

## Pre-training Data Quality and Quantity for a Low-Resource Language: New Corpus and BERT Models for Maltese

Micallef, K., Gatt, A., Tanti, M., van der Plas, L., & Borg, C. (2022, May 26). *Pre-training data quality and quantity for a low-resource language: New corpus and BERT models for Maltese*. arXiv. <https://doi.org/10.48550/arXiv.2205.10517>

GitHub : <https://github.com/MLRS/BERTu>

### Contexte

Le travail de Micallef et al. porte sur des tâches de compréhension et d'analyse linguistique automatique, en particulier l'étiquetage morphosyntaxique (*Part-of-Speech tagging*), l'analyse syntaxique des dépendances (*dependency parsing*) et la reconnaissance d'entités nommées (*Named Entity Recognition*), ainsi que l'analyse sémantique de sentiments (*sentiment analysis*). Ces tâches sont utilisées comme tests de référence (*benchmarks*) pour évaluer la qualité et la robustesse des représentations linguistiques apprises par les modèles BERT.

Micallef et al. s'inscrivent dans la continuité des recherches sur les modèles de langue pour les langues à faibles ressources. En effet, l'article mentionne à plusieurs reprises les travaux de Chau et al. (2020) et de Muller et al. (2021), qui ont montré que le pré-entraînement de mBERT sur une langue spécifique améliorait considérablement les performances. Les auteurs vont plus loin en créant un nouveau corpus monolingue de qualité pour le maltais, Korpus Malti v4.0, à partir duquel ils ont pré-entraîné deux modèles : BERTu, modèle monolingue entraîné de zéro, et mBERTu, modèle multilingue basé sur mBERT et affiné sur les textes maltais.

L'étude adopte une approche expérimentale et comparative, combinant analyses quantitatives et qualitatives pour évaluer l'effet de la taille et de la diversité du corpus sur les performances en morphosyntaxe et en sémantique. Les chercheurs ne visent pas à créer un nouvel outil applicatif, mais à mieux comprendre les conditions d'un pré-entraînement efficace pour une langue peu dotée.

Leurs résultats montrent que, même avec un corpus de taille modeste, un modèle monolingue bien entraîné peut atteindre – voire dépasser – les performances d'un modèle multilingue plus vaste, soulignant ainsi le rôle central de la qualité et de la représentativité du corpus utilisé lors de la phase de pré-entraînement dans la réussite des modèles de langue.

### Méthodologie

Les expériences ont été menées selon deux axes principaux, chacun visant à isoler un facteur clé du pré-entraînement : l'impact du domaine du corpus de pré-entraînement, d'une part, et l'impact de la taille du corpus de pré-entraînement, d'autre part.

Dans le premier cas, les chercheurs ont comparé deux configurations : l'une avec un pré-entraînement sur les seuls textes Wikipédia, l'autre sur un sous-ensemble mixte du Korpus Malti v4.0 constitué de 19 domaines variés (presse, débats parlementaires, textes juridiques, blogs, littérature, etc.). Cette comparaison visait à déterminer si la diversité thématique du corpus améliorait la performance du modèle par rapport à un corpus encyclopédique plus homogène.

Dans le second cas, afin de simuler différents niveaux de « richesse linguistique », les chercheurs ont créé différentes fractions (10%, 20%, 30%, ... jusqu'à 100%) de Korpus Malti v4.0. Ils ont ensuite entraîné autant de versions distinctes de leurs deux modèles BERTu et mBERTu sur ces sous-ensembles. Afin de garantir une comparaison équitable, ils ont ajusté les paramètres de pré-entraînement de façon à ce que chaque modèle voie le corpus le même nombre d'époques, c'est-à-dire de façon à ce qu'il parcourt les données le même nombre de fois que le modèle complet (100%). Cette précaution méthodologique a permis de s'assurer que les différences de performance observées provenaient uniquement de la quantité et de la diversité des données, et non d'un entraînement plus ou moins long.

## Résultats et évaluation

Les auteurs évaluent leurs modèles en comparant les prédictions des modèles à des jeux de données annotés manuellement pour les quatre tâches standard susmentionnées : analyse syntaxique en dépendances, étiquetage morphosyntaxique, reconnaissance d'entités nommées et analyse de sentiment. Ces tâches ont été choisies parce qu'elles mesurent différents niveaux de compréhension du langage, allant de la structure grammaticale à la compréhension sémantique.

Les sorties des modèles ont été comparées à des étiquettes de référence produites manuellement dans des corpus linguistiques maltais existants. Les performances ont été quantifiées à l'aide de métriques standardisées selon la nature de tâche : « Accuracy » pour l'étiquetage morphosyntaxique, « Labeled/Unlabeled Attachment Scores » pour l'analyse syntaxique des dépendances, « score F1 » pour la reconnaissance d'entités nommées et « macro F1 » pour l'analyse de sentiments.

Les scores ont été calculés automatiquement à partir des prédictions des modèles et des annotations de référence, puis moyennés sur cinq exécutions pour de réduire les effets aléatoires liés à l'entraînement. Ce protocole a permis de garantir la fiabilité statistique des résultats. L'analyse des performances permet ensuite d'examiner l'influence de deux variables principales : le domaine du corpus de pré-entraînement et la taille du corpus, ainsi que la différence entre le modèle monolingue (BERTu) et le modèle multilingue affiné (mBERTu).

Tout d'abord, les résultats montrent que les modèles entraînés sur le corpus mixte (19 domaines variés) surpassaient nettement ceux entraînés uniquement sur Wikipédia, en raison d'une plus grande diversité lexicale et syntaxique. Cette diversité permet aux modèles de mieux généraliser à différents types de textes. Une exception concerne la tâche de la reconnaissance d'entités nommées, où Wikipédia seul restait légèrement meilleur, étant donné que son vocabulaire correspondait davantage aux textes du jeu d'évaluation (issus de Wikipédia). Les

auteurs expliquent ce phénomène par une proximité de domaine entre le corpus d’ entraînement et le jeu d’évaluation (lui aussi issu de Wikipédia). Ils en concluent que la diversité thématique du corpus renforce la robustesse linguistique, sauf lorsque la tâche d’évaluation cible un domaine extrêmement spécifique.

Ensuite, les expériences portant sur la taille du corpus de pré-entraînement confirment que la quantité de données influence les performances, mais de manière non linéaire. Les modèles entraînés sur seulement 10 % du Korpus Malti v4.0 surpassent déjà le modèle mBERT d’origine, et les performances s’améliorent rapidement entre 10 % et 50 % des données avant de se stabiliser. Au-delà de ce seuil, les gains deviennent marginaux, voire nuls, suggérant que la qualité et la propreté du corpus comptent davantage que son volume brut. Ces résultats démontrent qu’un petit corpus bien nettoyé et bien équilibré peut être plus efficace qu’un corpus massif mais hétérogène.

Enfin, la comparaison entre les deux modèles révèle une complémentarité intéressante. En effet, le modèle monolingue (BERTu) dépasse le modèle multilingue affiné (mBERTu) sur les tâches sémantiques (analyse de sentiments, reconnaissance d’entités nommées), car il capture plus finement les nuances lexicales et les particularités morphologiques du maltais. À l’inverse, mBERTu conserve un léger avantage sur les tâches syntaxiques (étiquetage morphosyntaxique, dépendances) grâce au transfert interlinguistique issu de son pré-entraînement multilingue initial.

Les auteurs observent une amélioration moyenne de 10 à 20 points F1 pour les tâches sémantiques opérées par BERTu, tandis que les tâches syntaxiques progressent de manière plus modérée. Ces résultats suggèrent que les deux approches ne s’excluent pas mais se complètent : le modèle monolingue offre une compréhension linguistique fine et spécifique à la langue, tandis que le modèle multilingue est capable de généraliser les structures grammaticales communes à plusieurs langues.

Les résultats confirmant l’hypothèse principale des auteurs, à savoir qu’un pré-entraînement ciblé sur un corpus monolingue de qualité est plus bénéfique qu’un modèle multilingue pré-entraîné sur des données massives mais hétérogènes. L’étude illustre donc de manière empirique que la qualité du pré-entraînement (diversité, représentativité, nettoyage) joue un rôle déterminant dans la performance des modèles de langue pour les langues peu dotées.

### **Commentaire critique**

L’étude de Micallef et al. présente une méthodologie solide et transparente, tout en reconnaissant certains biais liés à la composition du corpus (surreprésentation de textes journalistiques et parlementaires) et à la proximité de domaine entre les données d’ entraînement et les jeux d’évaluation, en particulier pour la tâche de reconnaissance d’entités nommées. Ces limites n’invalident pas les conclusions, mais rappellent l’importance de la diversité des sources pour garantir la robustesse d’un modèle.

Il aurait néanmoins été intéressant d'évaluer ces modèles sur des tâches de traduction automatique, afin de vérifier si les bénéfices observés dans la compréhension du maltais se transposent à la génération bilingue, étape logique pour une application pratique des représentations apprises.

Dans l'ensemble, cet article constitue une excellente nouvelle pour les langues peu dotées : il montre qu'il est possible d'obtenir des modèles performants à coût relativement faible, à condition de disposer d'un corpus propre et bien structuré. Cette approche ouvre ainsi la voie à la création de ressources locales de qualité, pouvant être réutilisées pour d'autres langues européennes minoritaires, telles que le catalan.