HAL open science

ACCOLÉ : Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs – Nouvelles fonctionnalités

Emmanuelle Esperança-Rodier, Francis Brunet-Manquat

To cite this version:

Emmanuelle Esperança-Rodier, Francis Brunet-Manquat. ACCOLÉ : Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs – Nouvelles fonctionnalités. 2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT), 2020, Montrouge, France. pp.1-8. hal-03047150

HAL Id: hal-03047150

https://hal.science/hal-03047150v1

Submitted on 3 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ACCOLÉ : Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs – Nouvelles fonctionnalités

Emmanuelle Esperança-Rodier[1] Francis Brunet-Manquat[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France

emmanuelle.esperanca-rodier@univ-grenoble-alpes.fr, francis.brunet-manquat@univ-grenoble-alpes.fr

RÉSUMÉ

Cet article 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 – New Features

This article 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 Traductions Automatique, Annotation collaborative, Evaluation de la qualité de la TA, EPL

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

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, documentation : http://lig-membres.imag.fr/fbrunet/accole-plateforme-pour-ledition-collaborative-derreurs/) depuis un navigateur et ne nécessite aucune installation spécifique.

 ©2020 CNRS Attribution 4.0 International.

1.1 Gestion des projets d'annotations

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. Ainsi, 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). Des typologies déjà existantes, telles que Vilar et al. (2006) et DQF-MQM (Lommel et al., 2018) sont décrites de cette manière dans ACCOLÉ. Des corpus (français-anglais) sont déjà disponibles dans ACCOLÉ, comme des extraits du BTEC (Basic Travel Expression Corpus), des nouvelles journalistiques, des articles sur la réglementation européenne issus de l'Union Européenne, des brevets. De plus, la plate-forme permet de téléverser de nouveaux corpus et de saisir d'autres typologies d'erreurs.

1.2 Annotation d'erreurs

La plateforme ACCOLÉ propose de visualiser et d'annoter les erreurs d'un couple de phrases source/cible (mono-cible) (Figure 1).

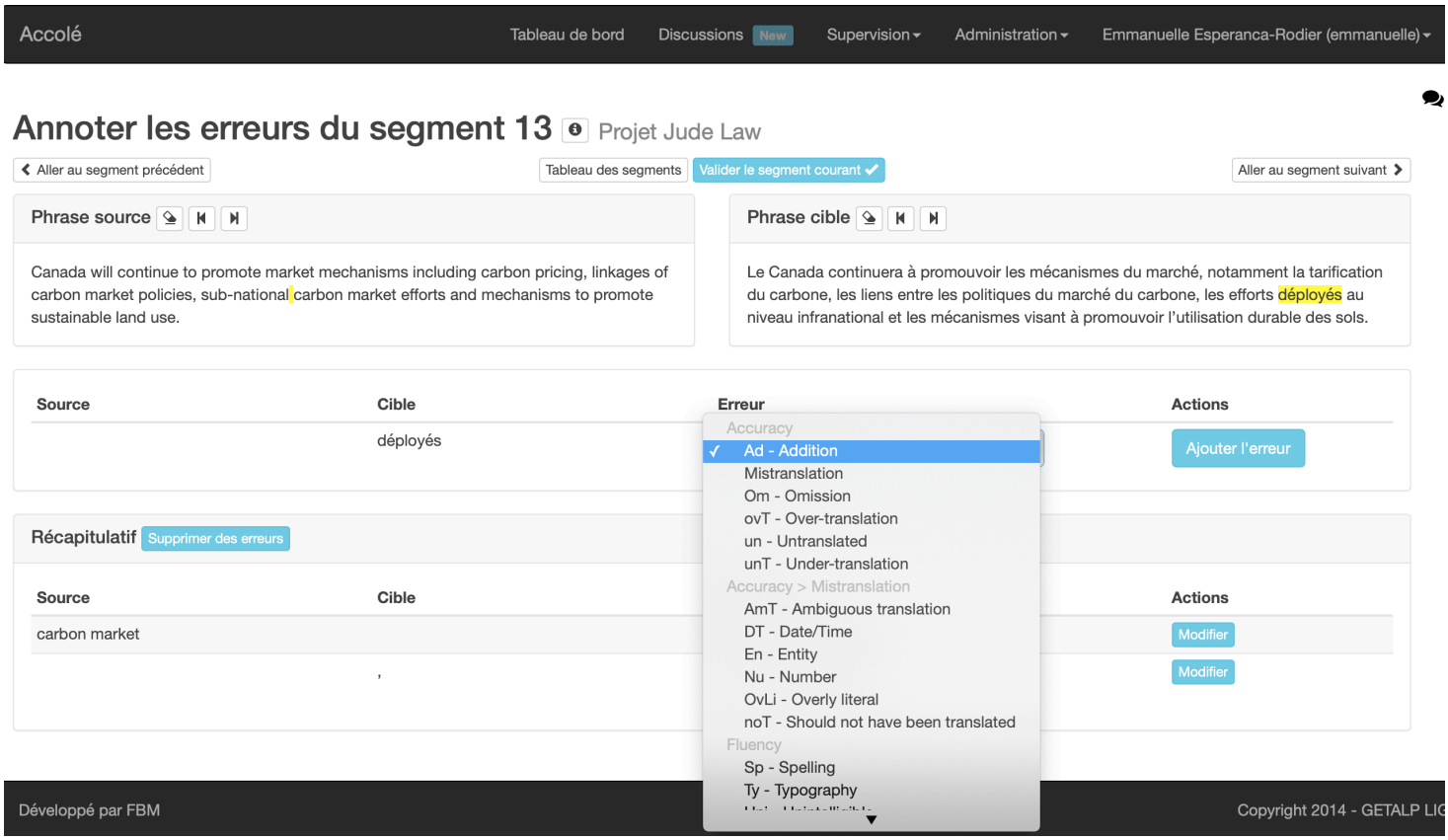

FIGURE 1 : Annotation d'une erreur sur la plateforme ACCOLÉ avec la typologie DQF-MQM (Lommel et al., 2018) dans un corpus mono-cible.

La figure 1 présente l'interface pour un corpus mono-cible proposée à l'annotateur. 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. Dans la figure 1, nous avons un exemple de mots ajoutés dans la traduction. « sub-national carbon market efforts » a été traduit par *« les efforts déployés au niveau infranational ». « déployés » a été ajouté dans la traduction alors qu'il n'est pas présent dans la source. L'espace entre « sub-national » et « carbon » a donc été sélectionné comme l'endroit équivalent à l'ajout dans la traduction de « déployés ». Cette façon d'annoter facilite la recherche de patrons pour un type d'erreurs, car il permet d'indiquer le contexte d'insertion, ou de suppression du mot respectivement dans la source et la cible.

Pour répondre aux problèmes d'accord inter-annotateurs (Popović, 2018), ACCOLÉ propose deux mécanismes pour aider l'annotateur dans sa tâche. 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.

1.3 Représentation des erreurs basée sur les SSTC

La plateforme utilise une représentation des données basée sur les SSTC (Structured String-Tree Correspondences, Boitet et Zaharin 1988). Une erreur est constituée d'une étiquette et d'un ensemble de SNODE (intervalle représentant la sous-chaîne dans la phrase source ou cible correspondante). Par exemple dans la figure 2, l'erreur portant sur "toute l'" et "any" est décrite par son étiquette Mauvais choix lexical (cat. Mot incorrect, sous-cat. Sens), par son positionnement dans la phrase source (SNODE [49-56] - sous chaîne entre le 49ème caractères et le 56ème) et la phrase cible (SNODE [46-48]). L'avantage d'utiliser ainsi les SNODE est de se passer d'une structure syntaxique pour décrire l'erreur.

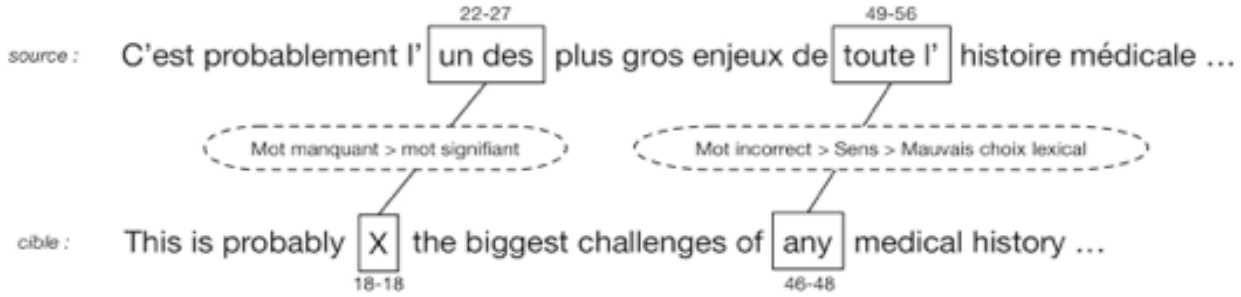

FIGURE 2 : exemple d'annotations

L'autre avantage est de pouvoir ajouter a posteriori des analyses morphosyntaxiques sur les phrases sources et cibles. Une ou plusieurs analyses (Stanford Tagger, Xerox Incremental Parser, etc.) pourront ainsi être rattachées à l'aide des SNODE aux phrases. A la fin de la tâche d'annotation, ses analyses serviront à rechercher des modèles d'erreurs (patrons morphosyntaxiques par exemple). L'idée est donc d'utiliser les erreurs comme une représentation pivot dans le mécanisme de recherche.

2 ACCOLÉ, nouvelles fonctionnalités

2.1 Annotation d'erreurs multi-cibles

Dans le cadre d'une expérimentation sur 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 la possibilité d'annoter plusieurs hypothèses de traduction correspondant à une seule phrase source et d'intégrer également une phrase de référence (Figure 3). Cette étude porte sur l'évaluation de la qualité des hypothèses de traduction, d'une même phrase source, issues de quatre systèmes différents de TA neuronale. Il s'agit d'effectuer une annotation des erreurs pour les quatre hypothèses de traduction en langue cible par rapport à la phrase source et à une référence de traduction. La typologie utilisée est une version adaptée de MQM-DQF (Lommel et al., 2018). Le corpus anglais-allemand annoté dans ce projet correspond à 200 segments issus des données IWSLT14.

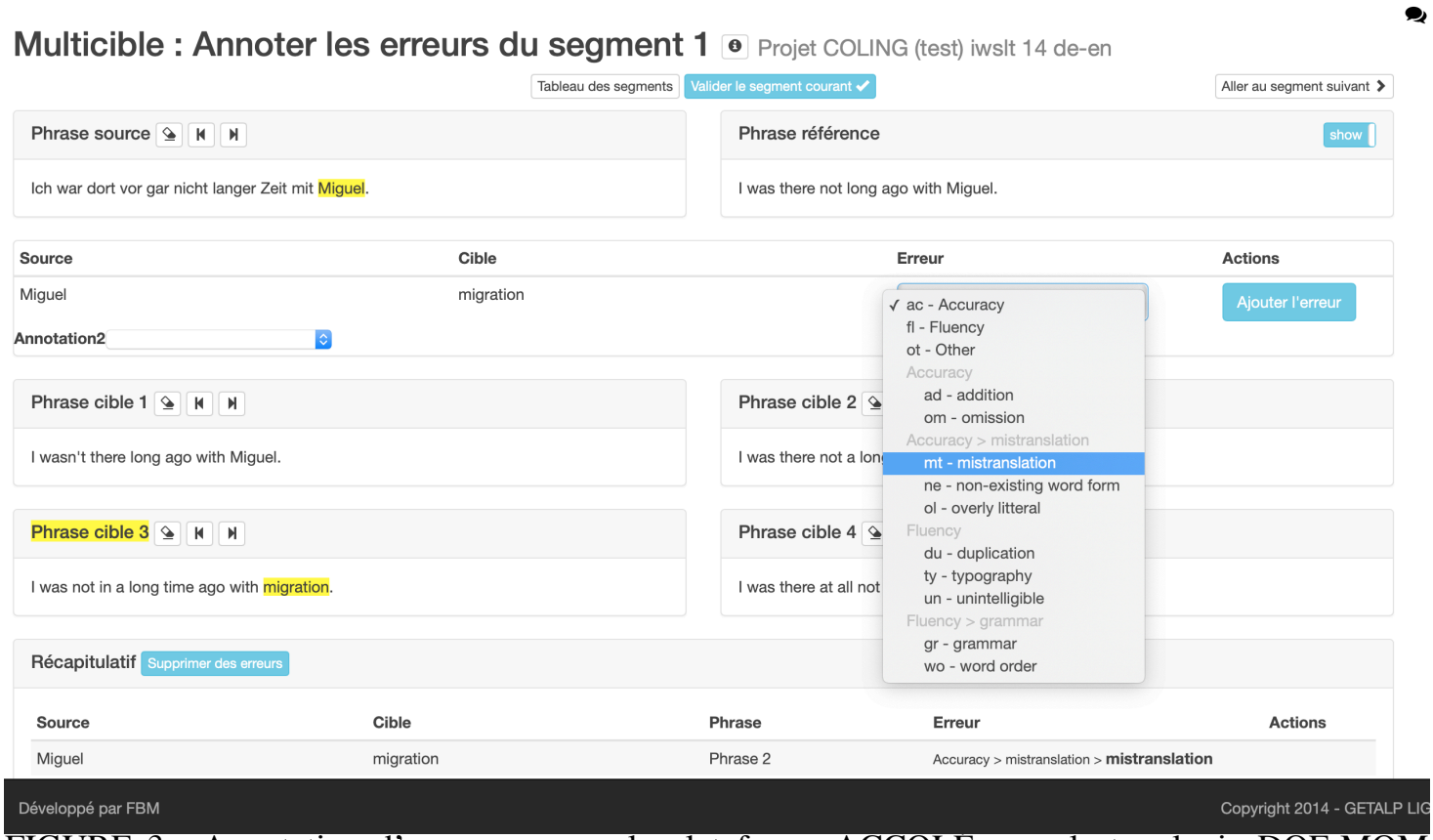

FIGURE 3 : Annotation d'une erreur sur la plateforme ACCOLE avec la typologie DQF-MQM (Lommel et al., 2018) dans un corpus multi-cible avec référence.

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. Dans l'exemple de la figure 3, « Miguel » a été traduit par « Miguel » dans la cible 1et la cible 4, et par *« migration » dans la cible 2 et la cible 3. Une erreur est donc annotée entre « Miguel » de la source, et *« migration » dans la cible 2 avec pour type « Accuracy>Mistranslation ». Comme le troisième moteur de traduction a engendré la même erreur de traduction, cette dernière est également annotée pour la cible 3 avec le même type d'erreur

2.2 Annotation d'Expressions Polylexicales

Dans le cadre d'un projet financé par NeuroCoG/Pôle Grenoble Cognition, nous avons adapté ACCOLÉ pour l'annotation d'Expressions Polylexicales (EPL). La typologie de types d'EPL intégrée à notre plateforme est telle que définie dans le travail de Tutin & Esperança-Rodier (2017). Elle est composée de 9 types : Collocations (C), Mots Fonctionnels (F), Formules de Routine (FR), Entités nommées (NE), Phrasèmes complets (PH), Pragmatèmes (PRAG), Proverbes (PROV),

Collocations fortes (SC) et enfin Termes Complexes (T). Chaque EPL est également annoté 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. Un projet financé par NeuroCoG/Pôle Grenoble Cognition, débutant en janvier 2021, nous permettra de vérifier de facto ce qui est le plus adapté pour l'annotateur.

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. Dans un projet, on peut choisir d'obtenir une proposition automatique des annotations (Figure 4) ou non.

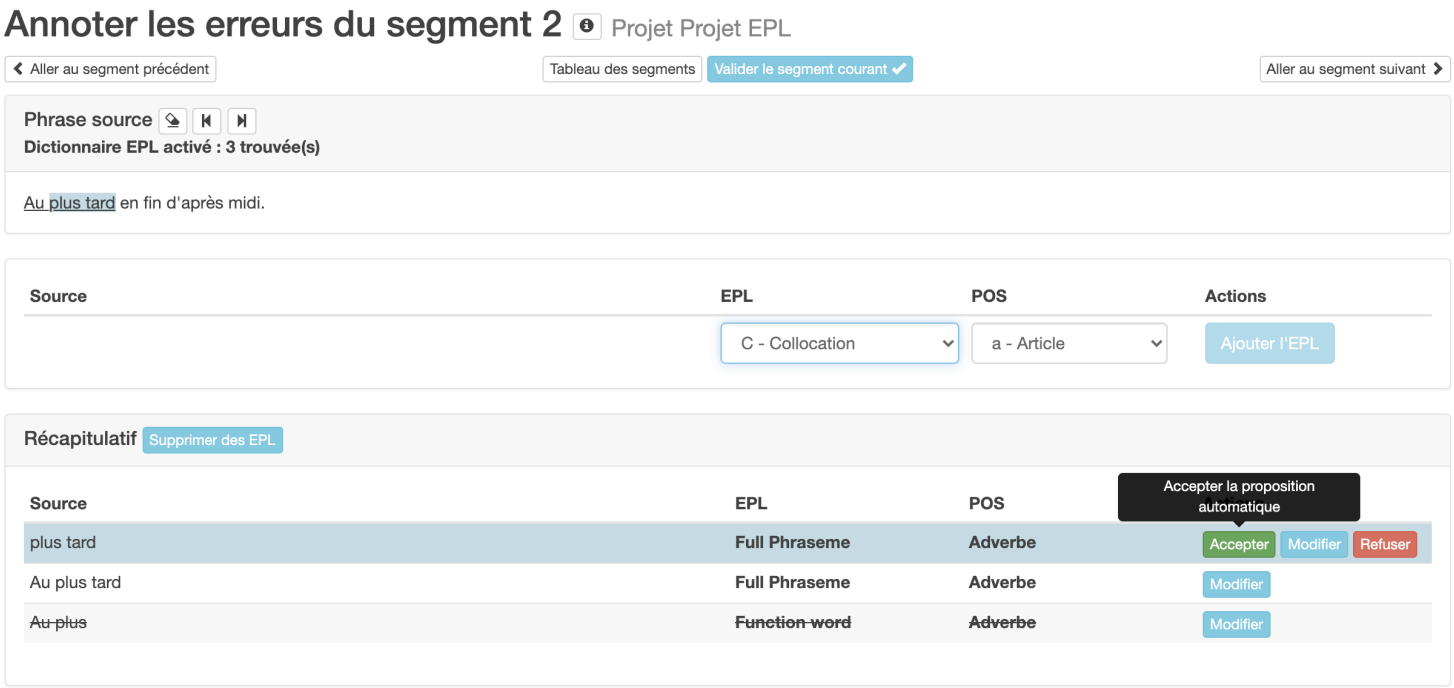

Figure 4 : Annotation d'un segment avec proposition automatique d'annotations à partir du dictionnaire

Dans la pratique, le dictionnaire est :

- soit utilisé par le mécanisme automatique pour déterminer toutes les EPLs d'un segment source et les proposer à l'annotateur dans l'éditeur (Figure 4). L'annotateur pourra les accepter, les refuser ou les modifier, c'est-à-dire les accepter mais en modifiant le type et/ou la partie du discours de l'EPL.

- soit le projet ne propose pas le chargement automatique des EPLs, mais l'interface d'édition précisera le nombre d'EPLs possibles (voir Figure 5), que l'annotateur décidera d'ajouter ou non.

Le mécanisme automatique repère les EPLs présentes dans le segment source à partir du dictionnaire. Le dictionnaire contient le mot du début d'une EPL. L'algorithme automatique cherche ensuite si les mots restants de l'EPL correspondent à la suite des mots du segment source. L'algorithme est pour le moment très simple car le dictionnaire est très simple. Après la fin des premières annotations, le dictionnaire sera amélioré (factorisation, lemmatisation, lien avec le format NIF (Hellman et al., 2013)). Il prendra en compte ces modifications en introduisant l'utilisation d'un lemmatiseur et, à plus long terme, proposera des annotations avec plus d'informations provenant du web sémantique.

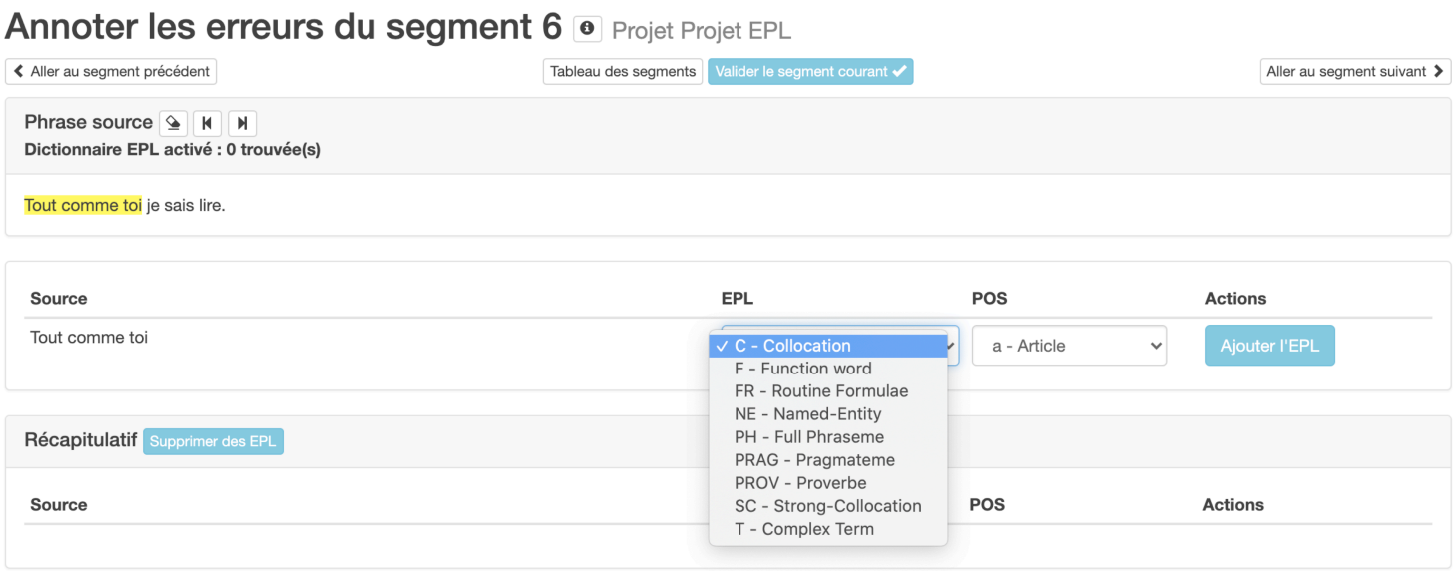

Figure 5 : Annotation d'une EPL de manière manuelle

L'analyse pré-syntaxique, quant à elle, propose déjà l'annotation en EPL du segment source. L'intégration à ACCOLÉ, dans un premier temps, consiste à faire correspondre l'analyse au formalisme d'ACCOLÉ, c'est-à-dire aux SNODEs. Puis dans un second temps, il s'agit d'ajouter les propositions d'EPL pour le segment source, et ce en faisant correspondre les types d'EPL avec les types utilisés pour l'annotation dans le projet et utilisés dans le dictionnaire.

La proposition d'annotation pour le segment source dans l'éditeur graphique sera la même que pour les propositions provenant du dictionnaire. Ces deux mécanismes ne diffèrent pas techniquement mais diffèrent dans le « workflow » de l'édition pour l'annotateur.

Nous avons obtenu un financement de NeuroCoG/Pôle Grenoble Cognition, pour réaliser avec ACCOLÉ l'annotation d'EPL dans un corpus monolingue français de 40,000 mots, que nous ferons traduire en polonais par différents moteurs de traduction automatique, notamment DeepL et Google. Nous étudierons ains le comportement de ces moteurs face à la traduction d'EPL du français vers le polonais. Cette étude nous permettra d'agrandir la quantité de corpus à disposition dans la plateforme ACCOLÉ, ainsi que d'ajouter le coupe français-polonais aux couples de langues jusque là étudiés. Nous serons également en mesure de créer un dictionnaire d'EPL en polonais.

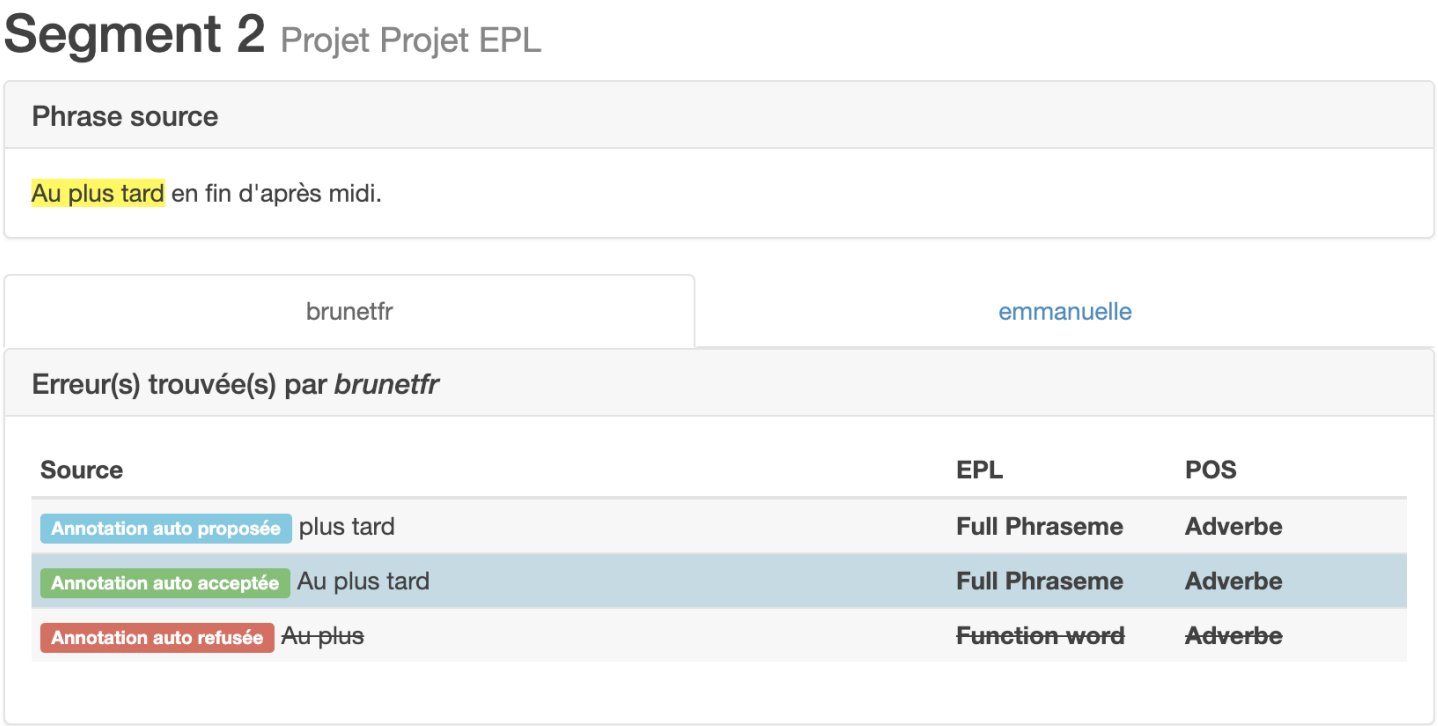

Figure 6 : Affichage des informations pour la Supervision

Enfin, le mode supervision de la plateforme (Figure 6), permet une visualisation des annotations d'EPL en fonction de leur provenance, si l'annotation a été réalisée manuellement ou bien si elle est issue d'une proposition automatique, et de leur statut, si l'annotateur l'a acceptée, modifiée, refusée ou bien si l'annotation a été proposée à l'annotateur mais est en attente d'acceptation, de refus ou de modification.

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 & Esperança, 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.

4 Conclusion

Dans cet article, nous présentons les fonctionnalités supplémentaires apportées à la plateforme ACCOLÉ. Certains corpus annotés disponibles ont déjà été utilisés pour une comparaison linguistique de la qualité de la traduction de différents systèmes de TA (Esperança-Rodier et Becker, 2018), ainsi que pour l'analyse de la qualité de systèmes de TA simultanée ou après complétion de la phrase entière (Elbayad et al., 2020). ACCOLÉ permet désormais d'étudier les phénomènes de traductions liés aux EPLs grâce à la production de corpus annotés en EPL afin de contribuer à la recherche en linguistique. Un projet d'annotation d'EPL et d'évaluation de TA français-polonais d'EPL vient d'être accepté par NeuroCog/Pôle Grenoble Cognition et débutera en janvier 2021.

ACCOLÉ permet également d'avoir une interface multi-cibles comme fréquemment utilisé lors des grandes campagnes d'évaluation. Outre l'implémentation de la recherche d'erreurs au sein des corpus par patrons morphosyntaxiques pour la partie annotation d'erreurs, nous étudions pour la partie annotation d'EPL, la possibilité de modifier le dictionnaire en cours de traitement : correction des entrées du dictionnaire, ajout d'entrées en fonction des annotations manuelles.

Enfin, nous comptons également intégrer des informations du web sémantique afin de proposer plus de format d'annotations, ou bien de données liées aux annotations.

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*.

Esperança-Rodier, E. et Becker, N. (2018). Comparaison de systèmes de traduction automatique, probabiliste et neuronal, par analyse d'erreurs. Proceedings of the 4th day on « *Traitement Automatique des Langues et Intelligence Artificielle » - TALIA 2018 Day of the plate-forme Intelligence Artificielle (PFIA 2018)*. Nancy, France, 6 juillet 2018. Edited by Didier Schwab et Pierre Zweignebaum.

Esperança-Rodier, E., Brunet-Manquat, F., et Eady, S. ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpora. *Translating and the computer 41, Nov 2019, Londres, United Kingdom*

Hellmann, S., Lehmann, J., Auer, S., Brümmer, M. Integrating NLP using Linked Data. *2th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia*

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, Cham*

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. and al. (2006) Error analysis of statistical machine translation output. *5th International Conference on Language Resources and Evaluation*. 97-702.