

MODÈLES DE TRADUCTION ÉVOLUTIFS

Frédéric BLAIN

23 Septembre 2013

Directeur : Holger SCHWENK
Co-encadrant : Jean SENELLART

Rapporteurs : Laurent BESACIER
Marc DYMETMAN

Examinateurs : Yannick ESTÈVE
Patrik LAMBERT



CONTEXTE SCIENTIFIQUE

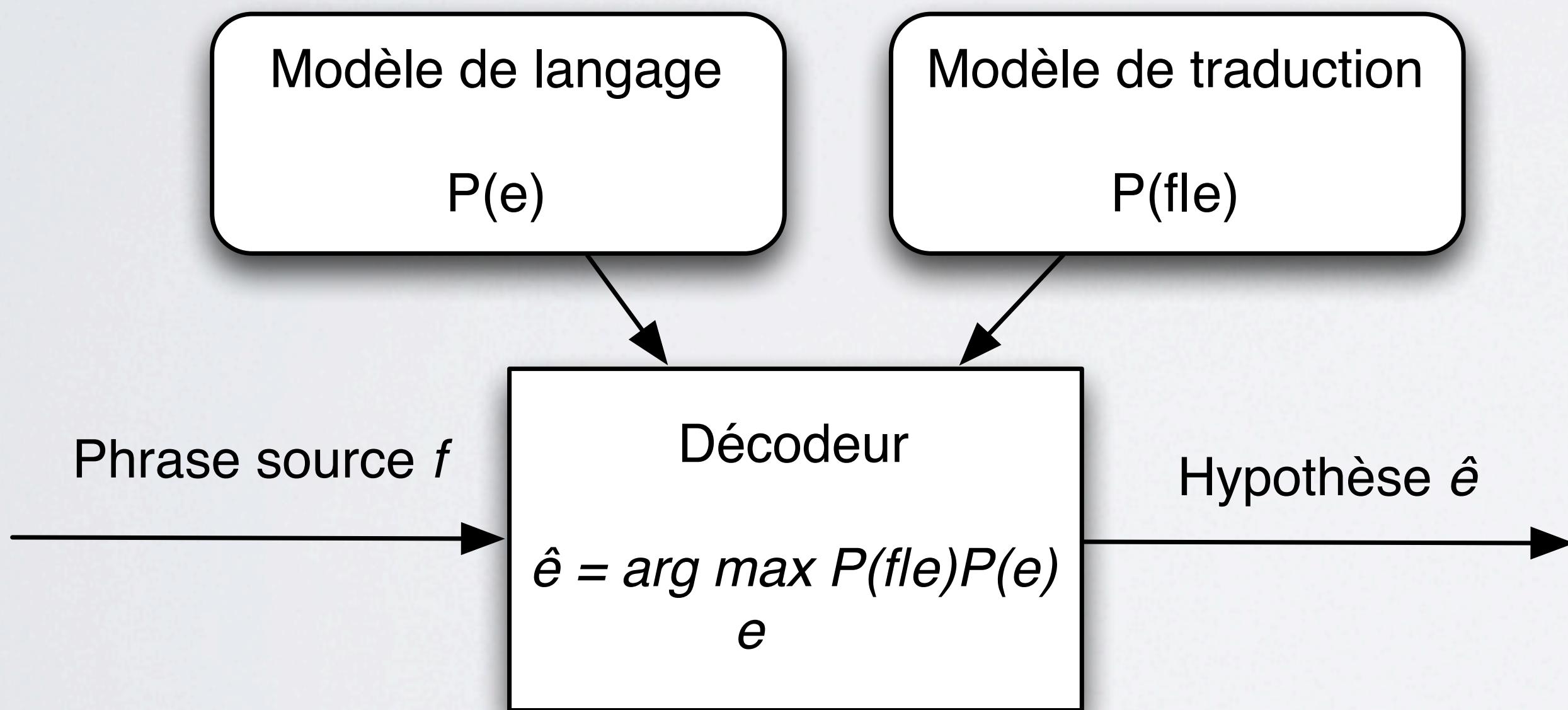
Thèse CIFRE-Défense



INTRODUCTION

Technologie de traduction automatique (TA)

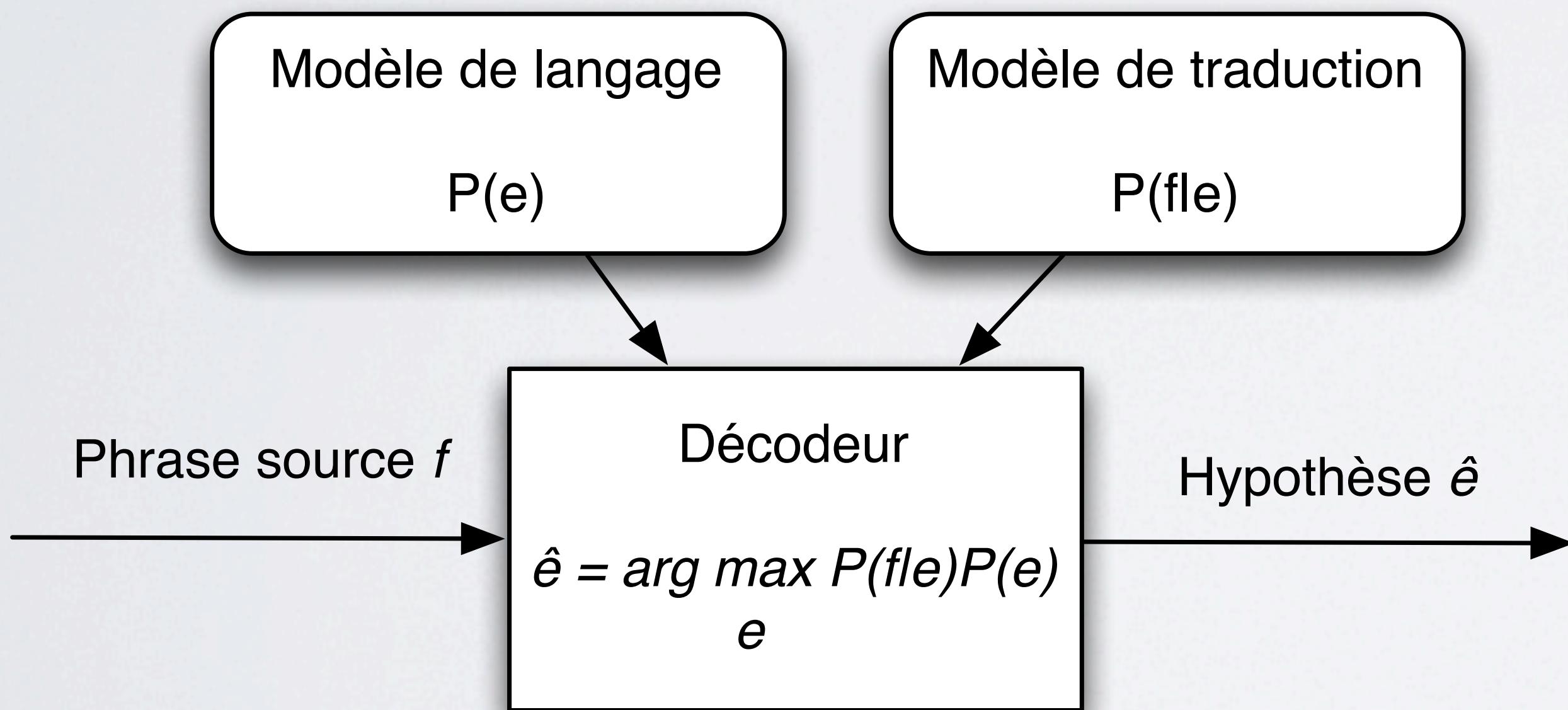
- ▶ Système de traduction automatique statistique (TAS) (EN>FR)



INTRODUCTION

Technologie de traduction automatique (TA)

- ▶ Système de traduction automatique statistique (TAS) (EN>FR)



Ressources en volume considérable...

Coût computationnel important...

Modèles statiques !

INTRODUCTION

Objectif

Développement de nouvelles méthodes pour la création de systèmes de traduction automatique **évolutifs**.

Axes de recherche

- ▶ Évolution du système de traduction liée aux retours utilisateurs
- ▶ Évolution dans le temps

INTRODUCTION

Cadre applicatif de la thèse

COSMAT - Service collaboratif de traduction de contenus scientifiques

Autres projets de recherche existants

TRACE - Traduction Robuste par Analyse et Correction d'Erreurs

FAUST - Feedback Analysis for User adaptive Statistical Translation

MATECAT - MT Enhanced Computer Assisted Translation

PLAN

La Post-Édition

Analyse qualitative de données post-éditées

Adaptation incrémentale d'un système de T.A.S

-- COSMAT : Service collaboratif de traduction de contenus scientifiques

Conclusions et perspectives futures

LA POST-ÉDITION

De quoi parle-t-on ?

LA POST-ÉDITION (PE)

Définition

« Processus visant à améliorer une hypothèse de traduction en sortie d'un système de traduction de manière à la **rendre publiable**, et ce avec un **minimum d'effort**. »

- ▶ De plus en plus utilisée dans l'industrie de la traduction
- ▶ Domaine de recherche des plus actifs actuellement

LA POST-ÉDITION

Ne pas confondre avec...

la « Post-Édition Statistique » (PES)

[Simard 2007], [Dugast 2007]

- ▶ Améliore la qualité en sortie du système de TA (meilleure cohérence syntaxique)
- ▶ L'hypothèse de traduction doit parfois tout de même être révisée...

LA POST-ÉDITION

Évaluer l'effort de PE, plusieurs approches...

- ▶ L'effort NE DOIT PAS être plus important que celui d'effectuer une traduction complète !

Critère n°1 : le temps ! [Specia, 2011]

- Consignes de post-édition (complète, légère)

Critère n°2 : l'utilisateur

- Activité au clavier (“Key stroke”) [Barrett et al., 2001]
- Comportement visuel devant la tâche (“Eye tracking”) [Doherty et al., 2010]

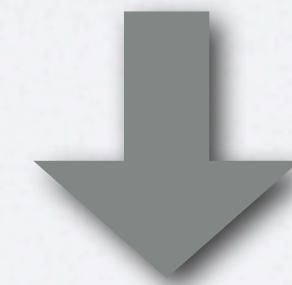
LA POST-ÉDITION

Limiter, voire réduire cet effort...

« la familiarité avec les motifs d'erreurs produites par un système de TA particulier est un facteur important dans la réduction du temps de post-édition »

[Martinez, 2003]

- ▶ La post-édition est considérée comme une classification d'erreurs de traduction



Appliquer un ensemble de règles de post-édition [Guzmàn, 2007] ?

LA POST-ÉDITION

Problématique

**Le système de TA ne bénéficie pas des retours utilisateurs...
il reproduira donc les mêmes erreurs !**

Propositions faites dans le cadre de cette thèse :

1. Analyser qualitativement des données post-éditées humainement

2. Adapter incrémentalement un système de TAS

LA POST-ÉDITION

Problématique

**Le système de TA ne bénéficie pas des retours utilisateurs...
il reproduira donc les mêmes erreurs !**

Propositions faites dans le cadre de cette thèse :

1. Analyser qualitativement des données post-éditées humainement
 - Qu'est-il important de retenir de ces révisions ?
2. Adapter incrémentalement un système de TAS

LA POST-ÉDITION

Problématique

**Le système de TA ne bénéficie pas des retours utilisateurs...
il reproduira donc les mêmes erreurs !**

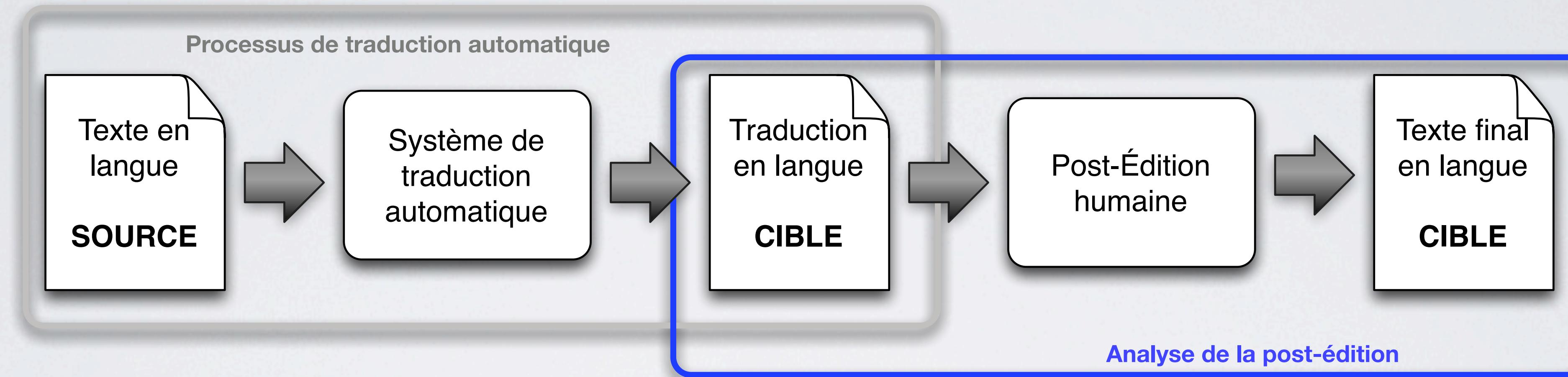
Propositions faites dans le cadre de cette thèse :

1. Analyser qualitativement des données post-éditées humainement
 - Qu'est-il important de retenir de ces révisions ?
2. Adapter incrémentalement un système de TAS
 - Comment peut-on bénéficier de ces nouvelles connaissances ?

ANALYSE QUALITATIVE DE DONNÉES POST-ÉDITÉES

Qu'est-il important de retenir des révisions du post-éditeur ?

ANALYSER LES RETOURS UTILISATEURS



Méthode proposée

- ▶ Comparaison entre une traduction et sa version post-éditée

ANALYSER LES RETOURS UTILISATEURS

Comment catégoriser les éditions en post-édition ?

- ▶ Annoter manuellement un ensemble réel de données post-éditées humainement
 - Données clients de Systran (Symantec, Autodesk)
 - Deux systèmes ENFR : à base de règles (RBMT), statistique (SMT)
- ▶ Définir une typologie de la post-édition
 - Inspirée de la classification d'erreurs de traduction
 - Ex: [Font-Llitjòs et al., 2005], [Vilar et al., 2006], [Dugast et al., 2007]
 - Adaptée au français

ANALYSER LES RETOURS UTILISATEURS

Première observation...

► **toutes les éditions ne concernent pas des erreurs du système**

Deux « niveaux » d'éditions :

- 1er niveau

=> Éditions liées à une erreur dans l'hypothèse de traduction

- 2nd niveau

=> Éditions induites par une édition de niveau 1 (i.e. des propagations)

Ex : accord en genre et en nombre du déterminant / adjectif / verbe associé à un nom (français)

ANALYSER LES RETOURS UTILISATEURS

EXEMPLE

Source

By default , the border is displayed .

Traduction

Par défaut , le bord est affiché .

**TER
37,5%**

Traduction post-éditée

Par défaut , la bordure est affichée .

ANALYSER LES RETOURS UTILISATEURS

EXEMPLE

Source

By default , the border is displayed .

Traduction

Par défaut , le bord est affiché .

**TER
12,5%**

Traduction post-éditée

Par défaut , la bordure est affichée .

ANALYSER LES RETOURS UTILISATEURS

Comment modéliser ces deux niveaux d'éditions ?

- ▶ Extraire automatiquement un ensemble d'éditions **minimales** et **logiques**
 - minimale => la plus petite édition indépendante réalisée par le post-éditeur
 - logique => par opposition aux éditions dites « mécaniques » en distance d'édition
(i.e. le nombre d'éditions correspond au nombre de différences observées)
- ▶ Objectif : correspondre davantage à l'intention du post-éditeur
 - Évaluation plus précise de la qualité en sortie d'un système de TA

ANALYSER LES RETOURS UTILISATEURS

Comment modéliser ces deux niveaux d'éditions ?

- ▶ Extraire automatiquement un ensemble d'éditions **minimales** et **logiques**
 - minimale => la plus petite édition indépendante réalisée par le post-éditeur
 - logique => par opposition aux éditions dites « mécaniques » en distance d'édition
(i.e. le nombre d'éditions correspond au nombre de différences observées)
- ▶ Objectif : correspondre davantage à l'intention du post-éditeur
 - Évaluation plus précise de la qualité en sortie d'un système de TA

« Actions de Post-Édition »

ANALYSER LES RETOURS UTILISATEURS

Seconde observation...

► **la PE est une tâche (très) répétitive !**

De nombreuses redondances dans les éditions observées

- Apprentissage possible dès le premier exemple
- Perspective de réduction importante de la tâche du post-éditeur

► Certaines classes d'éditions sont plus couramment observées que d'autres...

LES « ACTIONS DE POST-ÉDITION » (APE)

Typologie de la post-édition en APE (pour le français)

- Classes les plus couramment observées :
 - Groupe Nominal (GN) : *changement de déterminant, changement de sens nominal, accord en nombre, modification de la casse, choix d'adjectif*
 - Groupe Verbal (GV) : *accord grammatical, choix du sens verbal*
 - Changements de préposition
 - Changements de co-référence

RÉPÉTITIVITÉ DE LA TÂCHE

Classes les plus observées

lors de l'annotation manuelle

- Changements lexicaux : env. 90% des APE
 - Changements terminologiques : env. 60%
 - Changements de casse nombreux...
- => corpus de documentation technique

Classe	Sous-classe	Système RBMT		Système de TAS	
		#APE	%APE	#APE	%APE
Groupe Nominal (GN)					
<i>Choix du déterminant</i>	1	1.2%		3	2.2%
<i>Choix du sens du mot</i>	49	59%		84	62%
<i>Accord en nombre</i>	3	3.6%		0	0%
<i>Changement de casse</i>	19	23%		37	27%
<i>Changement adjectif</i>	2	2.4%		1	0.7%
Total	74	90%		125	92%
Groupe Verbal (GV)					
<i>Accord grammatical</i>	3	3.6%		2	1.5%
<i>Choix verbal</i>	3	3.6%		2	1.5%
Total	6	7.2%		4	3%
Changement de préposition	1	1.2%		0	0%
Changement de co-reference	2	2.4%		7	5%
TOTAL	83	100%		136	100%

RÉPÉTITIVITÉ DE LA TÂCHE

Impact des APE les plus fréquentes

Système RBMT				
<i>avant</i>	<i>après</i>	#occ.	%	
famille	usine	96	20%	
sol	atelier	65	13%	
plancher	sol	11	2%	
archive	actif	9	2%	
Total (top-4)		181	37%	
TOTAL (toutes)		488	100%	

Système de TAS				
<i>avant</i>	<i>après</i>	#occ.	%	
archive	actif	60	11%	
superposition	calque	39	7%	
archive	ressource	19	3%	
sol	atelier	13	2%	
Total (top-4)		131	23%	
TOTAL (toutes)		558	100%	

RÉPÉTITIVITÉ DE LA TÂCHE

Impact des APE les plus fréquentes

Système RBMT			
<i>avant</i>	<i>après</i>	#occ.	%
famille	usine	96	20%
sol	atelier	65	13%
plancher	sol	11	2%
archive	actif	9	2%
Total (top-4)		181	37%
TOTAL (toutes)	488	100%	

Système de TAS			
<i>avant</i>	<i>après</i>	#occ.	%
archive	actif	60	11%
superposition	calque	39	7%
archive	ressource	19	3%
sol	atelier	13	2%
Total (top-4)		131	23%
TOTAL (toutes)	558	100%	

RÉPÉTITIVITÉ DE LA TÂCHE

Impact des APE les plus fréquentes

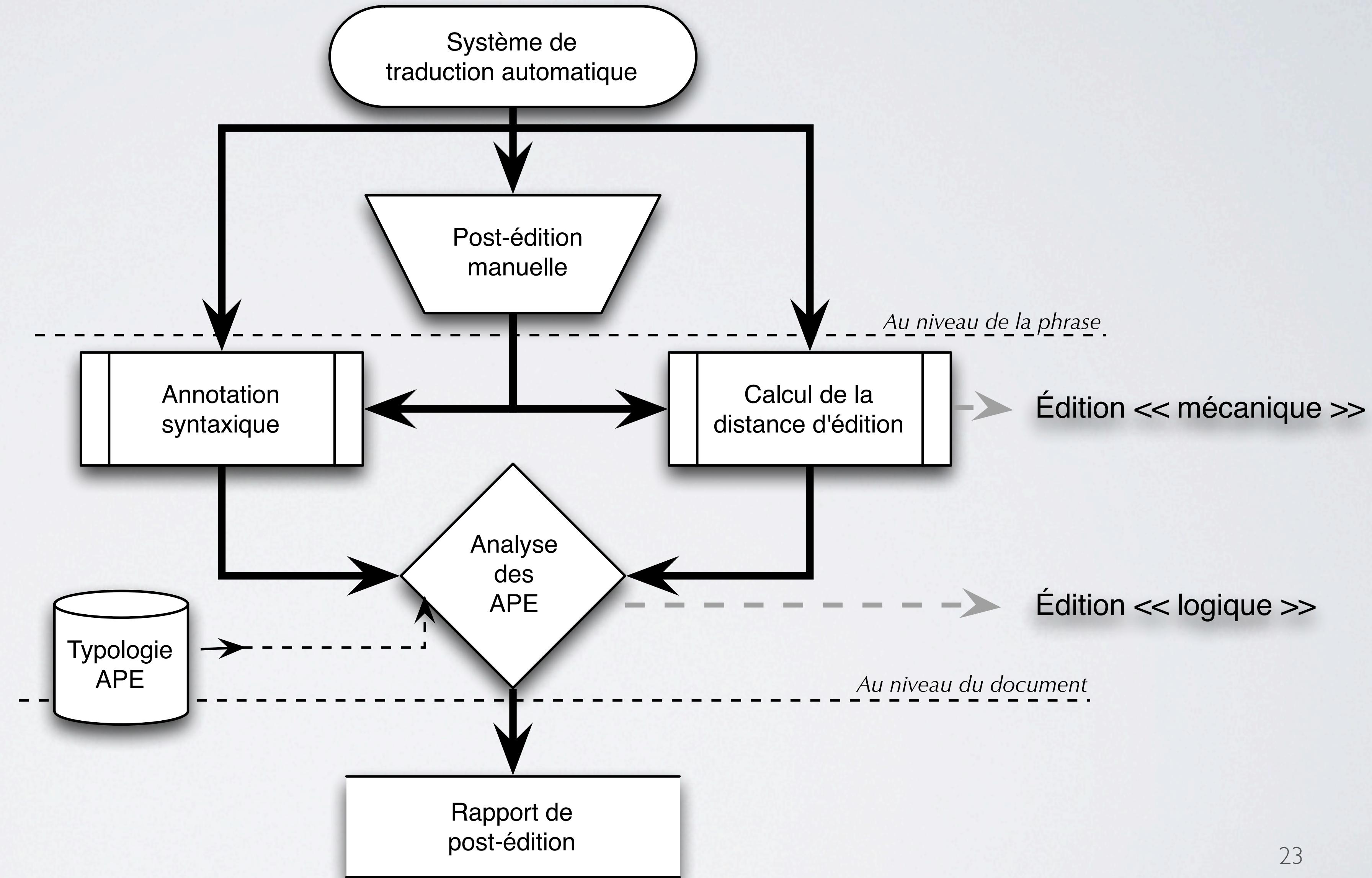
Système RBMT			
<i>avant</i>	<i>après</i>	#occ.	%
famille	usine	96	20%
sol	atelier	65	13%
plancher	sol	11	2%
archive	actif	9	2%
Total (top-4)	181	37%	
TOTAL (toutes)	488	100%	

Système de TAS			
<i>avant</i>	<i>après</i>	#occ.	%
archive	actif	60	11%
superposition	calque	39	7%
archive	ressource	19	3%
sol	atelier	13	2%
Total (top-4)	131	23%	
TOTAL (toutes)	558	100%	

OUTIL D'ANALYSE AUTOMATIQUE EN APE

« SmartDiff »

- Annotation syntaxique
=> Systran
- Distance d'édition
=> Translation Error Rate
- Analyse en APE
=> classifieur par règles



RÉSULTATS EXPÉIMENTAUX

Analyse automatique vs. Analyse manuelle (référence)

Classe	<i>Sous-classe</i>	Système RBMT				Système de TAS			
		#APE	#Match	%Prec.	%Rapp.	#APE	#Match	%Prec.	%Rapp.
Grp Nominal (GN)									
<i>Choix du déterminant</i>	15	1	7%	100%		16	1	6%	33%
<i>Chgt. de sens nominal</i>	89	35	40%	71%		97	69	71%	82%
<i>Changement du nombre</i>	3	0	0	0		4	0	0	0
<i>Changement de la casse</i>	18	12	67%	63%		27	25	93%	68%
<i>Choix d'adjectif</i>	0	0	0	0		1	0	0	0
Total	125	48	—	—		145	95	—	—

Grp Verbal (GV)

<i>Accord grammatical</i>	1	0	0	0		2	0	0	0
<i>Chgt. de sens verbal</i>	8	2	25%	67%		6	2	33%	100%
Total	9	2	—	—		8	2	—	—

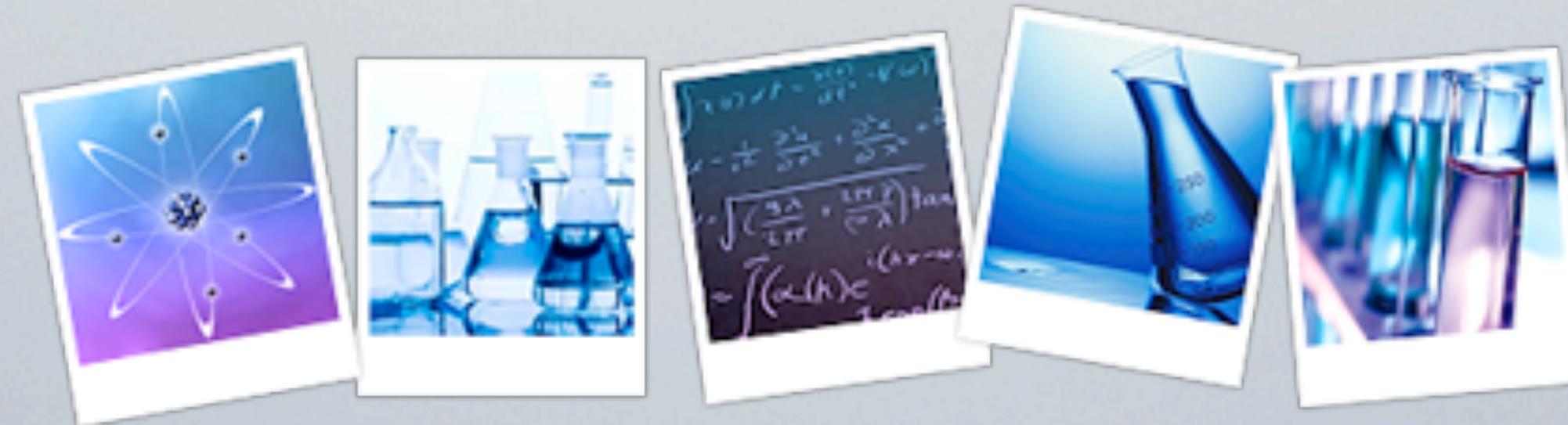
RÉSULTATS EXPÉIMENTAUX

Taux de couverture des APE observées

	Système RBMT		Système de TAS	
	#occ.	%couv.	#occ.	%couv.
Nombre d'éditions	3231	100%	3947	100%
Nombre d'APE	1133	35%	1340	34%
Nombre de propagation	169	5,2%	255	6,5%
Nombre de déterminant	40	1,2%	99	2,5%
Nombre de préposition	102	3,2%	97	2,5%
Nombre de verbe	27	0,8%	59	1,5%

ADAPTATION INCRÉMENTALE D'UN SYSTÈME DE T.A.S.

Comment bénéficier rapidement de nouvelles ressources ?



COSMAT

Service collaboratif de traduction automatique de contenus scientifiques

TRADUCTION DE CONTENUS SCIENTIFIQUES

Enjeux

Offrir la possibilité de faire **traduire** gratuitement **les articles scientifiques** du serveur « **HAL** » entre le **français et l'anglais**.

Différents acteurs identifiés

- public francophone : accéder à des ressources en anglais
- scientifiques français : traduire leurs travaux vers l'anglais pour communiquer
- scientifiques internationaux : accéder à des ressources francophones

UN SERVICE DE TRADUCTION...

Intégré dans la plateforme en ligne « HAL »

- 234k références au format PDF
- 30 domaines scientifiques => Physique (17,48 %), Informatique (20,88 %)

Associé à un outil de révision des traductions pour les utilisateurs

- Interface de post-édition développée par Systran
=> Doit permettre de collecter un corpus de données post-éditées

Par et pour les utilisateurs

- Adaptation incrémentale du service à partir des retours utilisateurs

APPROCHES SCIENTIFIQUES ET TECHNIQUES

Extraction de contenus structurés

- Format PDF => Grobid (code libéré durant COSMAT)

Moteurs de traduction

- Tâche : ENFR, FREN
- Hybrides, développés par le **LIUM** et Systran
 - Modules linguistiques développés par Systran
 - Reconnaissance d'entités scientifiques : formules, références, tableaux, etc.
 - Terminologie du domaine
 - Ressources bilingues et monolingues en domaine

ADAPTATION AU DOMAINE DU SYSTÈME

Extraction de données bilingues en domaine

- Références scientifiques de HAL majoritairement monolingues
- Exception : les thèses françaises avec un résumé en français ET en anglais
- Pas obligatoirement une traduction l'un de l'autre
- Filtrage par un alignement au niveau de la phrase (IBM | [Brown 1993])

► **Création d'un corpus scientifique unique : « Corpus COSMAT »**

=> Librement distribué : JHU SMT workshop 2012, projet européen « TransLectures »

ACTUALITÉS

Lancement officiel à l'automne 2013

- Sortie officielle prévue de HAL v3
- Collecte des premières données post-éditées...

LITTÉRATURE

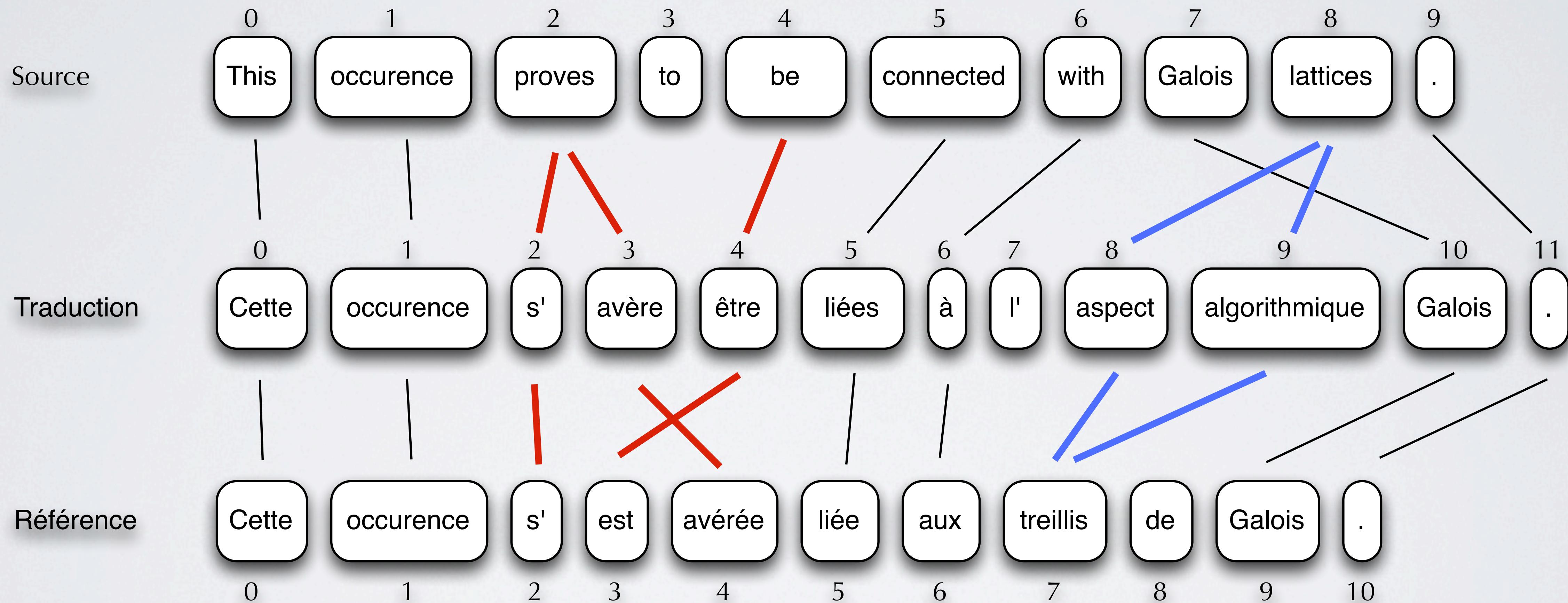
- ▶ Adaptation incrémentale par post-édition simulée [Hardt et al., 2010]
 - Utilisation de l'outil d'alignement des mots Giza
 - « Pour être pratique, l'adaptation incrémentale doit être réalisable en moins d'une seconde. »
- ▶ Stream-based translation model [Lavenberg et al., 2010]
 - Algorithme EM a besoin de beaucoup de données => pas adapté dans notre contexte industriel où l'on veut bénéficier de retours utilisateurs rapidement (au niveau du paragraphe par ex..)

Proposition

Extraire de nouvelles informations à partir de données post-éditées
sans employer l'outil d'alignement des mots Giza.

ALIGNEMENT DES MOTS SOURCE-RÉFÉRENCE

Exemple



PROTOCOLE D'ADAPTATION INCRÉMENTALE

« OnlineAdapt »

I. Traduction

Alignment source-traduction

2. Post-édition (simulée)

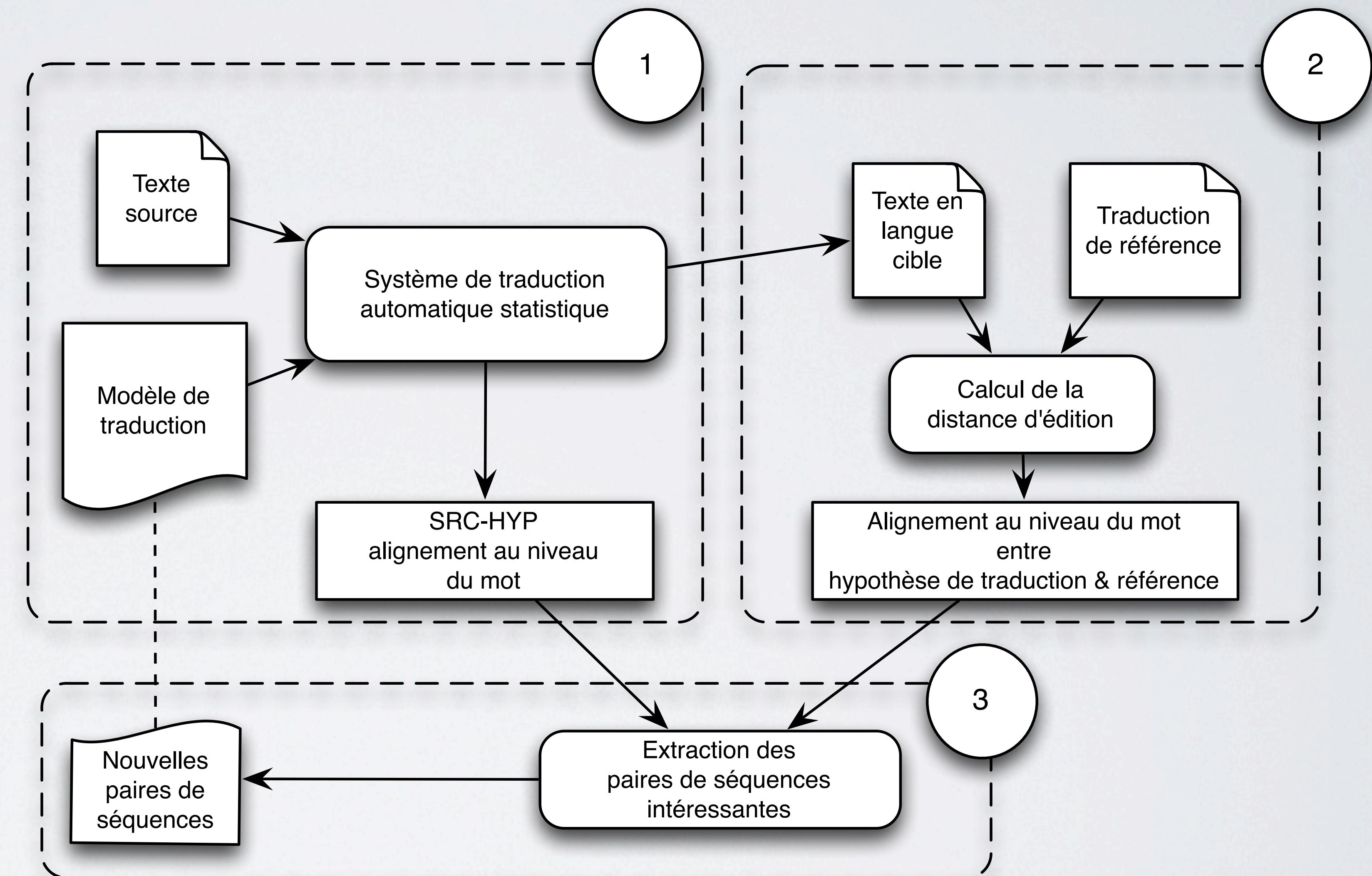
Distance d'édition

Alignment traduction-référence

3. Adaptation

Alignment source-référence

Extraction de nouveaux segments



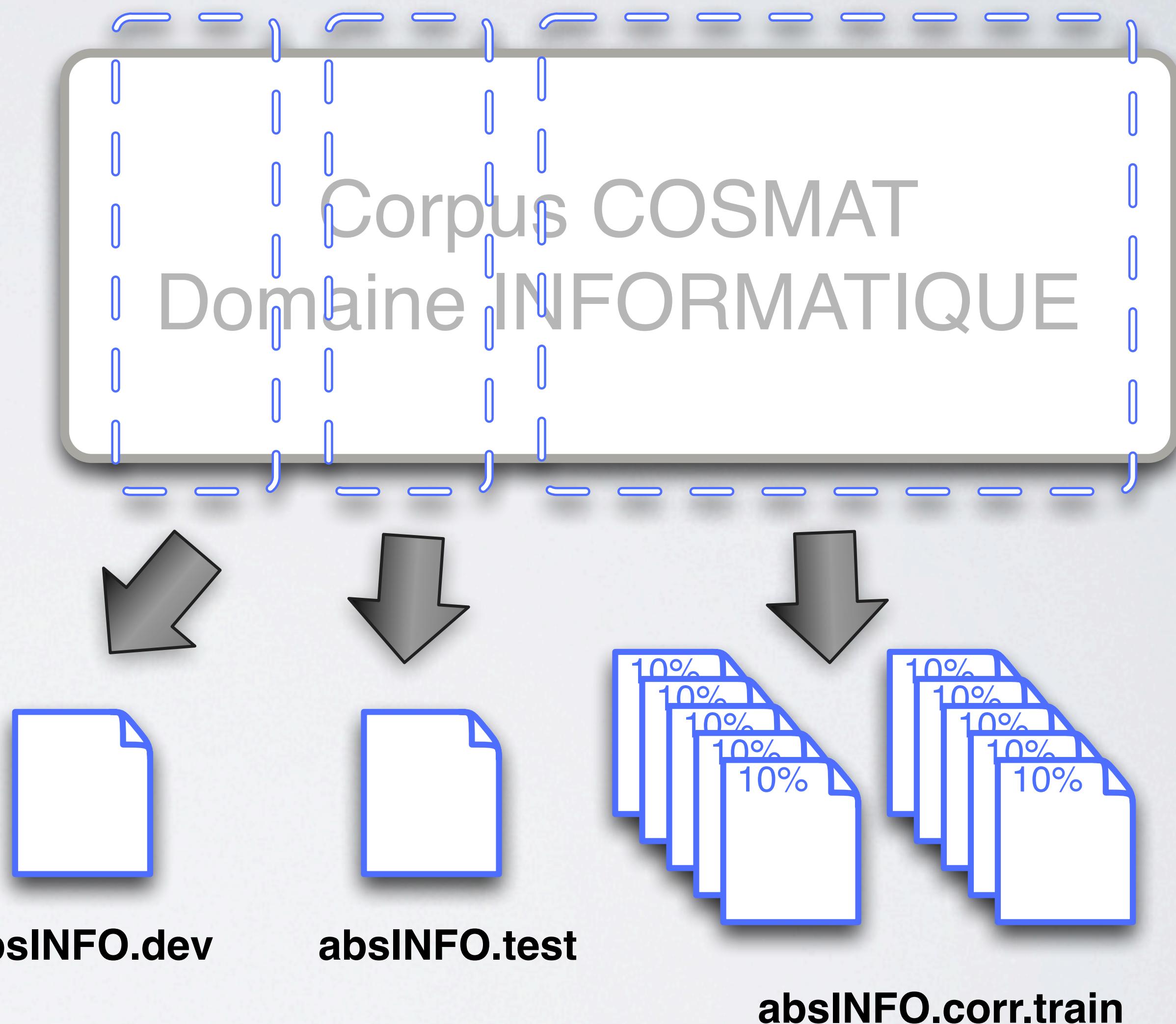
PROCESSUS EXPÉRIMENTAL

- ▶ Tâche : traduction ENFR de contenus scientifiques
 - Données expérimentales : Corpus COSMAT (résumés d'articles scientifiques)
 - Domaine : Informatique
- ▶ Système initial entraîné à partir d'alignements des mots réalisés avec Giza
 - Données hors domaine :
 - Europarl v7 (50M mots)
 - News Commentary v7 (2,8M mots)
 - Données en domaine :
 - Corpus COSMAT, domaine Informatique (absINFO - 500k mots)

SIMULATION DE POST-ÉDITION

AbsINFO (500k mots)

- AbsINFO.dev (75k mots)
- AbsINFO.test (75k mots)
 - Corpus de test utilisé pour valider qualitativement l'adaptation incrémentale
- AbsINFO.corr.train (350k mots)
 - Divisé en sous-corpus de 10% (35k mots) pour simuler la post-édition



PROCESSUS EXPÉRIMENTAL

Quatre systèmes ont été entraînés...

1. Giza

- Réentraînement complet pour chaque sous-corpus 10%, 20%, 30%, ... de absINFO.corr.train
- Considéré comme limite haute des résultats pour notre approche incrémentale
- Procédure **très** coûteuse en ressources computationnelles...

2. Giza incrémental

- Utilisation de la version incrémentale de Giza (état de l'art)
- Ajout incrémental de 10% de absINFO.corr.train à chaque itération

PROCESSUS EXPÉRIMENTAL

Quatre systèmes ont été entraînés...

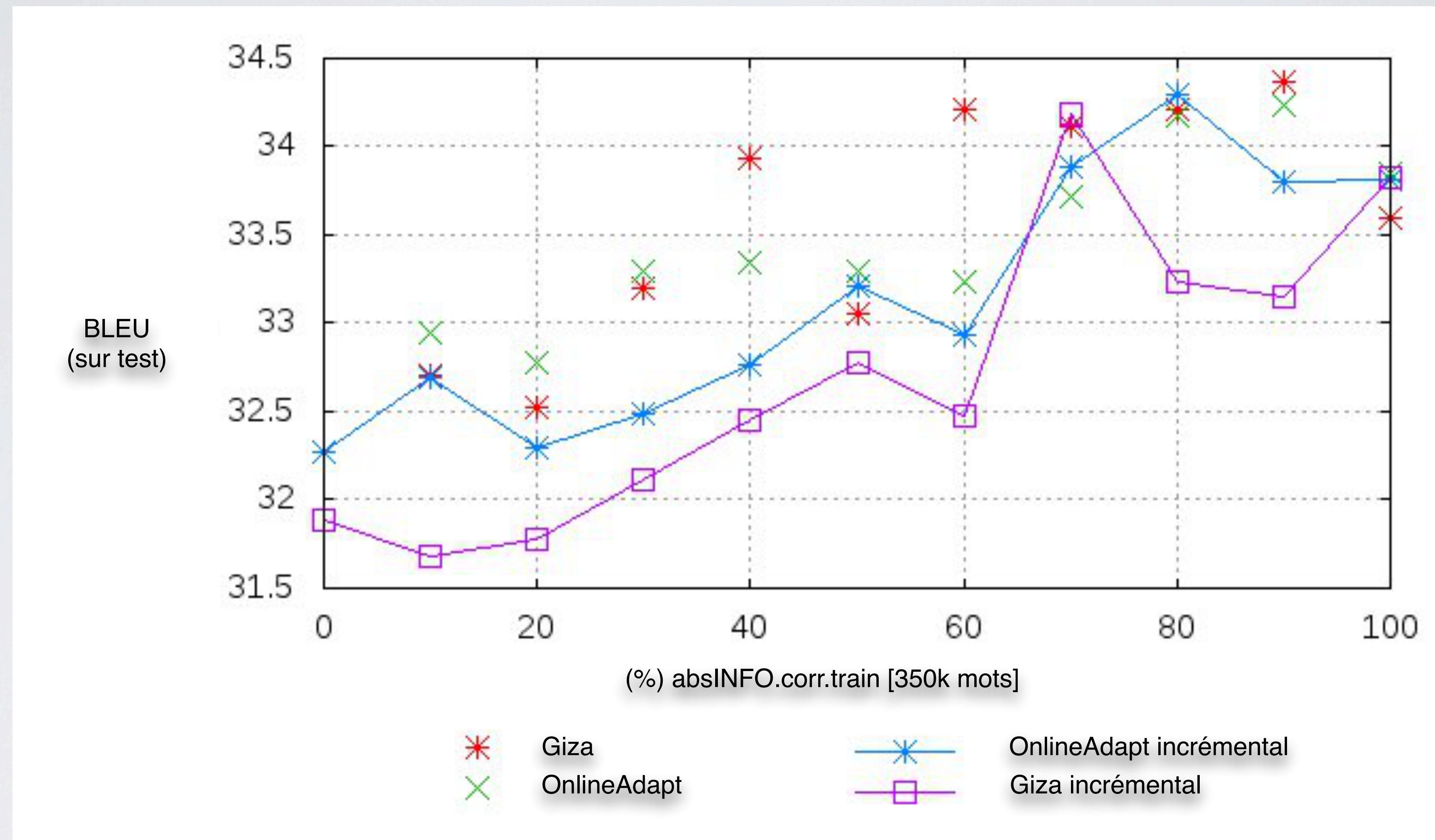
3. OnlineAdapt

- Comme « Giza » avec notre nouvel algorithme d'alignement source-référence
- Le système initial est toujours utilisé pour traduire les données additionnelles

4. OnlineAdapt incrémental

- Comme « Giza incrémental » avec notre nouvel algorithme d'alignement source-référence
- Les données additionnelles sont traduites avec le système entraîné à l'itération précédente

RÉSULTATS EXPÉIMENTAUX



FACTEUR « TEMPS »

« Inc-Giza »

- Extraction du (nouveau) vocabulaire
- Mise à jour des fichiers de cooccurrences (x2)
- Chargement des anciens modèles d'alignement + création des nouveaux alignements (x2)
- Extraction des nouveaux alignements

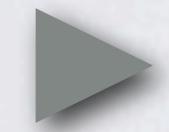
Alignment d'un sous-corpus de 35k mots en 14min (env.)

FACTEUR « TEMPS »

« OnlineAdapt »

- Alignements des mots source-traduction réalisés durant la traduction : coût nul
- Distance d'éditions traduction-référence avec TER : < 5sec pour 1,5k paires de phrases
- Alignements des mots source-référence : < 1sec pour 1,5k paires de phrases

**Alignment d'un sous-corpus de 35k mots
en quelques secondes !**

- 
- OnlineAdapt utilisable en temps-réel, ce qui n'est pas le cas de Giza-inc...

SYSTÈMES ENTRAINÉS

Peut-on réduire davantage le temps nécessaire ?

- ▶ Procédure d'extraction, d'évaluation des paires de séquences de mots, et d'apprentissage du nouveau modèle de traduction similaires aux deux approches (toolkit Moses)

Proposition : Utiliser plusieurs modèles de traduction

- ▶ 1 modèle générique + 1 modèle en domaine (enrichi à chaque itération)
 - Différentes combinaisons testées avec ces deux modèles de traduction

RÉSULTATS EXPÉIMENTAUX

Modèles de traduction

► repli vs. repli-inversé

- Repli :

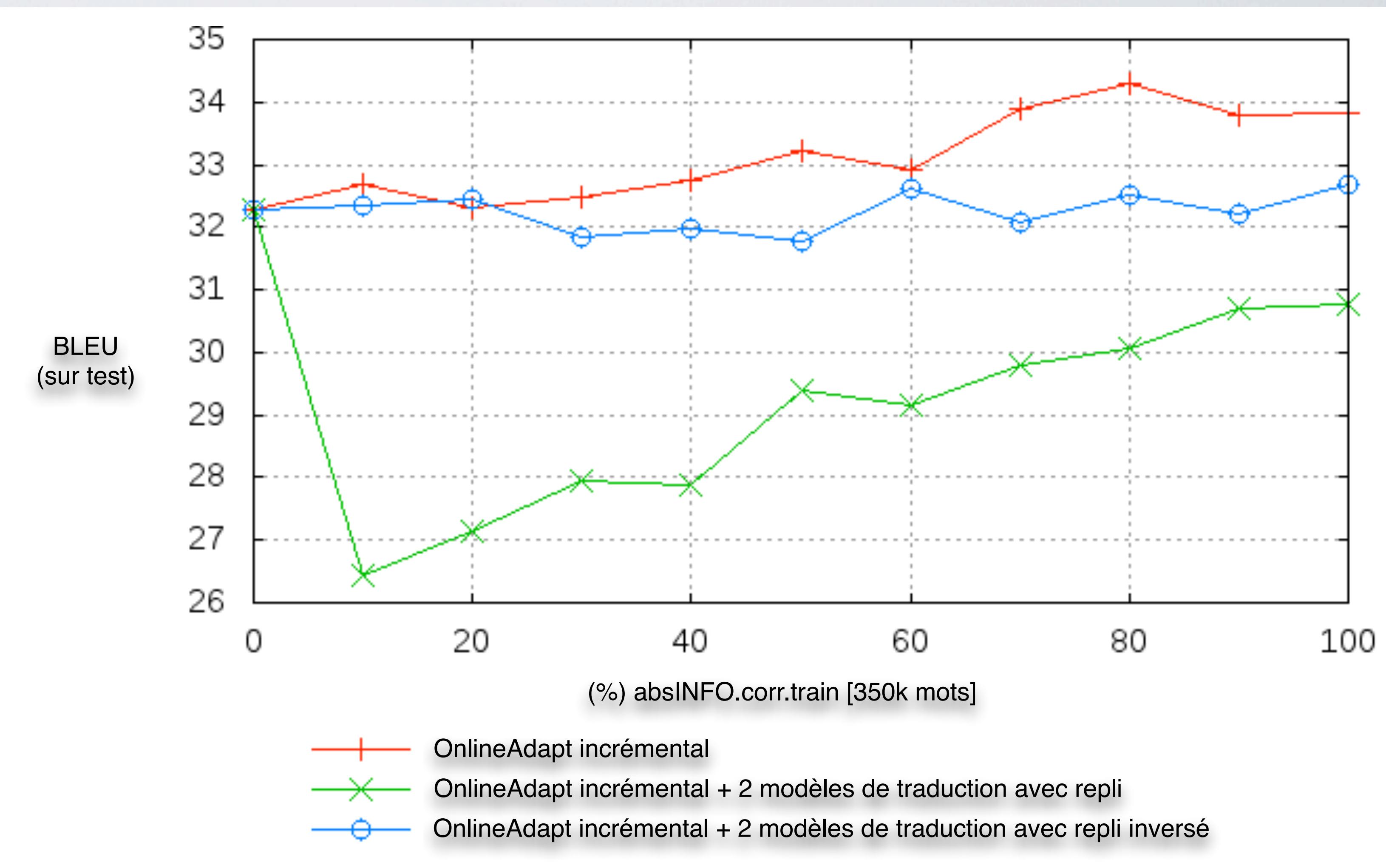
- I. PT en domaine

2. PT générique

- Repli-inversé :

- I. PT générique

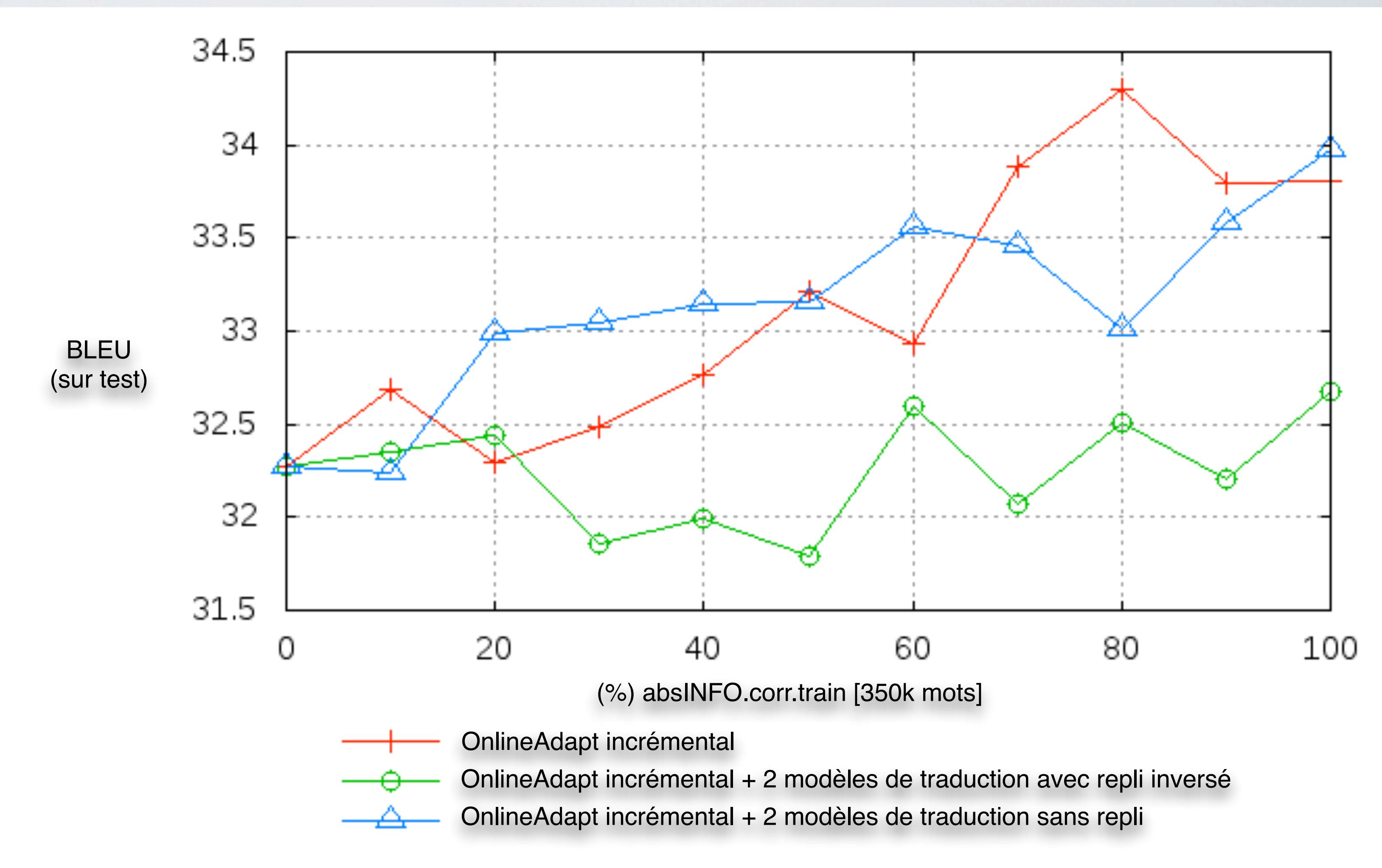
2. PT en domaine



RÉSULTATS EXPÉIMENTAUX

Modèles de traduction

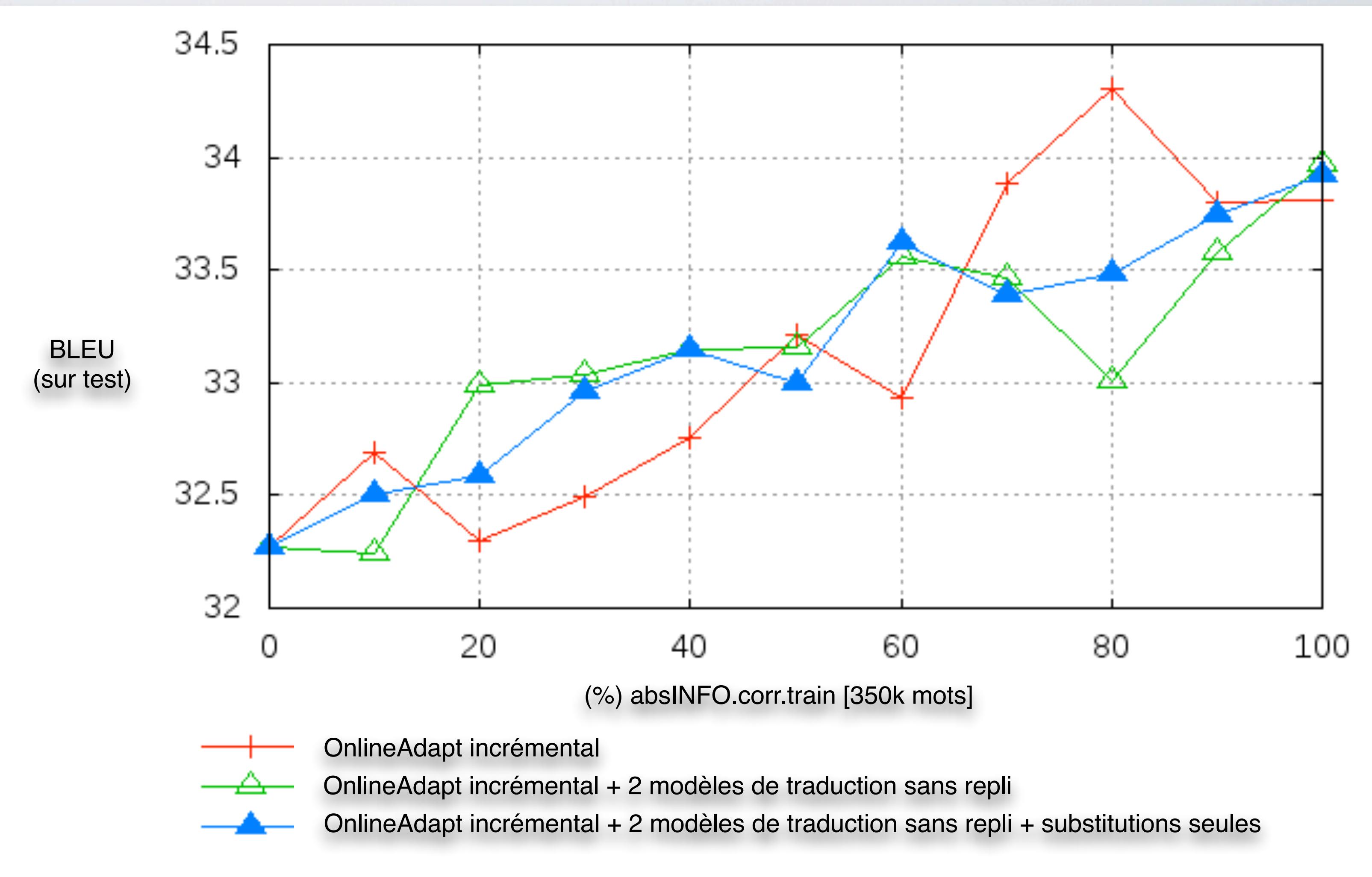
- ▶ repli-inversé vs. sans repli
- Repli-inversé :
 1. PT générique
 2. PT en domaine
- Sans repli : les 2 PT sont utilisées simultanément



RÉSULTATS EXPÉIMENTAUX

Modèles de traduction

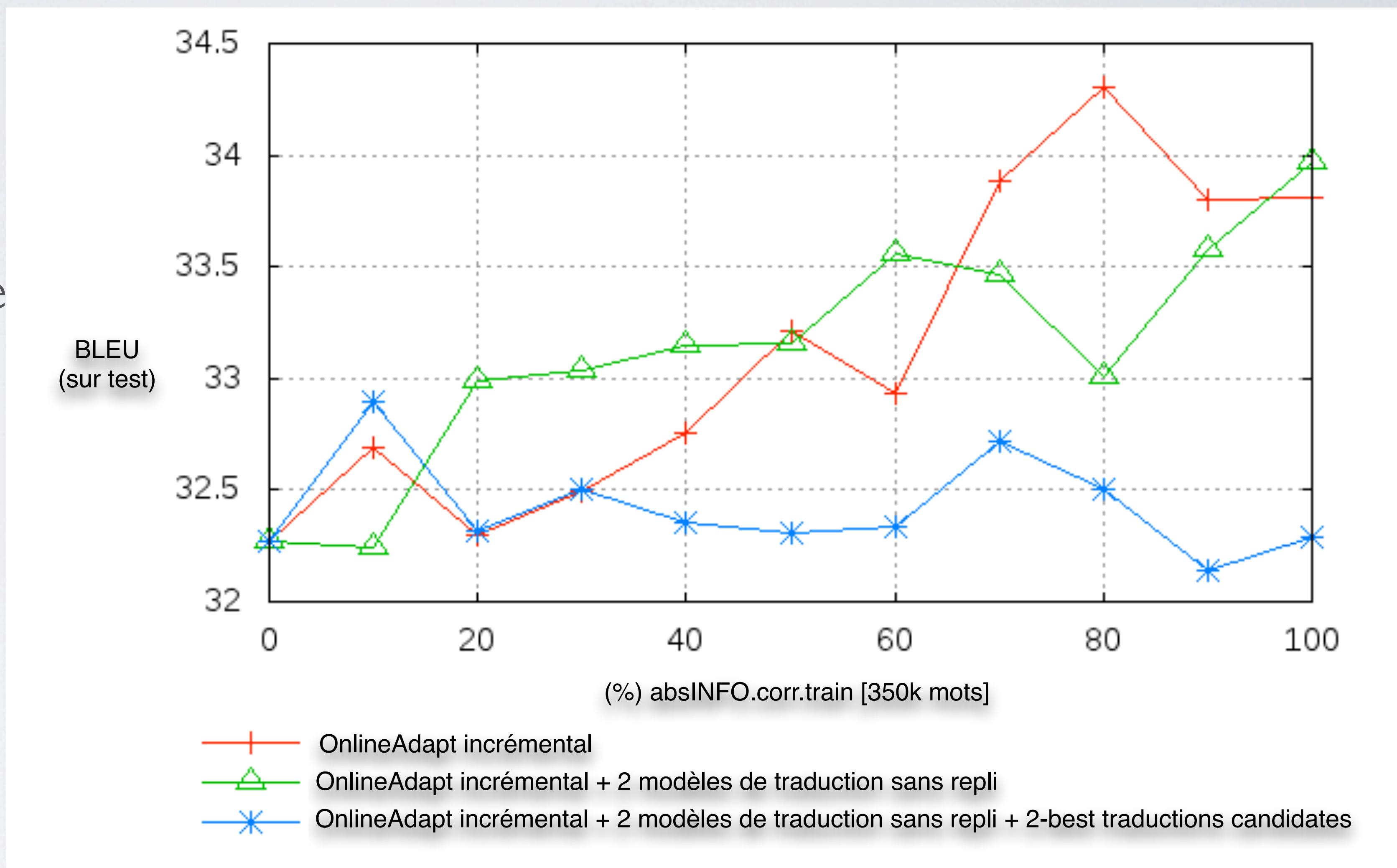
- ▶ sans repli vs. substitutions
- Seules les éditions de type « substitutions » sont utilisées pour générer les alignements source-référence



RÉSULTATS EXPÉIMENTAUX

Modèles de traduction

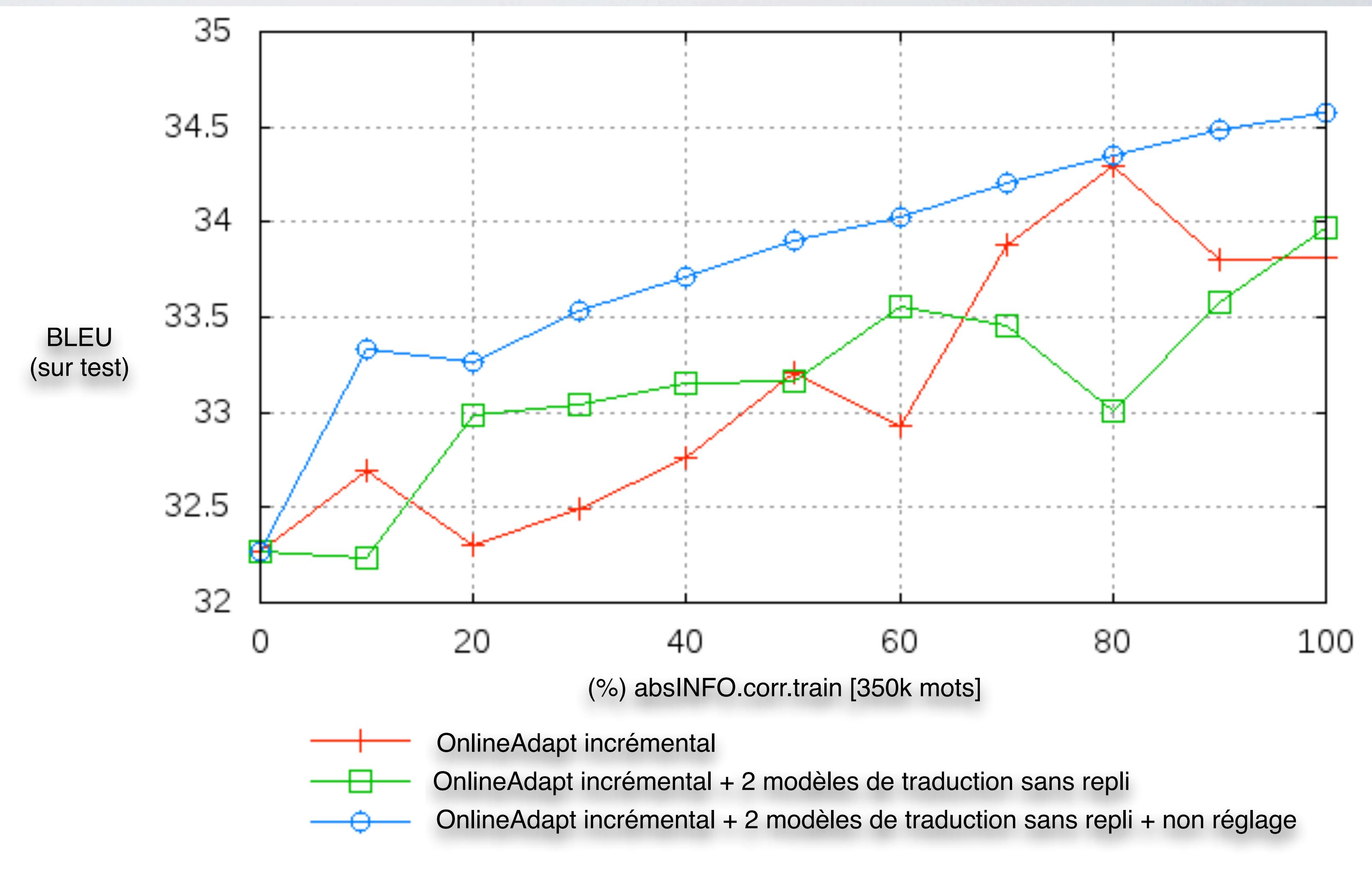
- ▶ sans-repli vs. n-best (n=2)
- Deux mêmes hypothèses de traduction mais alignements source-traduction possiblement différents



RÉSULTATS EXPÉIMENTAUX

Modèles de traduction

- ▶ sans repli vs. aucune optimisation
- Les poids associés aux PT ne sont pas réajustés après chaque itération...



RÉSUMÉ

Alignment source-référence

- Procédure plus rapide que l'état de l'art (quelques secondes vs. 14mins avec Inc-Giza)

Adaptation incrémentale en domaine

- Meilleure combinaison : 2 PT + sans-repli + pas d'optimisation à chaque itération
 - Pas d'optimisation = permet de bénéficier rapidement des données additionnelles
- Procédure adaptée et intégrable dans un processus de post-édition industriel.

CONCLUSIONS ET PERSPECTIVES

CONCLUSION

Contexte

- => Thèse CIFRE-Défense
- => Exploiter la post-édition dans un contexte industriel

Analyse qualitative de données post-éditées humainement

- => Nouvelle notion d'Actions de Post-Édition
- => Typologie de la post-édition pour le français
- => Procédure d'analyse automatique en APE d'un corpus post-édité
 - Précision et Rappel encourageants

CONCLUSION

Adaptation incrémentale d'un système de TAS

- => Nouvel algorithme d'alignement source-référence au niveau du mot
 - Utilisation temps-réel
- => Plus rapide que la procédure standard (quelques secondes seulement)
- => Procédure optimisée si on utilise deux modèles de traduction
 - 2 modèles de traduction + sans repli + pas d'optimisation entre chaque itération

CONCLUSION / PERSPECTIVES

COSMAT

=> Service de traduction de contenus scientifiques pour la communauté

=> Valorisations du projet

- Crédit d'un corpus scientifique bilingue ENFR unique (Physique, Informatique)
- Libération du code de l'outil Grobid

=> Lancement officiel avec HAL v3 en novembre 2013

- Collecte des premières données post-éditées
- Intégration de la procédure d'adaptation incrémentale

PERSPECTIVES

Analyse qualitative de données post-éditées humainement

=> Nouvelle métrique pour la post-édition ?

- Nouvelle version du TER basée sur les APE

=> Développement d'un outil d'annotation en APE sous licence open-source

- Générer un rapport en APE sur un corpus post-édité

- Idée plus précise sur la qualité d'un système de TA ?

- Réduction significative de l'effort de post-édition

PERSPECTIVES

Adaptation incrémentale d'un système de TAS

=> Renforcement de l'alignement des mots via l'interface de post-édition ?

- Historique des modifications de l'utilisateur
- Alignement possible au niveau des séquences de mots

=> Intégration de la procédure dans un contexte industriel ?

=> Intégration de la procédure dans le projet MateCat

- Post-doctorat au LIUM

PUBLICATIONS

- Blain Frédéric, Schwenk Holger, Senellart Jean. "*Incremental Adaptation Using Translation Information and Post-Editing Analysis*". International Workshop on Spoken Language Translation, Hong-Kong(China), Décembre 2012.
- Lambert Patrik, Schwenk Holger, Blain Frédéric. "*Automatic Translation of Scientific Documents in the HAL Archive*". Proceedings of the Eight International Conference on LREC, Istanbul(Turkey), Mai 2012.
- Lambert Patrik, Senellart Jean, Romary Laurent, Schwenk Holger, Zipser Floren, Lopez Patrice, Blain Frédéric. "*Collaborative Machine Translation Service for Scientific texts*". Proceedings of the demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Avignon(France), Avril 2012.
- Blain Frédéric, Senellart Jean, Schwenk Holger, Plitt Mirko, Roturier Johann. "*Qualitative Analysis of Post-Editing for High Quality Machine Translation*". Proceedings of the 13th Machine Translation Summit, Xiamen(China), Septembre 2011.

MERCI DE VOTRE ATTENTION

MODÈLES DE TRADUCTION ÉVOLUTIFS

Frédéric BLAIN

23 Septembre 2013

Directeur : Holger SCHWENK
Co-encadrant : Jean SENELLART

Rapporteurs : Laurent BESACIER
Marc DYMETMAN

Examinateurs : Yannick ESTÈVE
Patrik LAMBERT

