Linguistic Diversity in Natural Language Processing

Aarne Ranta* — **Cyril Goutte****

* *University of Gothenburg, Department of Computer Science and Engineering,* Aarne.Ranta@cse.gu.se
** *Conseil national de recherches Canada, Technologies numériques,* Cyril.Goutte@nrc.ca

ABSTRACT. *Although computational linguistics carries the promise of producing tools for processing and understanding a wide variety of languages, most of the work in NLP still focuses on a small number of languages, and in particular on English. The goal of this special issue is to promote linguistic diversity in NLP, by encouraging the publication of work on languages or language varieties less often studied, as well as methods that can easily and demonstrably be applied to those. Two articles are included in this special issue, one on language identification for building a resource for the Corsican language, the other on machine translation of two indigenous languages of northern Canada.*

RÉSUMÉ. *Bien que la linguistique informatique porte en elle la promesse d'outils aidant au traitement et à la compréhension d'une multitude de langues, la majorité des travaux en TAL porte encore sur un petit nombre de langues, et en particulier sur l'anglais. L'objectif de ce numéro spécial est de promouvoir la diversité linguistique en TAL en encourageant la présentation de travaux portant sur des langues ou variantes de langues moins souvent traitées, ainsi que sur des méthodes qui peuvent être aisément appliquées à celles-ci. Deux articles sont inclus dans ce numéro, l'un sur l'identification de la langue pour constituer une ressource pour la langue corse, l'autre sur la traduction de deux langues autochtones du nord du Canada.*

KEYWORDS: *linguistic diversity, less studied languages.*

MOTS-CLÉS : *diversité linguistique, langues peu traitées.*

1. Introduction

Research in Natural Language Processing (NLP) has largely focused on building various methods, models and tools for handling human language. From its original goal of giving computers the ability to understand and communicate with humans using spoken and written language interaction, it has naturally focused on languages researchers were most familiar with, and in particular on English. The rise of methods based on statistical learning and their reliance on significant amounts of linguistic resources has increased this trend, which the return of neural methods and move to deep learning has further reinforced.

Natural Language Processing systems, especially when they are developed using machine-learning-based techniques, have sometimes been claimed to be language agnostic, suggesting that expanding to more diverse languages may simply be a matter of retraining models on appropriate resources in the target language. However, NLP technology is typically developed on a handful of dominant languages which are sometimes related—when it is not created simply on English. It has been argued (Bender, 2011) that achieving genuine language independence requires a level of linguistic sophistication that is normally not included in NLP systems. A recent study of linguistic diversity (Joshi *et al.*, 2020) suggested a gradation of 6 language groups, from the virtually ignored, to the most dominant. The top two groups, on which most of the NLP work is performed, cover more than 4 billion speakers, but comprise only 25 languages, about 1% of the total number of languages considered (and significantly less than 0.5% of the about 7,000 human languages currently spoken). It is therefore conceivable that many of the linguistic features present in the 99+% of remaining languages are not considered and may pose significant, unforeseen challenges to methods in mainstream NLP. This raises the interesting question of how to complement the mainly computational concern of how methods scale up to more data, with the more pragmatic linguistic concern of how to work on more languages.

As textual resources are critical to feed many of the data-hungry statistical and neural methods, the limited availability of such resources for the vast majority of languages also creates challenges for linguistic diversity in NLP. Low resource NLP was the topic of a recent special issue of this journal (Bernhard and Soria, 2018) and many challenges were described and addressed there: obviously, the lack of resources, but also the heterogeneity both in terms of genre, time or topic, as well as the linguistic heterogeneity due to lack of language normalization or code mixing. These often lead to concerns with the quality of the resources, in addition to their quantity (Caswell *et al.*, 2022).

An additional and significant challenge for linguistic diversity is that it is often difficult to publish work performed on languages other than English. The prevalence of English NLP in the academic literature has long been supported by anecdotal accounts (Munroe, 2015; Mielke, 2016). One can easily speculate over the reasons for this situation. The availability of resources and benchmarks in English probably plays an important role, as it makes it easier to tackle an existing task and show progress using

a proposed new method. This is also due to a clear bias in the perception and assessment of novelty in our field. Novelty is one of the key criterion in many peer-review process, and there is a stronger focus of methodological novelty, while language novelty is typically assessed as "just applying an existing method to a new language".

The goal of this special issue is to favour language diversity in natural language processing by offering a venue for publishing this type of work. We believe this is a timely topic as well. The special theme track for the 2022 conference of the Association for Computational Linguistics is: "Language Diversity: from Low-Resource to Endangered Languages" (ACL, 2022), indicating that the concerns expressed above are shared by the most prominent organization in the field.

2. Summary of the Contributions

This special issue contains two articles addressing very different aspects of language diversity. The first one focuses on resource acquisition and processing tool creation with limited data for that purpose—in that specific case language identification tools. The second paper addresses the issue of building Machine Translation systems, and more specifically the challenges arising from the morphological complexity of polysynthetic languages.

2.1. *L'identification de langue, un outil au service du corse et de l'évaluation des ressources linguistiques*

The first article in this special issue deals with the topic of language identification for Corsican, a language considered endangered by UNESCO. Language identification is a task that is doubly relevant to the topic of this special issue, and offers both challenges and opportunities. First, because although it is a well-known task that has reached near-perfect performance on many languages, it is still challenging in particular when little material is available for training. Secondly, because it is a key language processing tool to filter and identify language-appropriate material in a large collection of documents, in order to build resources for less-studied languages. The paper explores both aspects, adapting and testing a large number of language identifiers on Corsican, and exploring the use of several of these tools to process existing linguistic resources.

2.2. *Towards a Low-Resource Neural Machine Translation for Indigenous Languages in Canada*

The second article is about machine translation for two indigenous languages: Inuktitut and Inuinnaqtun. Inuktitut has almost 40,000 speakers and an official status in Nunavut in Canada, whereas Inuinnaqtun is an endangered language with less than 1,000 native speakers. One obvious challenge of the project is, in particular for

Inuinnaqtun, the scarcity of data. The main focus of the paper is, however, on the polysynthetic nature of both of the languages, requiring a substantial effort in morphological segmentation as preprocessing of machine translation. The outcome is a thorough analysis and evaluation of different methods of segmentation, enabling a neural machine translation system for English-Inuktitut that outperforms the previous state of the art.

Acknowledgements

We wish to thank the editorial committee of the TAL journal for suggesting the topic of this special issue and inviting us to coordinate its scientific committee. We wish to thank more specifically the editors-in-chief for their always patient support during this process and in particular Emmanuel Morin for his invaluable help with SciencesConf. We are indebted to the reviewers and members of the scientific committee who accepted to join us for this special issue and volunteered their time in order to help us select the articles published here: Laurent Besacier (Naver Labs, France), Marine Carpuat (University of Maryland, USA), Leila Kosseim (Concordia University, Canada), Mathieu Mangeot (Université Savoie Mont Blanc, France), Yannick Parmentier (Université de Lorraine, France), Yves Scherrer (University of Helsinki, Finland), and Francis Tyers (Indiana University, USA).

3. References

ACL, "ACL 2022 Theme Track: 'Language Diversity: from Low-Resource to Endangered Languages'", https://www.2022.aclweb.org/post/acl-2022-theme-track-language-diversity-from-low-resource-to-endangered-languages, 2022 (visited February 2022).

Bender E. M., "On Achieving and Evaluating Language-Independence in NLP", *Linguistic Issues in Language Technology*, vol. 6, 2011.

Bernhard D., Soria C., "Traitement automatique des langues peu dotées", *Traitement automatique des langues*, vol. 59, no. 3, 2018.

Caswell I., Kreutzer J., Wang L., Wahab A., van Esch D., Ulzii-Orshikh N., Tapo A. A., Subramani N., Sokolov A., Sikasote C., Setyawan M., Sarin S., Samb S., Sagot B., Rivera C., Gonzales A. R., Papadimitriou I., Osei S., Suarez P. O., Orife I., Ogueji K., Niyongabo R. A., Nguyen T. Q., Muller M., Muller A., Muhammad S. H., Muhammad N. F., Mnyakeni A., Mirzakhalov J., Matangira T., Leong C., Lawson N., Kudugunta S., Jernite Y., Jenny M., Firat O., Dossou B. F. P., Dlamini S., de Silva N., Çabuk Balli S., Biderman S. R., Battisti A., Baruwa A., Bapna A., Baljekar P. N., Azime I. A., Awokoya A., Ataman D., Ahia O., Ahia O., Agrawal S., Adeyemi M., "Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets", *Transactions of the Association for Computational Linguistics*, vol. 10, p. 50-72, 2022.

Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M., "The State and Fate of Linguistic Diversity and Inclusion in the NLP World", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 6282-6293, 2020.

Mielke S. J., “Language diversity in ACL 2004 – 2016”, https://sjmielke.com/acl-language-diversity.htm, December 2016 (visited February 2022).

Munroe R., “Languages at ACL this year”, http://www.junglelightspeed.com/languages-at-acl-this-year/, July 2015 (visited February 2022).