*Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles*
*Lille, France, 28 juin au 2 juillet 2021*

# *ACCOLÉ : Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs, Multi-Cibles, et Annotation d'Expressions Polylexicales*

Emmanuelle Esperança-Rodier[1] Francis Brunet-Manquat[1]
Univ. Grenoble Alpes, CNRS, Grenoble INP[1], LIG, 38000 Grenoble, France
prénom.nom@univ-grenoble-alpes.fr

## RÉSUMÉ

Cette démonstration présente les avancées d'ACCOLÉ (Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs), qui en plus de proposer une gestion simplifiée des corpus et des typologies d'erreurs, l'annotation d'erreurs pour des corpus de traduction bilingues alignés, la collaboration et/ou supervision lors de l'annotation, la recherche de modèle d'erreurs dans les annotations, permet désormais d'annoter les Expressions Polylexicales (EPL) dans des textes monolingues en français, et d'accéder à l'annotation d'erreurs pour des corpus de traduction multi-cibles. Dans cet article, après un bref rappel des fonctionnalités d'ACCOLÉ, nous explicitons les fonctionnalités de chaque nouveauté.

## ABSTRACT

**ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpora, Multi-Target corpora and Multi Word Expression Annotation.**

This demonstration presents the recent advances in ACCOLÉ, which on top of offering simplified management of corpora and typologies of errors, annotation of errors in bilingual aligned corpora, collaboration and/or supervision during annotation, looking for error types in annotations, now permits to annotate Multi-Word Expressions (MWE) in French monolingual corpora, and to access error annotation for multi-target corpora. In this article, after reminding the regular features of ACCOLÉ, we will explain the features of each novelty.

MOTS-CLÉS : Annotations d'erreurs de Traduction Automatique, Annotation collaborative, Evaluation de la qualité de la TA, EPL

KEYWORDS: Annotations of translation errors, Collaborative annotation, Machine Translation Quality Assessment, MWE

[1] Institute of Engineering Univ. Grenoble Alpes

# 1 ACCOLÉ, une plateforme pour l'annotation d'erreurs

ACCOLÉ permet l'annotation manuelle des erreurs de traduction selon des critères linguistiques. L'idée sous-jacente est de pouvoir fournir à un utilisateur une aide dans le choix d'un système de TA à utiliser selon le contexte (compétences linguistiques et informatiques de l'utilisateur, connaissance du domaine du document source à traduire et la tâche pour laquelle il a besoin de traduire le document source.) Pour ce faire, ACCOLÉ doit permettre de détecter quels sont les phénomènes linguistiques qui ne sont pas traités correctement par le système de TA étudié. Les principales fonctionnalités de la plateforme ACCOLÉ sont la gestion simplifiée des corpus, des typologies d'erreurs, des annotateurs, etc. ; l'annotation d'erreurs ; la collaboration et/ou supervision lors de l'annotation ; la recherche de modèles d'erreurs (type d'erreurs dans un premier temps, patrons morphosyntaxiques ultérieurement) dans les annotations. Nous avons privilégié un accès simple à l'outil ainsi qu'au corpus. La plateforme ACCOLÉ est donc disponible en ligne (http://lig-accole.imag.fr,) depuis un navigateur et ne nécessite aucune installation spécifique.
Un projet d'annotation renvoie à une tâche d'annotation, en créant un couple associant un corpus et une typologie d'annotation. Ainsi, un même corpus pourra être associé à plusieurs typologies sous forme de plusieurs projets d'annotation. Le corpus ne sera chargé qu'une fois sur la plateforme. Les annotateurs ainsi que les superviseurs sont associés aux projets qu'ils doivent annoter par le responsable du projet. Les typologies d'erreur sont également gérées par le responsable du projet. Un type d'erreur est composé d'un nom, d'une catégorie (facultative), d'une sous-catégorie (facultative) et d'un code (raccourci clavier pouvant être utilisé lors de l'annotation).
L'annotation se fait en deux étapes. La première étape consiste à sélectionner, à l'aide de la souris, des mots dans la phrase source, et de leur équivalent dans la phrase cible, présentant une erreur de traduction. Il est possible de sélectionner des mots disjoints dans la source et dans la cible. Dans le cas de mots non traduits (omission), il faut sélectionner l'espace dans la cible, à l'endroit où le ou les mots sources aurait dû être traduits. Dans le cas d'addition, il faut sélectionner l'espace entre les mots sources, correspondant à la position du ou des mots qui ont été ajoutés dans la cible entre les traductions de ces mots sources. La seconde étape consiste à choisir le type d'erreurs soit à l'aide de la souris, soit à l'aide des raccourcis clavier, à associer au couple des mots sources/cibles préalablement sélectionnés.
ACCOLÉ est une plateforme d'annotation collaborative, elle est donc dotée de deux mécanismes d'aide à l'annotation. Le premier est un mécanisme de supervision permettant à un responsable de contrôler l'avancée de la tâche. Ce mécanisme encourage surtout la communication entre superviseur et annotateur par la possibilité de créer des fils de discussion pour un couple de phrase source/cible précis. La supervision autorise la demande de précisions sur un type d'erreurs, de pointer une erreur d'annotation, etc.. Le second mécanisme est le mécanisme collaboratif qui permet aux annotateurs de communiquer autour d'un couple phrase source/cible précis. Ce mécanisme est une option à activer dans le projet. Du fait de son aspect collaboratif, ACCOLÉ répond aux problèmes d'accord inter-annotateurs (Popović, 2018).

# 2 ACCOLÉ, nouvelles fonctionnalités

## 2.1 Annotation d'erreurs multi-cibles

Pour l'analyse de la qualité de systèmes de Traduction Automatique Neuronale en traduction simultanée ou après complétion de la phrase entière - online & offline NMT - (Elbayad et al., 2020), ACCOLÉ s'est étoffée de l'annotation de plusieurs hypothèses de traduction correspondant à une seule phrase source et de l'intégrer d'une phrase de référence.
L'annotation d'erreurs sur un corpus multi-cibles se déroule de la même façon que pour un corpus mono-cible. L'annotateur sélectionne à l'aide de la souris le couple d'occurrence source/cible 1 source/cible 2, source /cible 3… présentant une erreur de traduction. La seconde étape consiste à choisir le type d'erreurs soit à l'aide de la souris, soit à l'aide des raccourcis clavier, à associer au couple des mots sources/cibles préalablement sélectionnés. En plus de la source, l'annotateur a accès à une traduction de référence, comme le montre la Figure 1 ci-dessous.

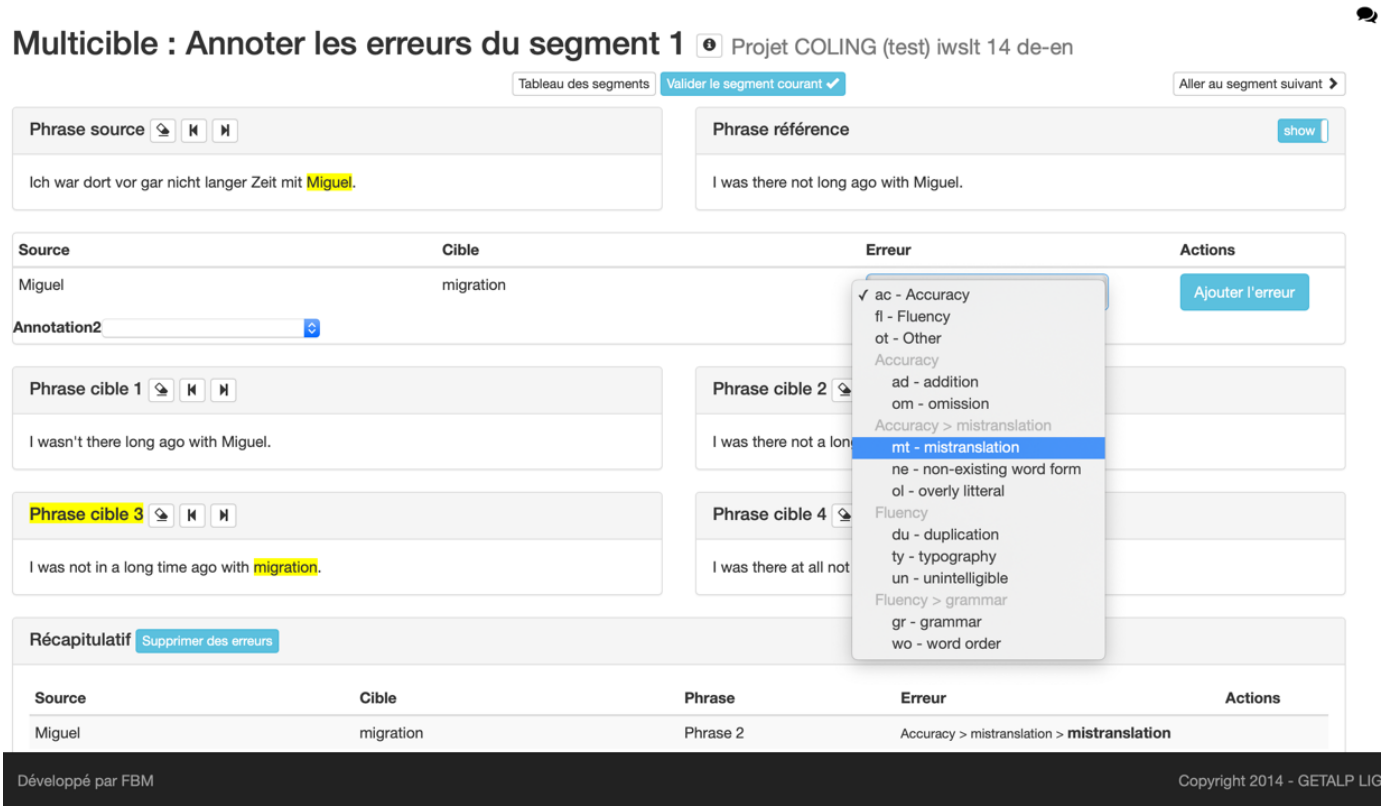


FIGURE 1 : Annotation d'une erreur sur la plateforme ACCOLÉ avec la typologie DQF-MQM (Lommel et al., 2018) dans un corpus multi-cible avec référence.

## 2.2 Annotation d'Expressions Polylexicales

Nous avons adapté ACCOLÉ pour l'annotation d'Expressions Polylexicales (EPL) comme l'illustre la Figure 2. La typologie de types d'EPL intégrée à notre plateforme est telle que définie dans le

travail de Tutin & al. (2017), composée de 9 types : Collocations, Mots Fonctionnels, Formules de Routine, Entités nommées, Phrasèmes complets, Pragmatèmes, Proverbes, Collocations fortes et enfin Termes Complexes. Chaque EPL est également annotée en partie du discours.
Nous avons envisagé que le corpus annoté soit monolingue ou bien bilingue. Toutefois, nous préférons la possibilité d'annoter la source en EPL de manière monolingue, de même que la cible. Si le corpus possède une traduction alignée du texte, alors il est possible d'annoter, à la fois dans la source et dans la cible, l'erreur repérée entre une EPL et sa traduction, afin de faire correspondre et comparer les annotations faites en première étape monolingue
Afin de faciliter la tâche d'annotation, un dictionnaire monolingue français d'EPL a été ajouté à ACCOLÉ, ainsi qu'un pré-traitement basé sur l'analyse syntaxique (Coavoux et Crabbé, 2017). Ainsi, ACCOLÉ permet d'annoter des EPL soit manuellement, en sélectionnant des mots à l'aide de la souris et en leur assignant un type, soit sur proposition de l'interface utilisant de manière automatique le dictionnaire et le pré-traitement, proposition qui sera à valider par l'annotateur.

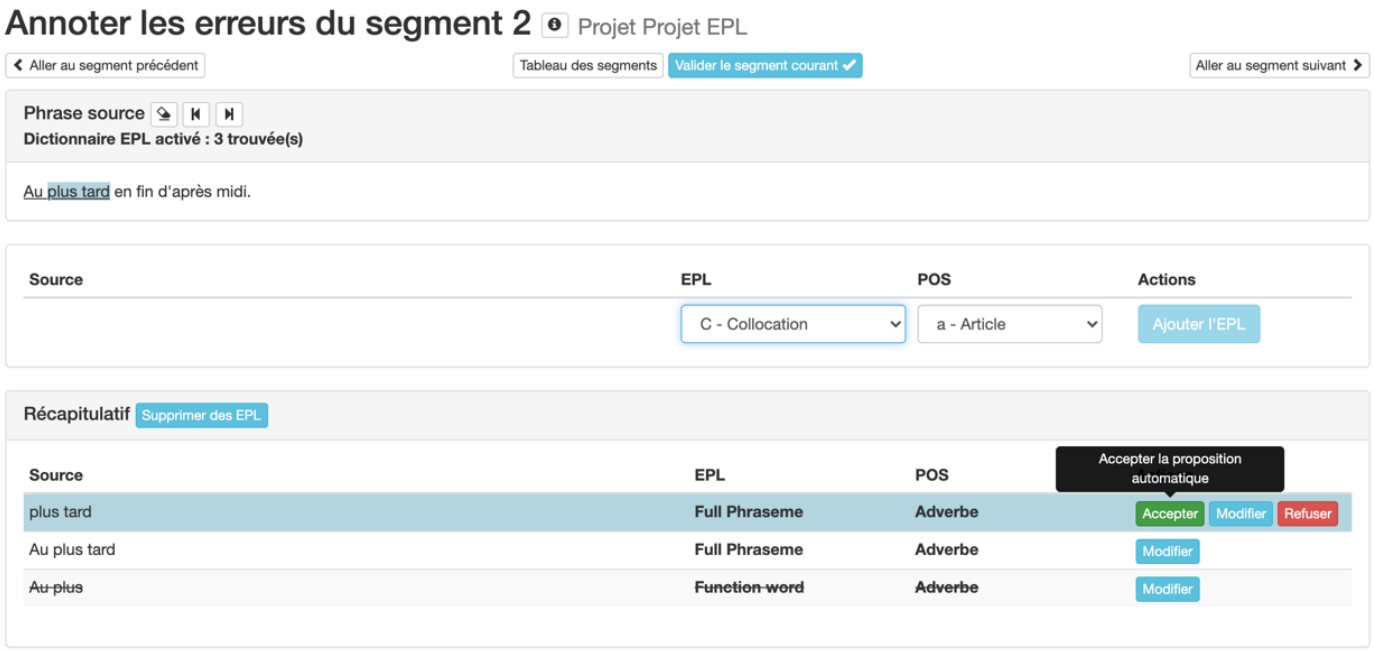


FIGURE 2 : Annotation d'un segment en EPL avec proposition automatique d'annotations

# 3 Données disponibles

ACCOLÉ propose 3 typologies d'erreurs, celle de Vilar et al. (2006), deux autres issues de MQM-DQF (Lommel, 2018) et 1 typologie d'annotation des EPL (Tutin et al, 2017) ainsi que 14 corpus FR-GB, 4 corpus monolingues FR/GB et 7 corpus GB-DE (allant des nouvelles journalistiques, à des documents techniques, des brevets, des extraits du BTEC (Basic Travel Expression Corpus) jusqu'à des documents sur le climat ou des textes médicaux) pour un total de 25 corpus (+66,6% en un an), ayant permis la création de 30 projets (+58%). Ceux-ci correspondent à 9 585 phrases (+40,5 %), 184 786 mots sources (+37,6%), 266 078 mots cibles (+132%), pour 34 558 annotations réalisées par 12 annotateurs natifs soit anglais, allemand ou français (+47%). Ces corpus sont

structurés selon les SNODEs (Boitet et al., 1988) et sont disponibles sur demande au format XML ou JSON. Une fonction permet de rechercher dans ces corpus les types d'erreurs. Au moment de la rédaction, nous continuons de travailler sur la recherche de modèle d'erreurs et plusieurs projets d'annotation sont en cours.

# Références


BOITET C. ET ZAHARIN Y. (1988). Representation trees and string- tree correspondences. In *Proceedings of international Conference on Computational Linguistics* COLING-88, 59-64.

COAVOUX M. ET CRABBÉ B. (2017). Représentation et analyse automatique des discontinuités syntaxiques dans les corpus arborés en constituants du français. *Actes de la 24e conférence sur le Traitement Automatique des Langues Naturelles, Jun 2017, Orléans, France. pp.77-92*

ELBAYAD M., USTASZEWSKI M., ESPERANÇA-RODIER E., BRUNET-MANQUAT F., VERBEEK, J. ET BESACIER L. (2020). Online Versus Offline NMT Quality: An In-depth Analysis on English–German and German–English. *Accepté à COLING 2020.*

LOMMEL A., ET ALAN K. M. (2018). Tutorial: MQM-DQF: A Good Marriage (Translation Quality for the 21st Century). *13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Papers). Vol. 2.*

POPOVIĆ, M. (2018). Error Classification and Analysis for Machine Translation Quality Assessment. *Moorkens J., Castilho S., Gaspari F., Doherty S. (eds) Translation Quality Assessment. Machine Translation: Technologies and Applications, vol 1. Springer.*

TUTIN A., ESPERANÇA-RODIER E. (2017). La difficile identification des expressions polylexicales dans les textes : critères de décision et annotation. "*La phraséologie française : débats théoriques et dimensions appliquées (didactique, traduction et traitement informatique)", Sep 2017, Arras, France.*

VILAR D., XU J., D'HARO L.F. ET AL. (2006). Error analysis of statistical machine translation output. *5th International Conference on Language Resources and Evaluation.* 97-702.