Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées

Résumé

Les sciences participatives, et en particulier la myriadisation (*crowdsourcing*) bénévole, représentent un moyen peu exploité de créer des ressources langagières pour certaines langues encore peu dotées, et ce malgré la présence de locuteurs sur le Web. Nous présentons dans ce travail les expériences que nous avons menées pour permettre la myriadisation de ressources langagières dans le cadre du développement d'un outil d'annotation automatique en parties du discours. Nous avons appliqué cette méthodologie à trois langues non standardisées, en l'occurrence l'alsacien, le créole guadeloupéen et le créole mauricien. Pour des raisons historiques différentes, de multiples pratiques (ortho)graphiques co-existent en effet pour ces trois langues. Les difficultés posées par l'existence de cette variation nous ont menée à proposer diverses tâches de myriadisation permettant la collecte de corpus bruts, d'annotations en parties du discours, et de variantes graphiques.
L'analyse intrinsèque et extrinsèque de ces ressources, utilisées pour le développement d'outils d'annotation automatique, montrent l'intérêt d'utiliser la myriadisation dans un cadre linguistique non standardisé : les locuteurs ne sont pas ici considérés comme un ensemble uniforme de contributeurs dont les efforts cumulés permettent d'achever une tâche particulière, mais comme un ensemble de détenteurs de connaissances complémentaires. Les ressources qu'ils produisent collectivement permettent de développer des outils plus robustes à la variation rencontrée.
Les plateformes développées, les ressources langagières, ainsi que les modèles de *taggers* entraînés sont librement disponibles.

Mots-clés : Myriadisation ; Traitement automatique des langues ; Langues peu dotées ; Langue non standardisées ; Corpus annoté ; morphosyntaxe ; annotation manuelle

Crowdsourcing linguistic resources for natural non-standardised languages processing

Summary

Citizen science, in particular voluntary crowdsourcing, represents a little experimented solution to produce language resources for some languages which are still little resourced despite the presence of sufficient speakers online.
We present in this work the experiments we have led to enable the crowdsourcing of linguistic resources for the development of automatic part-of-speech annotation tools. We have applied the methodology to three non-standardised languages, namely Alsatian, Guadeloupean Creole and Mauritian Creole. For different historical reasons, multiple (ortho)-graphic practices coexist for these three languages. The difficulties encountered by the presence of this variation phenomenon led us to propose various crowdsourcing tasks that allow the collection of raw corpora, part-of-speech annotations, and graphic variants.
The intrinsic and extrinsic analysis of these resources, used for the development of automatic annotation tools, show the interest of using crowdsourcing in a non-standardized linguistic framework: the participants are not seen in this context a uniform set of contributors whose cumulative efforts allow the completion of a particular task, but rather as a set of holders of complementary knowledge. The resources they collectively produce make possible the development of tools that embrace the variation.
The platforms developed, the language resources, as well as the models of trained taggers are freely available.

Keywords : Crowdsourcing ; Natural language processing ; Less-resourced languages ; Non-standardized languages ; Annotated corpora ; Part-of-speech ; Manual annotation

UNIVERSITÉ SORBONNE UNIVERSITÉ

ÉCOLE DOCTORALE :
École Doctorale V – Concepts et langage
Maison de la Recherche, 28 rue Serpente, 75006 Paris, FRANCE

DISCIPLINE : Mathématiques, informatique et applications aux sciences de l'Homme