**Résumé de l’article : Casablanca — Data and Models for Multidialectal Arabic Speech Recognition**

**Problématique :**  
Malgré les avancées de la reconnaissance vocale, la majorité des dialectes arabes restent sous-représentés dans les jeux de données publics. Ce manque freine le développement de systèmes robustes capables de traiter la diversité linguistique du monde arabe, notamment les dialectes peu ou pas dotés de ressources comme le mauritanien ou l’émirati.

**Approche proposée :**  
Les auteurs présentent **Casablanca**, un jeu de données supervisé de **48 heures** de discours en huit dialectes arabes (algérien, égyptien, émirati, jordanien, mauritanien, marocain, palestinien et yéménite). Ce corpus a été constitué via une collecte communautaire de vidéos YouTube, puis **annoté manuellement** par 27 locuteurs natifs. Chaque segment est étiqueté avec :

* transcription orthographique,
* dialecte,
* genre du locuteur,
* et présence de code-switching .

Les auteurs proposent également des **modèles de base** (baselines) pour la reconnaissance vocale, l’identification du dialecte et du genre, et une évaluation approfondie de plusieurs modèles multilingues (Whisper, SeamlessM4T, MMS).

**Technologies utilisées :**

* **Annotation** : LabelStudio, pyannotate, pydub pour la segmentation audio.
* **Modèles de reconnaissance vocale** : Whisper (v2 et v3), SeamlessM4T, MMS, et des variantes fine-tunées sur des dialectes.
* **Évaluation** : taux d’erreur sur les mots (WER) et sur les caractères (CER), y compris pour les segments en code-switching.

**Perspectives de recherche :**

* **Élargir la couverture dialectale** en ajoutant plus de dialectes et sous-dialectes (micro-dialectes).
* **Réduire le biais de genre**, encore présent dans le corpus.
* **Améliorer la transcription automatique dans des contextes de code-switching**, où les performances des modèles restent faibles.
* **Utiliser Casablanca comme base pour d’autres tâches NLP**, comme la synthèse vocale, la traduction dialectale ou la classification socio-linguistique.