

Ioana Vasilescu, Lori Lamel (TLP)

Phonologie et RAP: évaluer le statut des phonèmes d'une langue avec la RAP

Contexte:

Cette proposition porte sur une *utilisation novatrice d'un système de RAP en tant qu'outil linguistique permettant de mettre en évidence la fonctionnalité de certaines oppositions phonémiques dans une langue, à partir de leur représentation dans la parole tout venant*. Il s'agit de « faire de la phonologie » à travers le paramétrage d'un système de RAP qui décidera, lui, et non pas le linguiste expérimentateur, du statut et du nombre des unités phonémiques de la langue. Nous avons adopté récemment cette approche dans le but de décider si l'inventaire phonémique du roumain, une langue peu étudiée dans sa version orale spontanée comportait 6 ou 7 unités (Renwick et al., 2016). Nous avons utilisé un système de RAP paramétré de sorte qu'on puisse évaluer lesquelles des oppositions phonémiques étaient les plus fonctionnelles (Vasilescu et al., 2014). Le WER et l'évaluation du nombre et type de vocabulaire affecté par les différentes configurations nous a permis de trancher sur un système « efficace » de 6 voyelles. Le *contexte* de ce travail est celui de la *phonologie dite de laboratoire* (laboratory phonology), qui aspire à définir les contrastes et les éléments sonores systémiques d'une langue grâce à des études « grand corpus », à partir de données traitées automatiquement (en s'appuyant sur des alignements automatiques et en utilisant des modélisations statistiques). Si les études « grand corpus » à partir desquelles les linguistes construisent des hypothèses phonologiques pertinentes représente aujourd'hui une méthode consacrée, la phonologie reste une démarche empirique, résultat des déductions du linguiste. Nous proposons d'utiliser la RAP pour objectiver l'analyse distributionnelle et le rôle phonémique des sons dans une langue.

Proposition :

Cette AI porte sur l'utilisation de systèmes de RAP pour différentes langues dont le nombre et le statut des unités dans le système fait débat. Nous allons nous focaliser sur les inventaires vocaliques. Nous allons prendre en compte des langues romanes pour lesquelles certaines oppositions phonémiques vocaliques font débat. Il s'agira du français (voyelles périphériques mi-ouvertes/fermées et voyelles centrales), italien (voyelles périphériques mi-ouvertes/fermées) et catalan (voyelles périphériques mi-ouvertes/fermées). Le choix de langues d'une même famille nous permettra de dresser une typologie sur ce critère (nous y rajouterons le roumain que nous avons traité). Nous allons appliquer la méthodologie décrite dans (Renwick et al., 2016) : différents jeux de phonèmes et oppositions seront considérées selon les hypothèses concernant le statut des voyelles sus-citées. WER et analyse des vocabulaires concernés seront indicateurs de l'état des phonèmes dans les langues. Nous ferons appel à des données et systèmes déjà présents au Limsi.

Nous nous attendons aux résultats suivants :

- mise en évidence des statuts phonémiques des voyelles dans plusieurs langues romanes
- hiérarchie des oppositions/perte des oppositions intra- et inter-langues
- plus largement, objectivation de l'analyse phonologique, validation de la méthode à travers les langues, éventuellement amélioration des jeux des phonèmes pour la RAP

Travail effectif :

Ce projet correspond à un stage ingénieur de 5 à 6 mois. Il s'agira de paramétrer des systèmes existants, de mener des expériences de reconnaissance vocale, d'évaluer le taux d'erreur et le vocabulaire sur lequel il porte (encadrement Lori Lamel). La démarche sera suivie par une analyse linguistique portant sur l'analyse lexicale et morpho-syntaxique des vocabulaires concernés par les oppositions phonémiques analysées, une typologie intra- et inter-langues et une analyse phonologique des résultats (encadrement Ioana Vasilescu)

Références :

Renwick, M.E.L., Vasilescu, I., Dutrey, C., Lamel, L., Vieru, B. 2016. Marginal contrast among Romanian vowels : evidence from ASR and functional load. In Proceedings of Interspeech, San Francisco, USA.
Vasilescu, I., Vieru, B., and Lamel, L., 2014. Exploring Pronunciation Variants for Romanian Speech-to-text Transcription. In Proceedings of SLTU, St. Petersburg, Russia.