Le corpus spécifique traite de l'aérospatial, les données ont été extraites sur des sites d'aérospatial coréenne tels que Korea Aerospace Industries, ou encore KOREA AEROSPACE RESEARCH INSTITUTE, et Ministry of Science and ICT à cela s'ajoute des articles journalistiques venant pour la plupart de The Kyunghyang Shinmun, un quotidien coréen.
Le corpus général est composé de plusieurs corpus parallèles récupérés sur le web et de la partie que j'ai réussi à extraire du corpus Sejong obtenu grâce à une de nos enseignantes dans le cadre d'un autre projet.

Malheureusement, nous avons eu des difficultés à nous répartir le travail en ayant une membre d'un groupe ne faisant pas du tout partie de notre cursus, ne pouvant venir à tous les cours, et n'ayant aucune notions en informatiques et une autre membre qui fait partie du parcours IM et a donc un emploi du temps et des disponibilités différentes des nôtres. La communication étant peu présente, on a pour la majorité du projet dû avancer chacune de notre côté malgré quelques discussions. Mais, voyant que mes camarades avaient dû mal à avancer et étaient dépassées par les projets de leur autre filière, je leur ai transmis via discord mes scripts pour l'alignement, la tokenisation, le nettoyage de corpus et celui de tokenisation, en sachant qu'elles devraient bien sûr l'adapter à leur paire de langue ou peut-être choisir des outils plus adaptés ou plus faciles de compréhension pour elles.

Dans mon zip se trouveront aussi :

- Les scripts suivants :
    - convert.sh m'a permis de convertir les fichiers hwp en texte brut.
    - extract.py m'a permis de récupérer le contenu des fichiers textes obtenus suite à la conversion.
    - nettoie.py m'a permis de nettoyer le corpus de motifs que je considérais comme du bruit.
    - tokenisation.py m'a permis de tokéniser les corpus.

- Les notebooks suivants :
    - Alignement_final qui montre le choix final que j'ai choisi pour l'alignement
    - Alignement_test qui montre l'utilisation de l'outil de segmentation de phrase kss et de KoNLP qui s'est révélé infructueux.
    - ss_tokenisation qui traite la sous-tokenisation.