Panel de discussion sur le thème des langues peu dotées

DELPHINE BERNHARD

Université de Strasbourg

dbernhard@unistra.fr

Ce panel de discussion a eu pour but de faire émerger des questionnements et de susciter des échanges entre participants aux journées jointes. Pour ce faire, nous avons utilisé l'outil https://speakup.info.

Dans une première phase, les participants ont rassemblé un grand nombre de questionnements, parmi lesquels :

- Comment énumérer toutes les ressources déjà existantes pour une langue peu dotée (et ses langues les plus proches) ?
- Quid de la pérennité des outils utilisés ? (Outils de récolte comme LIG-Aikuma, ASR avec ELPIS, etc)
- Comment définir si une langue est peu ou bien dotée ?
- Big Tech and other large companies are not particularly interested in supporting low-resource languages and mostly develop models for top n languages spoken in the world. How can we draw public interest to this topic?
- Comment encourager les chercheurs et chercheuses, en particulier en France, à s'intéresser davantage à ces langues ?
- Les modèles de langue multilingues résolvent-ils le problèmes des langues rares ?
- Quelles ressources doit-on créer pour les langues peu dotées en premier temps ?
- Est-ce qu'il y a un biais possible à essayer de traiter toutes les langues peu dotées avec une même approche ?
- Comment travailler sur des formes écrites de langue orales (sans conventions d'écritures, forte variabilité) ?

- Comment favoriser davantage les collaborations entre linguistes et « informaticiens » (au-delà du GDR) ?
- À quel point est-il nécessaire d'avoir des méthodes adaptées ?
- Quels sont les biais introduits par les modèles multilingues ?
- Les "gros" modèles de langues sont ils une opportunité ou une menace pour les langues peu dotées ?
- Quels outils seraient utiles pour les linguistes en documentation ?
- Peut-on exploiter la science participative pour les recherches sur les langues peu dotées ?
- Comment traiter des documents multilingues avec code switching ?
- Il y a-t-il une utilité à sélectionner les langues dans lesquelles sont entraînée les modèles multilingues avant de faire du transfert vers une langue peu dotée (eg langues très similaires à la langue cible)?
- Comment gérer l'évolution des langues peu dotées au cours du temps ?
- À quel point faut il connaître les spécificités culturelles ?
- Du point de vue du TAL, est-ce que c'est le manque de ressources qui constitue un obstacle ou les particularités de la langue ? Si c'est le manque de ressources, les études par ablation sur des langues bien dotées suffisent elles ?
- Comment favoriser l'interdisciplinarité au sein, et en dehors du laboratoire (linguistes, informaticiens, sociologues, philosophes, etc.) ?
- Est ce que les langues mortes sont considérées comme peu dotées ?
- Comment mobiliser des ressources (e.a. financières, postes) à long terme et pérennes pour le travail sur les différentes langues peu dotées ?
- What are the minimum requirements of a corpus for an endangered language?

Dans une seconde phase, les participants ont sélectionné certaines questions, puis les ont débatues par petits groupes pour faire émerger des éléments de réponse. Voici un résumé des échanges :

Groupe 1 (notes prises par Shu Okabe)

– Comment travailler sur des formes écrites de langue orales, sans conventions d'écritures, avec une forte variabilité ?

Cela dépend de ce que l'on entend par "travailler" : étudier ou analyser ? La variabilité a des origines multiples :

- données provenant d'internet
- fautes de frappe
- marqueur d'identité
- écriture "phonétique" (comme on entend)
- polygraphie
- emprunts

Les discussions ont notamment cherché à déterminer en quoi la variabilité est un problème, pour qui : les humains vs les machines ?

Pour formaliser une forme d'écriture, plusieurs moyens existent, par exemple :

- corpus parallèles
- simplification pour réduire la variabilité

– Comment élaborer une écriture pour une langue non écrite ? Dès le choix du script, de lourdes questions techniques, scientifiques, politiques et éthiques se posent !

Une décision est déjà opérée par le choix d'une écriture phono-graphique (écriture syllabaire ou alphabétique).

Une même forme peut changer en fonction de la géographie et de la culture.

Le choix de l'orthographe est fait par les linguistes (ex : alphabet latin, alphasyllabaire) mais aussi par les contraintes techniques.

Les choix des modes de communication (SMS, etc.) sont également influés par des contraintes techniques.

Le choix d'une écriture est une adaptation à la langue, aux contraintes techniques des émetteurs et des récepteurs.

Groupe 2 (notes prises par Frédéric Béchet)

– Les "gros" modèles de langues sont ils une opportunité ou une menace pour les langues peu dotées ?

Ils sont une opportunité pour les raisons suivantes :

- il y aura toujours une langue proche ou des caractéristiques communes avec la langue cible ;
- plus les modèles sont gros, plus on peut espérer capter des invariants de la langue ;
- l'affinage ou "fine tuning" permet d'accéder à un niveau de performances qui serait inaccessible si l'on se limitait à de l'apprentissage monolingue sur des corpus de taille limitée ;
- les gros modèles multilingues seraient moins biaisés (cf présentation de Karën Fort) ;
- le fait de traiter un très grand nombre de données sur de nombreuses langues permet d'analyser les caractéristiques de ces « gros »modèles et de révéler les possibles « biais » ou « failles » des modèles qui passeraient peut être inaperçu sur de plus petits modèles ;
- mutualisation des efforts au niveau technique / plateforme ;
- "démocratisation" du développement de modèles pour une nouvelle langue il « suffit » de collecter des données et de faire du « fine tuning » d'un gros modèle.

Ils sont également une menace pour les raisons suivantes :

- les gros modèles sont "biaisés" envers les langues les mieux représentées ;
- risque de perte d'intérêt pour les langues peu dotées si l'écart de performances avec les langues bien dotées devient trop grand (par ex. si l'ASR est parfaite en anglais, il y aura moins de financement sur ce thème pour les langues peu dotées)
- risque d'avoir des langues à 2 vitesses (ou 3 ou 4), entre les langues plus ou moins bien dotées ;

- “risque” de standardisation des méthodes de traitement des langues, qui peut-être ne s’appliqueraient pas à toutes les langues ? ;
- si une langue ne rentre pas dans les “canons” des gros modèles, même les linguistes de terrain vont favoriser les langues qui sont bien outillées ;
- est-ce que la notion d’espace multilingue a du sens pour des langues tellement différentes des langues “dominantes” (par ex. langue amazonienne / rapport a la nature)
- “appauvrissement” de la description des langues, car aucune représentions explicites dans les “gros” modèles.

Groupe 3 (notes prises par Benoit Favre)

– Est-ce que n’importe qui peut travailler sur les langues peu dotées (par exemple sans forcément avec des connaissances de la langue) ? A quel point est-ce qu’on devrait prendre en compte la dimension éventuellement politique/culturelle ?

La première réponse est non. Le chercheur doit connaître la langue sur laquelle il travaille

C’est plus facile si on est locuteur soi-même. Sinon, il faut mobiliser des locuteurs.

Être locuteur peut générer des biais qu’on risque de faire passer dans les recherches. On est aussi biaisé par notre langue première.

Il faut privilégier une approche participative, en gardant la nécessité d’expertise pour aller outre les biais.

Il faut réunir tous les acteurs de chaîne : locuteurs, linguistes et TALeux.

Doit-on prendre en compte l’aspect politique / culturel ?

Aspect éthique : ne pas impliquer des locuteurs de la langue peut être considéré comme une appropriation / oppression culturelle (par ex. perception d’une technologie comme résolvant les problématiques d’une langue rare, peut-être mal interprétée par les décideurs)

La langue et la culture font partie d’un même tout.

– Quels outils seraient utiles pour les linguistes en documentation ?

- collecter
- explorer (ex. concordancier)
- faire sur de l’oral tout ce qu’on peut faire sur l’écrit (par ex. acoustic unit discovery)
- annoter, aide à l’annotation

Groupe 4 (notes prises par Maximum Coavoux)

– Comment définir si une langue est peu ou bien dotée ?

"Bien / mal doté" est un concept de TAL. Les linguistes de terrain ne réfléchissent pas en terme de doté/mal doté, ils parlent plutôt de langues rares. Dans une perspective terrain, il s’agit d’une langue qui n’est pas une des 300/600 langues bien dotées.

Se pose également la question des langues avec beaucoup de ressources mais avec des corpus de mauvaise qualité.

Quelques critères de définition :

- la langue dispose-t-elle d’un status d’"endangerment"
- quel est l’éloignement de la langue par rapport à des langues bien dotées (ex: galicien)

- combien de personnes se consacrent à cette langue ?
- est-ce qu'il y a une grammaire / une description de cette langue?
- la langue nécessite-t-elle d'utiliser des méthodes spécialisées ?
- y a-t-il des difficultés à accomplir un objectif avec la langue ?
- la langue est elle traitée par Google / par Wikipedia ?