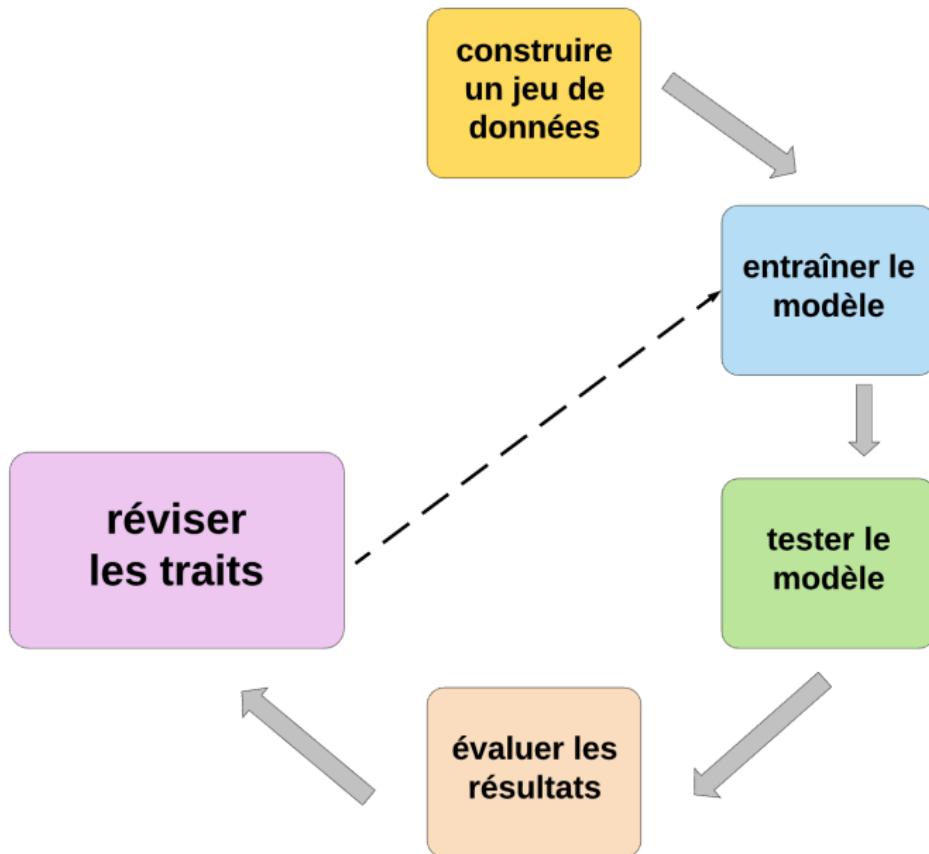


Résultats : exactitude moyenne sur 5 plis

Classification en 5 classes non littérales



Ajouter des traits sensibles au contexte



Motivation

- Focaliser sur la classification des classes non littérales
- Exploiter l'information contextuelle, réviser les traits
- Comparer avec le système précédent
traits indépendants du contexte

Exemple de traduction dépendante du contexte

became the basis for → *ont inspiré* (traduction possible ?)

Exemple de traduction dépendante du contexte

became the basis for → **ont inspiré** (possible en contexte !)

Steve's columns became the basis for a book, which was turned into a movie.

Les chroniques de Steve ont inspiré un livre, qui a été adapté à l'écran.

Exemple de traduction dépendante du contexte

became the basis for → **ont inspiré** (possible en contexte !)

Steve's columns became the basis for a book, which was turned into a movie.

Les chroniques de Steve ont inspiré un livre, qui a été adapté à l'écran.

Jeu de données (selon le glissement de sens)

Équivalence (équivalence + transposition)	710
Implication textuelle (généralisation + particularisation)	384
Lié en thématique (modulation, mod+trans)	305

Inférence lexicale monolingue sensible au contexte

Shwartz et Dagan (2016), Vyas et Carpuat (2017)

Contextes	Relation en contexte
<i>Roughly 1,500 gold and silver pieces were found [...]</i>	Indépendant
<i>The labor leader Kevin Rudd was found to have gone to a strip club during a trip</i>	

Inférence lexicale monolingue sensible au contexte

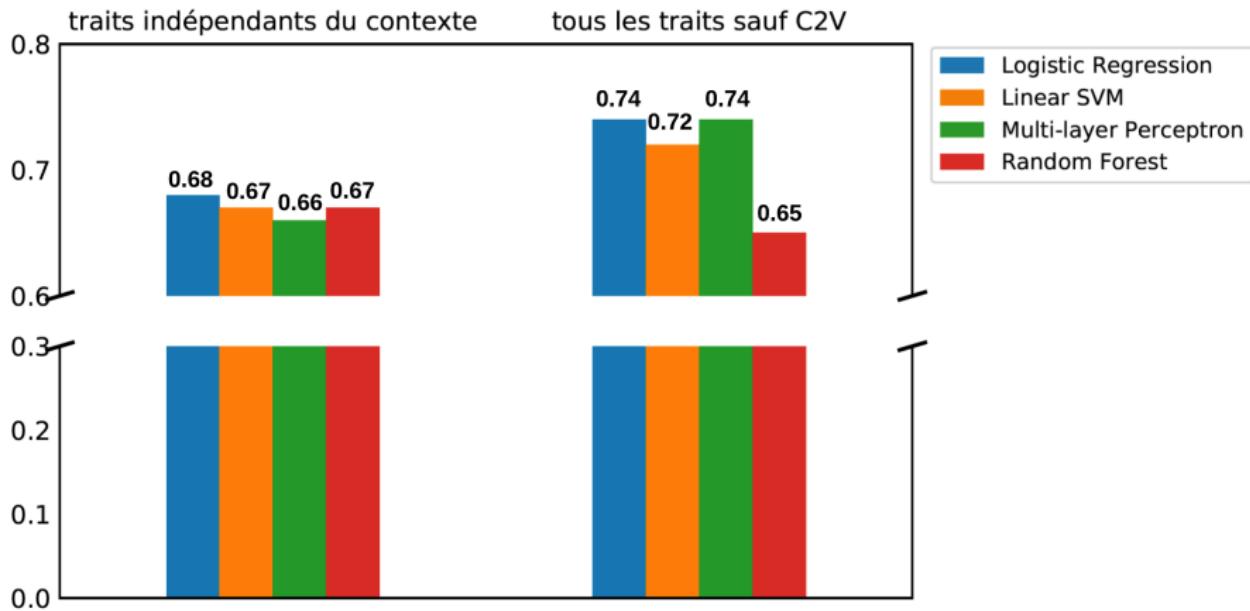
Shwartz et Dagan (2016), Vyas et Carpuat (2017)

Contextes	Relation en contexte
<i>Roughly 1,500 gold and silver pieces were found [...]</i>	Indépendant
<i>The labor leader Kevin Rudd was found to have gone to a strip club during a trip</i>	

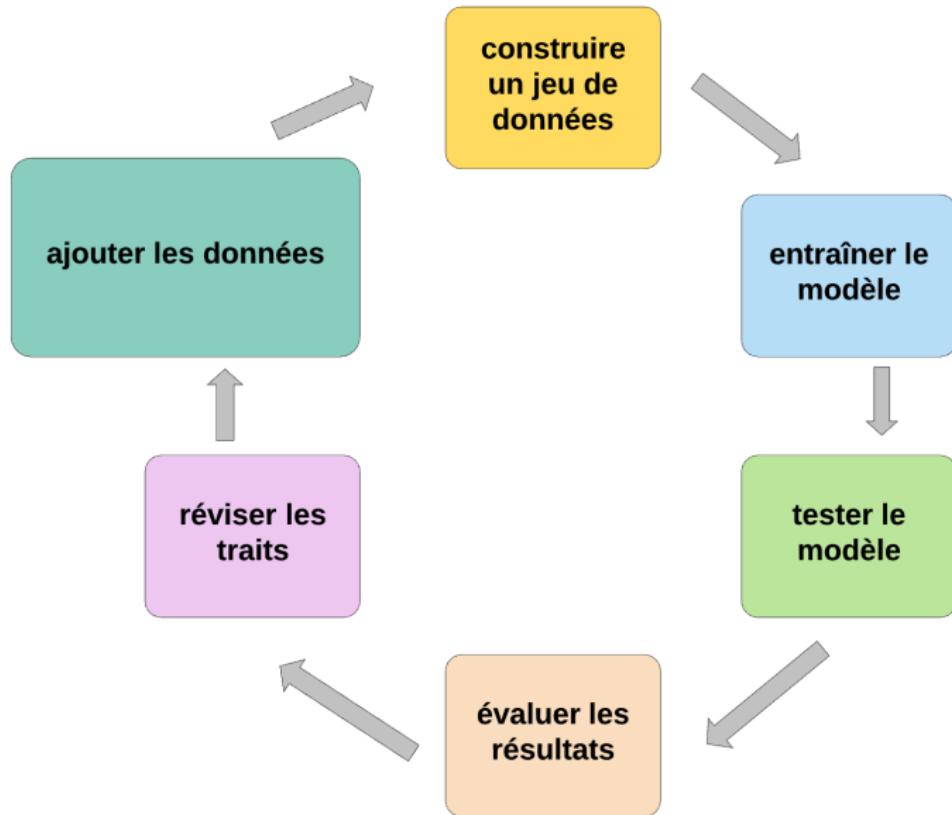
Adaptation des traits pour notre tâche cross-lingue

Représentation du contexte : ajouter des traits avec ELMo (Peters et al., 2018) et context2vec (Melamud et al., 2016)

Résultats (F1 moyenne pondérée)



Extension des études au couple anglais-chinois

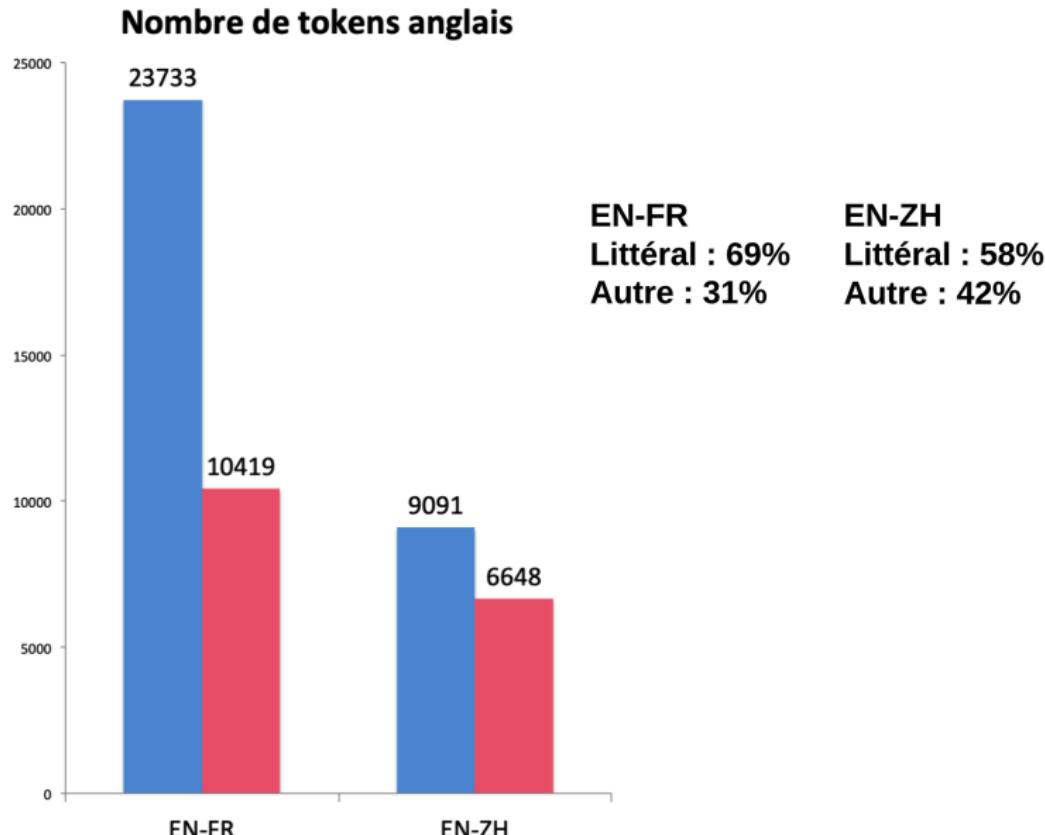


Annotation du corpus²

- Constituer un corpus de multi-genres
- Adapter et enrichir le guide d'annotation
- Tester l'universalité de la typologie des procédés
- kappa de Cohen (deux annotations sur 100 paires de phrases):
0,58 (modéré)

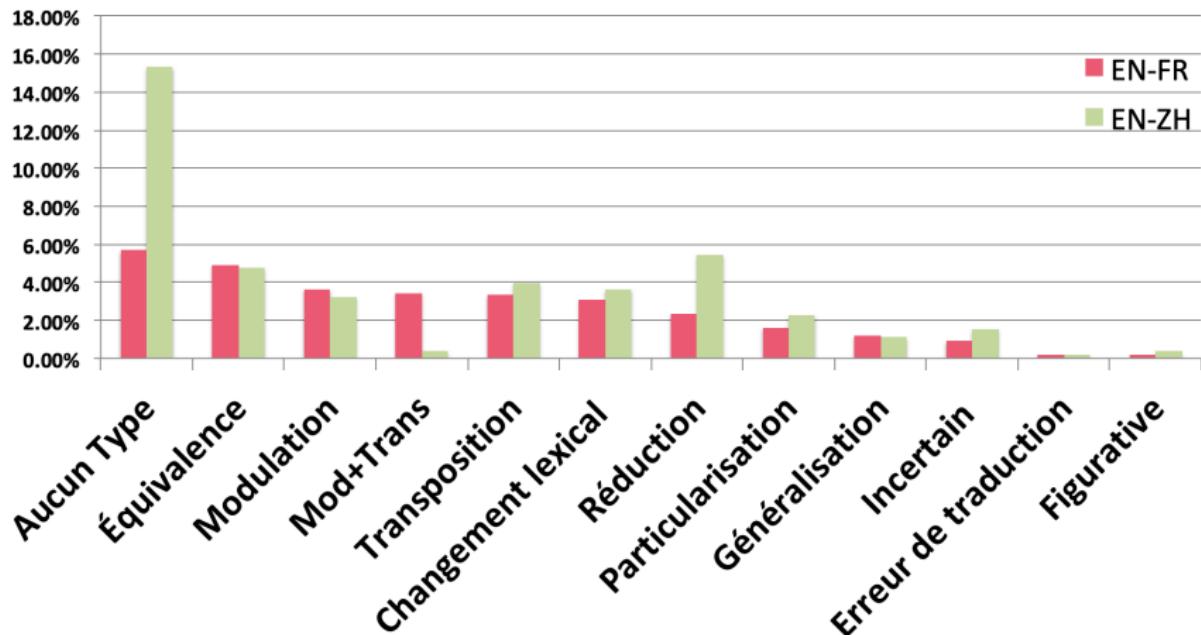
²Travail en collaboration avec Lufei Liu et Xinyi Zhong (master traduction)

Statistiques d'annotation (contraste EN-FR, EN-ZH)



Statistiques d'annotation (contraste EN-FR, EN-ZH)

Pourcentage de tokens anglais par catégorie



Pistes de validation

Aide à la construction de ressource de paraphrase

- Influence de différentes langues pivot
- Influence de différents procédés de traduction

Mesure d'évaluation complémentaire pour

- l'alignement automatique de mots
- la qualité de la traduction automatique

Texte :

Sur ce tronçon de rail, il y a un feu de signalisation défectueux. Enfin, j'imagine qu'il doit être défectueux, parce qu'il est presque toujours rouge ; on s'y arrête quasiment tous les jours, parfois quelques secondes, parfois plusieurs minutes **d'affilée**.

Requête : "d'affilée"

Liste de candidats proposés :

successivement

consécutivement

de suite

sans interruption

(deuxième groupe : vide)

durer

} **paraphrases strictes**
+ **général ou + spécifique**
rélié à

Requête : “d'affilée”

Paraphrase stricte “successivement”

 [...] travailler plus de quatre heures **d'affilée**.

 [...] work for more than four hours **consecutively**.

 [...] accrue pension rights earned **consecutively**.

 [...] accumuler des droits à la retraite gagnés **successivement**.

Requête : “d'affilée”

Réécriture liée sémantiquement “durer”

 [...] avec un troisième gouvernement minoritaire **d'affilée**.

 [...] with a third minority government **in a row**.

 [...] for four quarters **in a row**.

 [...] 连续 [...] 四季 [...]

 [...] 连续几个小时 [...]

 [...] for hours **on end**.

 [...] for weeks **on end** during the height of summer.

 [...] quand la chaleur peut **durer** pendant plusieurs semaines.

Expériences en compréhension écrite

- Participation des étudiants chinois adultes
- 15 niveau A2, 11 niveau B2
- Évaluation du prototype par questionnaire

Expériences en compréhension écrite

- Participation des étudiants chinois adultes
- 15 niveau A2, 11 niveau B2
- Évaluation du prototype par questionnaire

De par le monde, il existe des tas d'endroits où l'on peut observer des phénomènes étonnantes. Ainsi, j'ai pu voir à la lisière d'une ville, dans un pâté de maisons isolé, des chaumières qui semblaient désaffectées, où pourtant habitaient des personnes qui semblaient inaptes au travail. On y voyait un ivrogne, ayant son calepin à la main, errer avec dans son sillage certains de ses compagnons. Quand il n'écrivait pas, il racontait des récits édifiants sur les prouesses accomplies par des personnages dans sa tête, ce qui faisait grincer des dents à ses amis. Ils lui en voulaient de les effrayer ainsi. Il s'avère que tous ces malheureux vivaient ainsi dans un perpétuel ennui. Pour leur sauver la mise, il aurait fallu qu'ils prennent leurs problèmes à bras le corps pour remédier à cette triste situation. Ils en étaient tout bonnement incapables, si quelqu'un ne venait pas mettre au point avec eux un moyen de s'en sortir. Mais prôner n'importe quel travail n'est pas une solution ! Un jour, un homme les entraîna pour travailler pour lui contre une poignée de pièces. Mais cette solution trop facile était un piège : cet homme était un voleur et il les incitait à jouer les fouineurs pour dévoiler des secrets. Ayant bravé les interdits, ces malheureux se sont trouvés en garde à vue. En plus, dans la foulée, leurs demeures ont été endommagées par une pluie battante. Mais bon sang ! Quelle tragédie !

Expériences en compréhension écrite

- Participation des étudiants chinois adultes
- 15 niveau A2, 11 niveau B2
- Évaluation du prototype par questionnaire

De par le monde, il existe des tas d'endroits où l'on peut observer des phénomènes étonnantes. Ainsi, j'a_____ autour du monde _____ té de maisons isolé, des chaumières qui semblaient dé_____ autour du globe _____ personnes qui semblaient inaptes au travail. On y voyait un iv_____ dans le monde entier _____ avec dans son sillage certains de ses compagnons. Quand il _____ à travers le monde _____ts sur les prouesses accomplies par certains pe_____ aux quatre coins du monde _____s à ses amis. Ils lui en voulaient de les effrayer ainsi. Il s'avère que tous ces malheureux vivaient ainsi dans un perpétuel ennui. Pour leur sauver la mise, il aurait fallu qu'ils prennent leurs problèmes à bras le corps pour remédier à cette triste situation. Ils en étaient tout bonnement incapables, si quelqu'un ne venait pas mettre au point avec eux un moyen de s'en sortir. Mais prôner n'importe quel travail n'est pas une solution ! Un jour, un homme les entraîna pour travailler pour lui contre une poignée de pièces. Mais cette solution trop facile était un piège : cet homme était un voleur et il les incitait à jouer les fouineurs pour dévoiler des secrets. Ayant bravé les interdits, ces malheureux se sont trouvés en garde à vue. En plus, dans la foulée, leurs demeures ont été endommagées par une pluie battante. Mais bon sang ! Quelle tragédie !

Expériences en compréhension écrite

- Participation des étudiants chinois adultes
- 15 niveau A2, 11 niveau B2
- Évaluation du prototype par questionnaire

De par le monde, il existe des tas d'endroits où l'on peut observer des phénomènes étonnantes. Ainsi, j'ai travelled over the world et je have seen de curious things. J'ai vu strange houses, des isolated houses, des chaumières qui semblaient inhabited. Des personnes qui semblaient unfit for work. On y voyait un man avec his sash certains de ses compagnons. Quand il y avait problems sur les roads accomplies par des personnages strange. Il a été seen à ses amis. Ils lui wanted de les scare ainsi. Il a été seen dans un perpetual ennui. Pour leur saving la misery, il a été seen bras le body pour remedy à cette triste situation. Ils en étaient totally incapable, si quelqu'un ne venait pas put them in order avec eux un moyen de s'en sortir. Mais advocacy n'est pas une solution ! Un jour, un homme les entraîna pour travailler pour lui contre a pile of coins. Mais cette solution trop facile était un piège : cet homme était un voleur et il les incitait à jouer les tricks pour dévoiler des secrets. Ayant defied les interdits, ces malheureux se sont trouvés en garde à vue. En plus, in the end, leurs demeures ont été endommagées par a heavy rain. Mais bon sang ! Quelle tragédie !

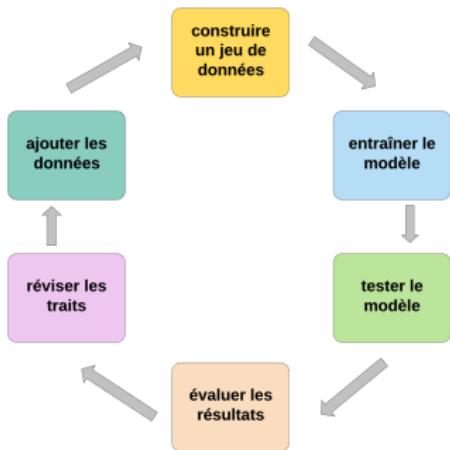
Expériences en compréhension écrite

- Participation des étudiants chinois adultes
- 15 niveau A2, 11 niveau B2
- Évaluation du prototype par questionnaire

De par le monde, il existe des tas d'endroits où l'on peut observer des phénomènes étonnantes. Ainsi, j'ai over the world dé from around the world iv autour du monde il autour du globe da dans le monde entier Il s à travers le monde au aux quatre coins du monde té de maisons isolé, des chaumières qui semblaient personnes qui semblaient inaptes au travail. On y voyait un avec dans son sillage certains de ses compagnons. Quand ts sur les prouesses accomplies par des personnages à ses amis. Ils lui en voulaient de les effrayer ainsi. ainsi dans un perpétuel ennui. Pour leur sauver la mise, il bras le corps pour remédier à cette triste situation. Ils en étaient tout bonnement incapables, si quelqu'un ne venait pas mettre au point avec eux un moyen de s'en sortir. Mais prôner n'importe quel travail n'est pas une solution ! Un jour, un homme les entraîna pour travailler pour lui contre une poignée de pièces. Mais cette solution trop facile était un piège : cet homme était un voleur et il les incitait à jouer les fouineurs pour dévoiler des secrets. Ayant bravé les interdits, ces malheureux se sont trouvés en garde à vue. En plus, dans la foulée, leurs demeures ont été endommagées par une pluie battante. Mais bon sang ! Quelle tragédie !

⇒ Retours positifs et conseils pour la conception

Conclusion et perspectives



- Cycle de travail pour valider l'hypothèse
- Typologie de procédés de traduction et annotation du corpus ([RECITAL 2018; LR4NLP@COLING 2018](#))
- Reconnaissance automatique ([CICLING 2019; TALN 2019; soumission ACL 2020](#))
- Extension des études sur anglais-chinois ([soumission LREC 2020](#))
- Validations dans d'autres cadres de recherche ([EIAH 2019](#))

À court terme

Étudier les procédés de traduction dans des corpus littéraires

Améliorer la classification parmi les procédés non littéraux

Transférer des traits pour le couple anglais-chinois

Développer l'outil d'aide à la compréhension écrite

À court terme

Étudier les procédés de traduction dans des corpus littéraires

Améliorer la classification parmi les procédés non littéraux

Transférer des traits pour le couple anglais-chinois

Développer l'outil d'aide à la compréhension écrite

À plus long terme

Alléger le besoin de l'annotation manuelle

Déetecter les frontières de traduction non littérale

Contrôler sémantiquement la méthode par pivot (paraphrase)

Merci pour votre attention !

Traits sensibles au contexte

- Plongements lexicaux d'ELMo anglais et français, ou la moyenne des plongements pour représenter des segments
- Similarité cosinus entre des plongements d'ELMo bilingues
- Plongements de phrases de contexte anglais et français, où le mot ou le segment en question est remplacé par un blanc
- Des représentations de mot masquées basées sur Context2Vec et MUSE, suivant le travail de [Vyas et Carpuat \(2017\)](#)

Exemples de confusion d'annotation

Littéral vs Équivalence :

unrest → *des troubles, fueled* → *a alimenté*

Littéral vs Généralisation :

that service → *lesquels*

Littéral vs Transposition :

energy farming → *fermes à énergie*

Transposition vs Mod+Trans :

reproductive health services → *le contrôle des naissances*

Particularisation vs Mod+Trans :

there's a lot that (has to come together)

→ *beaucoup de choses (doivent se produire en même temps)*

F-mesures moyennes pour les classes non littérales

équivalence	généralisation	particularisation	modulation	contient-transposition
0,51 ± 0,02	0,25 ± 0,09	0,56 ± 0,05	0,36 ± 0,08	0,68 ± 0,02

Tableau 6.5 – F-mesures moyennes sur les cinq plis pour chaque procédé non littéral

Expérience avec l'ingénierie des traits

Pour trouver les meilleurs hyperparamètres, 10% de données sont séparées comme test, et une validation croisée à trois plis est exécutée sur 90% de données d'entraînement.

Ensuite ces hyperparamètres ont été utilisés avec la validation croisée sur l'ensemble des données.

Les données de test séparées : en cours de préparation (il faut annoter en 3 passes).

Tailles différentes de la ressource PPDB

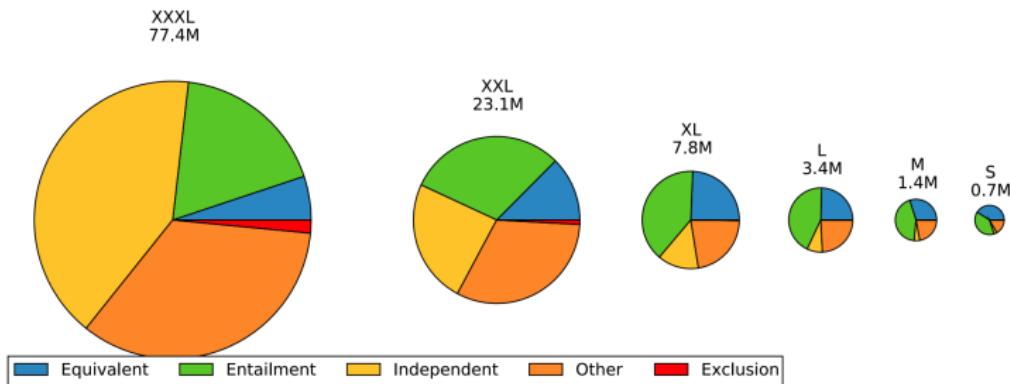


Figure 2: Distribution of entailment relations in different sizes of PPDB. Distributions are estimated from our manual annotations of randomly sampled pairs. PPDB-XXXL contains over 77MM paraphrase pairs (where the majority type is independent), compared to only 700K in PPDB-S (where the majority type is equivalent).

Pondération lexicale bi-directionnelle sur les mots pleins

(Koehn et al., 2003)

Exemple :

are a better match to → *correspondent mieux à*

$$\text{lex}(e|f, A) = \prod_{i=1}^{\text{length}(e)} \frac{1}{|\{j|(i,j) \in A\}|} \sum_{\forall(i,j) \in A} w(e_i|f_j)$$

e : segment source

f : segment cible

A : l'ensemble des mots alignés

$w(e_i|f_j)$: probabilité de traduction lexicale

F1 moyenne pondérée

'micro F1': *Calculate metrics globally by counting the total true positives, false negatives and false positives.*

'macro F1': *Calculate metrics for each label, and find their unweighted mean. This does not take label imbalance into account.*

'weighted': Calculer les métriques pour chaque étiquette et trouver leur moyenne, pondérée par le support (le nombre d'instances vraies pour chaque étiquette). **Cela modifie la "macro" pour tenir compte du déséquilibre de l'étiquette**; il peut en résulter un score F1 qui n'est pas compris entre la précision et le rappel.

Étude de contrôle : deux annotations indépendantes

100 paires de phrases (3 055 tokens EN et 3 238 tokens FR)

- **Frontière stricte** (les segments comparés ont la même frontière)
- **Frontière flexible** (la partie différente est *Littéral*) :

A1: *still pursue* → continue à rechercher

A2: *still* pursue → continue à rechercher

- **Les cas exclus** (frontière différente, et catégories ∈ non-littéral) :

A1: *but what's interesting is* → *par contre , ... est intéressante*

A2: *but what's interesting is* → *par contre , ... est intéressante*

	κ	%EN tokens
frontière stricte	0,67	72,60%
frontière flexible	0,62	85,56%

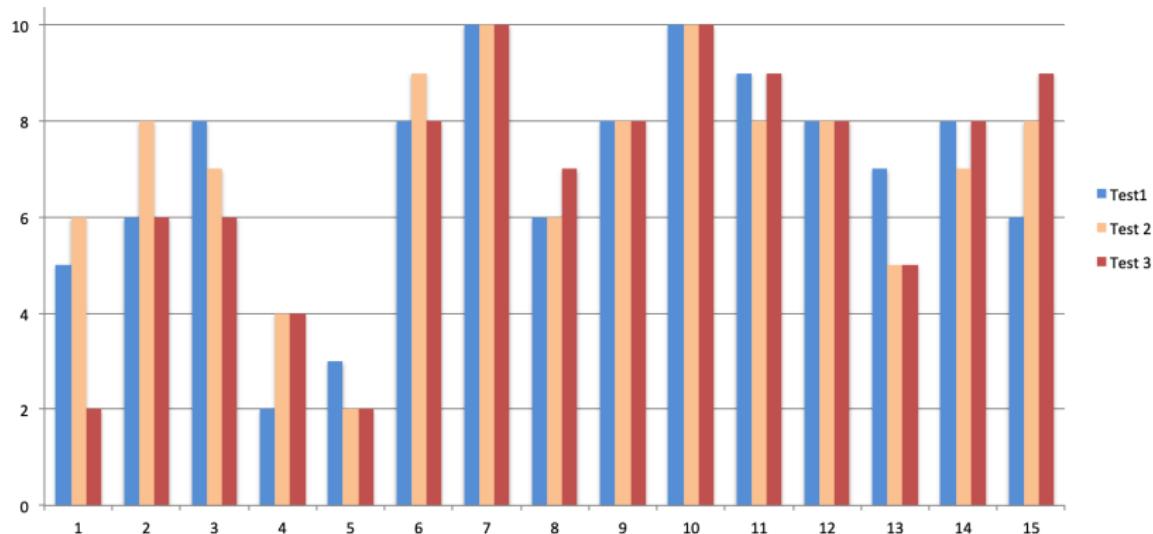
Contribution des traits (classification binaire 1:1)

Trait	F1 Moyenne (Littéral)	F1 Moyenne (Non littéral)
pos_vecteur_comptage	0,78	0,75
posCosinus_tous_les_mots	0,69	0,69
posCosinus_mots_pleins	0,68	0,64
pos_changement_patron	0,70	0,34
distance_Levenshtein	0,75	0,75
(ratio)_longueur_token	0,74	0,71
analyse_constituant	0,66	0,50
analyse_dépendance_interne	0,76	0,73
analyse_dépendance_externe	0,61	0,62
ConceptNet_Embedding	0,73	0,73
ConceptNet_lien	0,70	0,78
ConceptNet_pourcentage_indirect	0,70	0,62
différence_probabilité_traduction	0,77	0,78
entropie_traduction	0,60	0,62
pondération_lexicale	0,78	0,80
analyse_PoS	0,78	0,75
surface	0,72	0,70
analyse_syntaxique	0,76	0,76
ressource_ConceptNet	0,77	0,78
alignement_de_mot	0,84	0,85
tous les traits	0,87	0,87
tous les traits - les traits en vert	0,87	0,88

Contribution des traits (5 classes non littérales)

Trait	F1(micro-av)	F1(macro-av)	F1 (E)	F1 (G)	F1 (P)	F1 (M)	F1 (T)
pos_vecteur_comptage	0,52	0,44	0,51	0,16	0,54	0,34	0,65
posCosinus_tous_les_mots	0,39	0,29	0,25	0,00	0,45	0,23	0,54
posCosinus_mots_pleins	0,39	0,25	0,06	0,00	0,42	0,19	0,58
pos_changement_patron	0,37	0,20	0,43	0,00	0,00	0,00	0,56
distance_Levenshtein	0,36	0,26	0,32	0,00	0,31	0,18	0,49
(ratio)_longueur_token	0,39	0,34	0,36	0,21	0,39	0,26	0,49
analyse_constituant	0,36	0,19	0,00	0,00	0,41	0,00	0,56
analyse_dépendance_interne	0,45	0,39	0,37	0,21	0,47	0,31	0,59
analyse_dépendance_externe	0,35	0,26	0,27	0,05	0,33	0,15	0,51
ConceptNet_EMBEDDING	0,32	0,27	0,28	0,02	0,38	0,25	0,40
ConceptNet_lien	0,32	0,18	0,13	0,00	0,30	0,00	0,46
ConceptNet_pourcentage_indirect	0,29	0,16	0,22	0,00	0,00	0,18	0,42
différence_probabilité_traduction	0,38	0,32	0,35	0,13	0,38	0,26	0,50
entropie_traduction	0,35	0,27	0,36	0,00	0,39	0,15	0,44
pondération_lexicale	0,32	0,24	0,36	0,02	0,28	0,15	0,41
analyse_PoS	0,51	0,43	0,48	0,15	0,52	0,34	0,64
surface	0,38	0,34	0,36	0,23	0,37	0,27	0,47
analyse_syntaxique	0,48	0,40	0,38	0,18	0,52	0,30	0,63
ressource_ConceptNet	0,34	0,28	0,24	0,02	0,46	0,23	0,44
alignement_de_mot	0,45	0,38	0,44	0,15	0,51	0,26	0,54
pos + surface + syntaxique	0,54	0,47	0,49	0,28	0,54	0,38	0,67
tous - ConceptNet	0,54	0,47	0,51	0,27	0,54	0,37	0,67
tous les traits	0,55	0,47	0,50	0,25	0,55	0,37	0,67

Expérience en compréhension écrite (chinois, niveau A2)



Expérience en compréhension écrite (chinois, niveau B2)

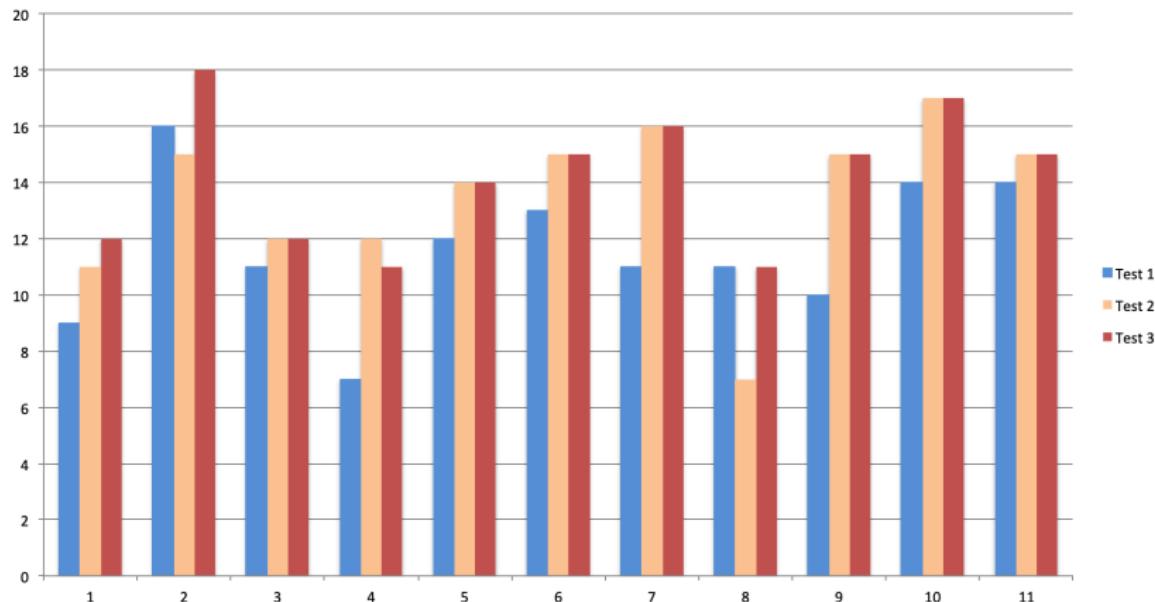
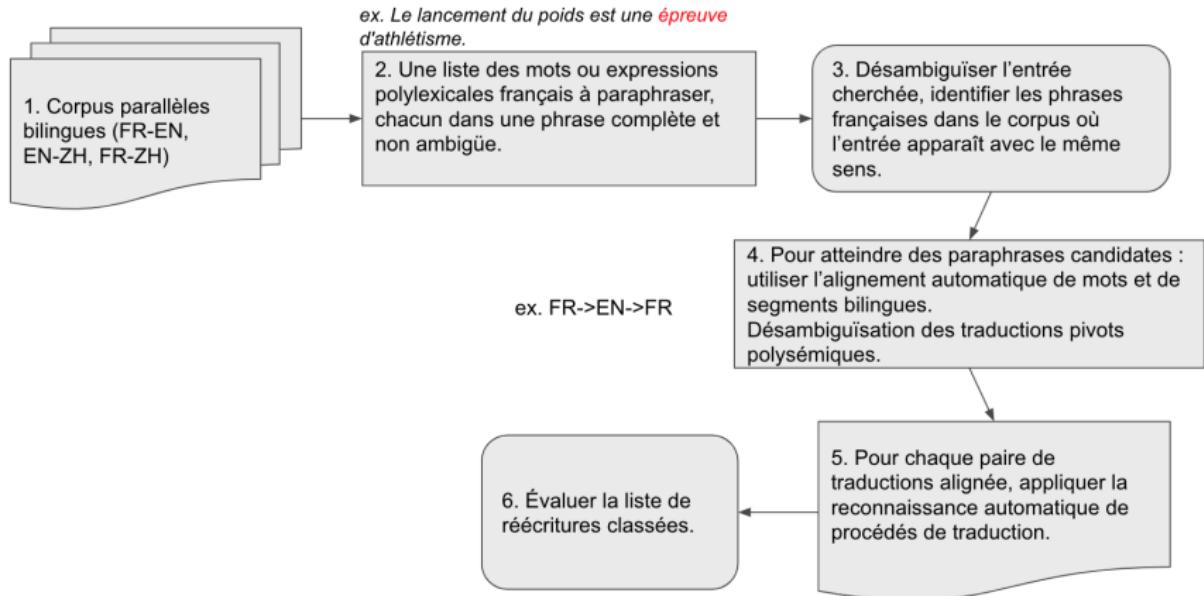
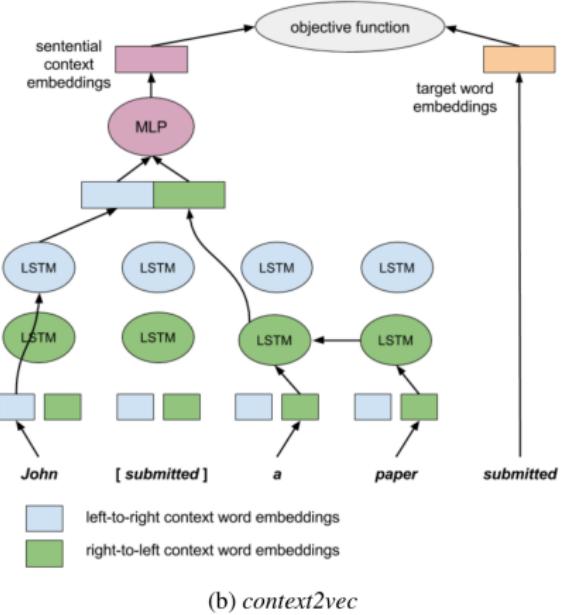
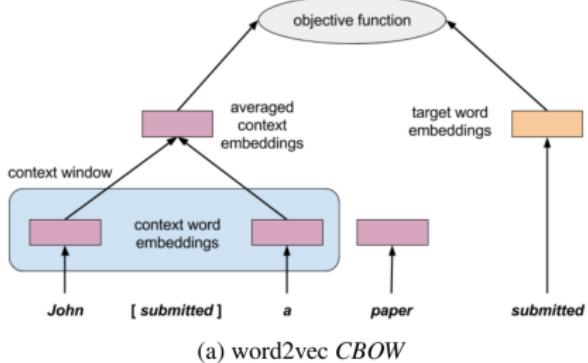


Schéma de travail pour le développement de l'outil



Architecture de Context2Vec



(Melamud et al., 2016)

Architecture de ELMo

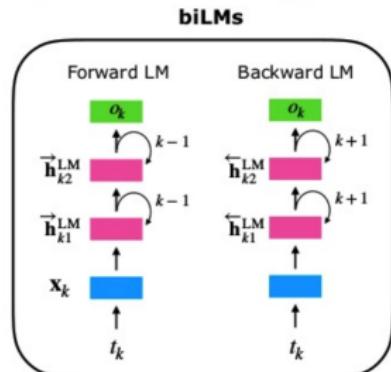


ELMo is a task specific representation. A down-stream task learns weighting parameters

$$\text{ELMo}_k^{\text{task}} = \gamma^{\text{task}} \times \sum \left\{ \begin{array}{l} s_2^{\text{task}} \times h_{k2}^{\text{LM}} \\ s_1^{\text{task}} \times h_{k1}^{\text{LM}} \\ s_0^{\text{task}} \times h_{k0}^{\text{LM}} \\ (\{x_k; x_k\}) \end{array} \right. \times \begin{array}{c} \text{Concatenate hidden layers} \\ \xrightarrow{\quad [\vec{h}_{kj}^{\text{LM}}; \hat{h}_{kj}^{\text{LM}}] \quad} \end{array}$$

Unlike usual word embeddings, ELMo is assigned to every *token* instead of a type

ELMo represents a word t_k as a linear combination of corresponding hidden layers (inc. its embedding)



(Peters et al., 2018)