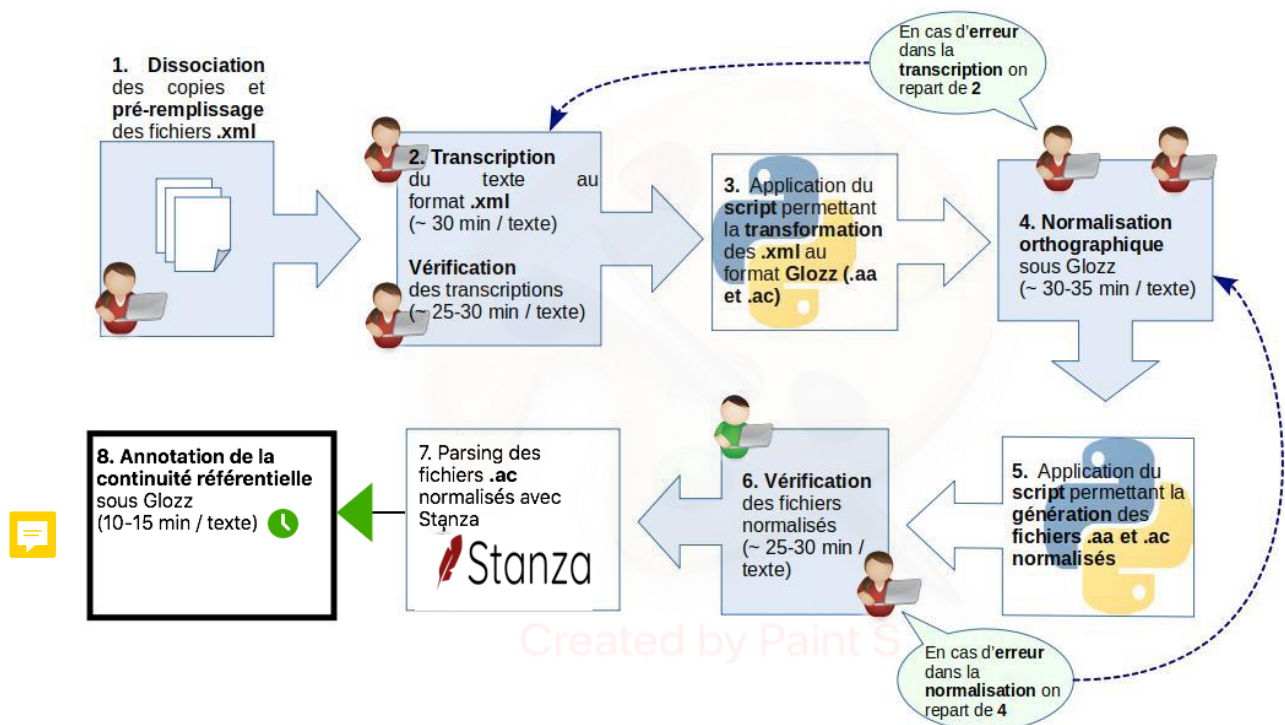




Ces étapes sont illustrées dans le schéma ci-dessous.



0. **Collecte** des copies : les élèves produisent un texte suivant une consigne (expliquée ci-dessous) et les copies sont numérisées au format .png ;
1. **Préparation** des copies : il s'agit de dissocier les scans reçus et de préparer les fichiers .xml, qui seront utilisés pour la transcription ;
2. **Transcription** et **vérification** des transcription : la transcription et la vérification sont effectuées par deux personnes différentes ;
3. **Génération** des fichiers pour l'annotation des erreurs d'orthographe : un script python permet de transformer les fichier .xml au format exploitable par l'interface d'annotation utilisée ;
4. **Normalisation orthographique** : il s'agit d'annoter les erreurs d'orthographe et en indiquer la version corrigée. Si les annotateurs rencontrent des erreurs provenant de la transcription, on revient à l'étape 2 pour corriger les .xml ;

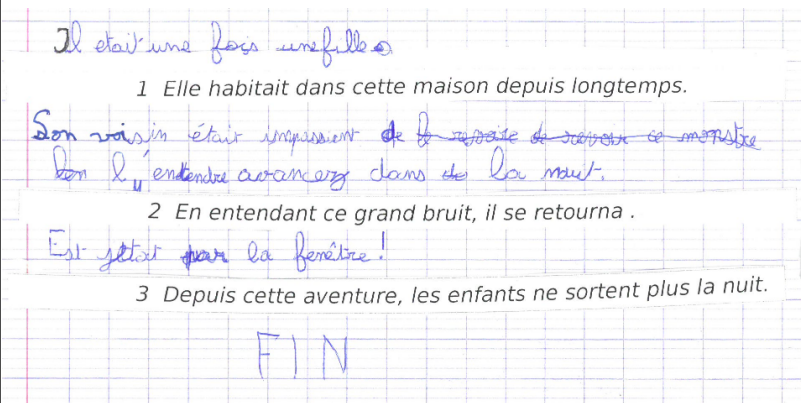
5. **Génération** des fichiers normalisés : un script python permet de générer les fichiers normalisés, c'est-à-dire sans erreurs d'orthographe ;
6. **Vérification** des fichiers normalisés : la vérification est effectuée par une personne différente de celle qui a réalisé l'annotation. Si des erreurs de normalisations sont rencontrées, on revient à l'étape 4 ;
7. **Parsing** des fichiers normalisés : pour le parsing nous utilisons *Stanza* (Peng Qi, Yuhao Zhang, Rui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations. 2020.).
8. **Annotation de la continuité référentielle** des textes normalisés grâce à l'interface d'annotation *Glozz* (Widlöcher A. and Mathet Y., 2009).

Consigne et collecte des copies

L'originalité et la particularité du corpus RésolCo réside dans le fait que les textes produits répondent tous à une même consigne d'écriture.

Celle-ci est une tâche-problème imposant aux élèves la résolution de problèmes de cohésion textuelle (*Garcia-Debanc et Bonnemaison, 2014* ; *Garcia-Debanc et Bras, 2016*)

L'objectif de cette consigne, fournie ci-dessous, est de provoquer chez le scripteur la mise en oeuvre de stratégies de résolution des problèmes de cohérence soulevés par l'intégration de trois phrases.

<p>Consigne : Racontez une histoire dans laquelle vous insérerez, séparément et dans l'ordre donné, les trois phrases suivantes :</p> <div style="border: 1px solid black; padding: 2px; margin: 5px;">Elle habitait dans cette maison depuis longtemps.</div> <div style="border: 1px solid black; padding: 2px; margin: 5px;">Il se retourna en entendant ce grand bruit.</div> <div style="border: 1px solid black; padding: 2px; margin: 5px;">Depuis cette aventure, les enfants ne sortent plus la nuit.</div> <p><i>Vous pouvez découper les bandelettes contenant les phrases ci-dessous ou bien recopier chaque phrase avec soin à l'identique de celles qui vous sont données.</i></p>	<p>Ci-dessous un exemple de copie récoltée en 2016 dans une classe de CE2.</p> 
--	---

Les trois phrases en jeu dans cette consigne impliquent des stratégies discursives variées, amenant le scripteur à gérer plusieurs continuités référentielles et planifier son discours afin d'assurer la cohérence de son texte (*Garcia-Debanc et al. 2017*).

Les copies sont ensuite numérisées au format .png pour la collecte des données.

Il est possible de contribuer à la récolte des copies en utilisant la consigne et les documents nécessaires à la collecte des textes. Cette consigne et ces documents sont disponibles au format .pdf au lien suivant : [documents pour la récolte des textes d'élèves](#).

Transcription des textes d'élèves

Afin de transformer les copies en une ressource exploitable par la communauté scientifique, et y appliquer des méthodes de linguistique de corpus et de TAL, il est nécessaire de passer par une étape de transcription des scans de copies.

Le format choisi pour la transcription est le format XML, selon la norme TEI-P5.

Toute transcription est assortie de métadonnées fournissant des informations sur la collecte et la numérisation de la copie, sur les conditions d'écriture et sur l'école qui a participé à la récolte des copies.

Les transcriptions sont anonymisées.

Pour ce qui concerne le corps du texte, l'objectif de la transcription est de reproduire le plus fidèlement possible le texte de l'élève ou de l'étudiant. Afin d'obtenir une reproduction fidèle, la mise en page ligne par ligne est renseignée ainsi que toute trace du processus d'écriture comme les ratures, les ajouts, les soulignements, etc. Aucune erreur d'orthographe n'est corrigée lors de la transcription.

Toute transcription est vérifiée par une personne différente de celle qui a transcrit le texte.

Pour plus de détails sur les éléments transcrits et les balises utilisées vous pouvez consulter :

- la version actuelle du [guide de transcription](#) proposé dans le cadre du projet [E:Calm](#) qui sera publié une fois finalisé sur le site du projet.
- le [canevas](#) d'une transcription au format XML qui liste tous les éléments saisis dans le *teiHeader* qui contient les métadonnées associées à chaque copie.
- l'exploration du [corpus transcrit](#)

Normalisation des textes d'élèves

Afin de faciliter l'analyse des erreurs orthographiques et pour pouvoir appliquer des méthodes de linguistique de corpus et de TAL, une étape de normalisation des transcriptions de copies a été prévue.

Cette phase consiste dans l'étiquetage des erreurs d'orthographe. Pour annoter ces erreurs on a choisi d'utiliser l'interface d'annotation [Glozz](#) (Widlöcher A. and Mathet Y., 2009).

Pour ce qui concerne l'annotation des erreurs d'orthographe, il a été décidé de ne pas classer les erreurs, en effet, la catégorisation des erreurs orthographiques sera effectuée par les experts du projet E-Calm qui travaillent sur cet aspect-là.

La normalisation du corpus RésolCo est effectuée par des annotateurs francophones.

En cas d'incertitude, les annotateurs peuvent indiquer une incertitude dans la détection, dans la correction de l'erreur ou bien une incertitude sur les deux.

Si plusieurs solutions de correction sont possibles, elles sont indiquées.

Toute normalisation est vérifiée par une personne différente de celle qui a annoté le texte.

Pour plus de détails sur les éléments normalisés et les décisions prises vous pouvez consulter :

- *la version actuelle du guide de normalisation* proposé dans le cadre du projet *E:Calm* qui sera publié une fois finalisé sur le site du projet.

Annotation de la continuité référentielle.

L'étude de la cohérence discursive dans les écrits d'élèves et d'étudiants fait partie des tâches principales au coeur du projet E-calm. Afin d'obtenir un aperçu de la maîtrise et de l'évolution de cette compétence parmi les rédacteurs, nous nous sommes concentrés sur des éléments de cohésion directement liés à la cohérence globale du récit et qui nous serviront d'indicateurs pertinents : les continuités référentielles (ou CR).

Afin d'analyser ces éléments de cohésion discursive, une annotation de chaque "maillon" (ou référence) des chaînes de continuité référentielle correspondant aux trois référents humains de la consigne a été effectuée pour les copies normalisées grâce à l'interface d'annotation *Glozz* (Widlöcher A. and Mathet Y., 2009).

L'alignement de ces textes annotés avec les textes parsés sous Stanza nous permet de récupérer les informations morpho-syntaxiques pour effectuer une analyse plus fine des chaînes de continuité référentielle.

Pour plus de détail concernant l'annotation ou pour explorer le corpus normalisé et annoté veuillez consulter la page *Exploration CR*