Ce projet de thèse se concentre sur les biais de genre en traduction automatique, en proposant d'identifier les mécanismes d'acquisition de ces biais par les modèles. La recherche de solutions pour réduire ces biais est alors vue comme une application de cet effort de compréhension.

L'approche proposée s'appuie sur des langues artificielles pour mener des expériences en conditions contrôlées, et ainsi apporter de la causalité aux observations faites sur ces biais.

Originalité L'étude des biais des modèles est une branche déjà très active du domaine, mais ce projet a l'originalité de s'intéresser spécifiquement à leur acquisition, afin de dépasser les interprétations souvent simplistes qui amalgament biais des données et biais des modèles. Le focus mis sur le biais de genre est plutôt en phase avec la littérature, mais la plus-value de ce choix est bien justifiée dans le descriptif, en ce qu'il offre ici un cadre expérimental intéressant plus qu'un objectif en soi.

L'approche adoptée est originale à trois égards. Tout d'abord, le regard linguistique porté sur ces phénomènes est une plus-value appréciable par rapport à l'existant. Sans être inédites, les expériences contrôlées sur des langues artificielles restent une approche peu explorée et qui devrait permettre une analyse plus rigoureuse des mécanismes en jeu que l'existant. Enfin, l'idée de traductions multiples en sortie du système, au-delà de permettre un allègement des biais, a le potentiel d'offrir un éclairage nouveau sur le principe même de traduction (ces difficultés affectant aussi les humains).

Faisabilité Le cadre des travaux et la démarche envisagée ont l'avantage de permettre une progression constante tout au long du projet. L'inscription dans une littérature riche permet de s'appuyer sur une solide base de connaissances sur les phénomènes de biais de genre, et offre donc un point de départ assuré pour explorer cette nouvelle méthodologie. L'objectif de fond (comprendre les mécanismes d'acquisition de biais) est très ambitieux, mais en matière d'interprétabilité tout nouveau résultat est un pas certain en avant, donc il est cohérent d'avoir un objectif à géométrie variable qui pourra s'ajuster aux résultats obtenus. La démarche très expérimentale qui est envisagée garantit de fait que des résultats seront obtenus – et même des résultats négatifs seraient intéressants.

Peut-être les intuitions liées aux "facteurs plus complexes" (qu'un simple déséquilibre) élaborées en stage de M2 auraient pu être davantage décrites dans le document, afin de fiabiliser le démarrage de thèse sur des premières pistes d'exploration bien cadrées. Le fait que ce soient des travaux déjà en cours est toutefois un plus.

Qualité générale Ce projet porte sur un sujet important, autant pour la science que la société, et dont les enjeux sont bien décrits. La justification de cette thèse s'appuie sur une solide intuition, mais aussi sur des faits, observations empiriques et arguments linguistiques. Les premières idées ébauchées quant à l'interaction entre biais et manque d'information (qui ne crée pas forcément le biais, mais le rend problématique du fait de l'obligation de choix qu'a le modèle) seraient intéressantes à creuser davantage quant à leur impact sur la méthodologie.

Le planning prévisionnel aurait pu être un peu plus détaillé, toutefois il semble cohérent au regard du dynamisme de ce domaine (ces modèles en général, et l'étude de leurs biais d'autant plus) que le planning de thèse soit lui-même mouvant pour s'ajuster aux avancées concomitantes de la communauté.

En résumé, tout en s'inscrivant dans une tendance globale, ce projet apporte un regard critique sur l'existant qui lui permet d'adopter un angle conceptuel et méthodologique nouveau, et me semble donc très prometteur.