DiLiTAL – Diversité linguistique et TAL

Fadoua Atta-Allah[1], Fatima Agnaou[2], Khalid Ansar[3], Aicha Bouhjar[2], Siham Boulaknadel[1], Malika Chakiri[4], Hammou Fadili[5], Jamal Frain[1], Jovan Kostov[6, 7], Alice Millour[8], Satenik Mkhitaryan[9], Michael Zock[10]

(1) CEISIC, IRCAM, Rabat, Maroc
(2) CRDPP, IRCAM, Rabat, Maroc
(3) CAL, IRCAM, Rabat, Maroc
(4) Université Paris Descartes, 12, rue de l'Ecole de Médecine, 75006, Paris
(5) Laboratoire CEDRIC - CNAM, 2, rue Conté, 75003, Paris
(6) E.A. 4154 PLIDAM - INALCO, 2, rue de Lille, 75007, Paris
(7) PERL - USPC, 8, place Paul Ricœur, 75013, Paris
(8) E.A. 4509 STIH – Université Paris Sorbonne, 1, rue Victor Cousin, 75005, Paris
(9) E.A. 2520 ERTIM – INALCO, 2, rue de Lille, 75007, Paris
(10) LIF – UMR 7279, 163, avenue de Luminy, F-13288, Marseille

{agnaou,ansar,ataaallah,bouhjar,boulaknadel,frain}@ircam.ma, chakirimalika@yahoo.fr, jovan.kostov@gmail.com, alice.millour@paris-sorbonne.fr, satenik.mkhitaryan@inalco.fr, mikael.zock@lif.univ-mrs.fr

RESUME

L'atelier DiLiTAL se donne pour objectif de sensibiliser la communauté scientifique non seulement aux enjeux et aux difficultés rencontrées dans le traitement des langues peu dotées (LPD), mais aussi à l'intérêt de créer des outils génériques permettant de traiter un très grand nombre de langues tout en identifiant les besoins spécifiques en fonction de leurs particularités. Cet atelier vise à mettre en commun les différentes méthodes et techniques utilisées pour dynamiser la construction et la mutualisation des ressources, ainsi que le transfert des savoirs et des savoir-faire.

ABSTRACT

Workshop Linguistic diversity and NLP - DiLiTAL
The goal of the DiLiTAL workshop is to raise awareness among the scientific community of the issues and difficulties encountered in the processing of under-resourced languages (ULL). We also want to point out the necessity of creating generic tools in order to process a large number of languages by identifying their particularities. This workshop's aim is to pool the different methods and techniques used to boost the construction of resources, as well as the transfer of knowledge and experience in the processing of under-resourced languages.

MOTS-CLES : langues peu dotées, diversité, linguistique, TAL.
KEYWORDS: less resourced languages, diversity, linguistics, NLP.

1 Présentation générale de l'atelier

Savoir communiquer est l'un des fondements de nos sociétés. Nos survies et notre coexistence son intimement liées à notre faculté de nous faire comprendre, ce qui peut poser problème à cause de nos différences culturelles et linguistiques et de nos expériences de vie, sensibilités, points de vue etc. Dans de très nombreux pays, on pratique plusieurs langues dont le statut est différent : langue officielle, langue régionale etc. Cet état engendre une grande disparité au niveau des outils et des ressources en traitement automatique des langues (TAL). Dans ce domaine, les langues « mineures/minorées » sont communément connues sous le nom de « langues peu dotées » (LPD). On note également que les décideurs s'y intéressent généralement peu, ce qui est un handicap incontestable à l'heure de la globalisation.

Cet atelier vise à susciter une réflexion débouchant sur un élargissement des travaux en TAL pour prendre en compte d'autres langues que celles habituellement traitées. L'attention de l'atelier DiLiTAL est portée sur la nature des outils et des ressources et sur la manière de les adapter aux LPD, mais aussi sur la question de formation de spécialistes capables de les élaborer, de les pérenniser et de les réutiliser dans une visée d'enseignement et de diffusion de ces langues, selon les initiatives mises en place par les grandes organisations internationales (Nations-Unies, Conseil de l'Europe, etc.).

La nécessité de traiter automatiquement les LPD découle des besoins à la fois scientifiques et humanitaires (santé, éducation, culture, littératie, etc.), mais également des enjeux d'ordre politique (accès à l'information et à l'enseignement). L'atelier DiLiTAL se donne pour objectif de sensibiliser la communauté scientifique à ces enjeux et aux difficultés rencontrées dans le traitement des LPD, mais aussi à l'intérêt de créer des outils génériques permettant de traiter un très grand nombre de langues, tout en identifiant les besoins spécifiques en fonction des particularités des différentes langues (typologie).

2 Axes thématiques de l'atelier DiLiTAL

L'atelier DiLiTAL a réuni des interventions qui présentent des travaux théoriques et/ou des applications concrètes qui s'articulent autour des thématiques suivantes :

2.1 Ressources et corpus : production, standardisation et archivage

L'ambition de cet axe est d'explorer les initiatives actuelles et futures qui ont pour but de collecter et de structurer des données langagières (spécialisées ou généralistes) des LPD. Ces données (lexiques, corpus etc.) sont souvent utilisées pour l'entrainement des étiqueteurs morphosyntaxiques qui représentent, à leur tour, une étape préalable à des tâchas plus complexes comme l'analyse syntaxique ou la traduction. Nous nous interrogeons également sur l'accessibilité » et la portabilité des données linguistiques, qui s'avèrent être des problèmes majeurs dans les travaux consacrés aux LPD et c'est pour cette raisons que les contributions concernent principalement la création des ressources *open source*.

2.2 Outils pour le traitement des LPD

Les interrogations de ce second axe portent sur la pertinence de l'utilisation d'outils existants pour les LPD et sur la manière dont ils gèrent le multilinguisme. En effet, depuis l'arrivée d'UTF-8 (Unicode), le TAL s'est doté de la possibilité de diversifier son terrain d'action-recherche permettant de traiter d'autres langues que celles dites « majoritaires » (et bien dotées), comme l'anglais, l'espagnol, le français, le chinois etc. Dans une telle perspective, il nous a semblé nécessaire de réfléchir à une amélioration de la gestion du multilinguisme par l'inclusion de nouvelles graphies et de nouveaux standards, tout en identifiant les contraintes méthodologiques auxquelles se heurtent les chercheurs dans le cadre d'un travail de recherche sur une LPD avec des outils existants.

2.3 Questions sociales, culturelles et éthiques

DiLiTAL a été également une occasion d'examiner les possibilités de coopération transfrontalière pour encourager l'échange d'expériences et pour dynamiser la recherche en TAL dans les pays où les LPD sont en usage. Il nous semble important d'entamer une réflexion concernant des aspects éthiques de la collecte, de l'archivage et du traitement des données linguistiques qui, eux, se trouvent ancrés dans des contextes sociétaux où les codes et les mœurs sont méconnus ou sensiblement différents de ceux du chercheur. Le but principal de cet axe est de réfléchir à une méthodologie de terrain qui ne vise pas uniquement à ménager la science, mais qui prend aussi en compte le contexte global, à savoir, le sujet parlant qui fait partie d'une culture, voire d'une catégorie sociale. Cet atelier a été l'occasion de discuter du TAL en termes de moteur des changements linguistiques. En effet, ce problème se pose dans le cadre des commissions terminologiques qui jouent un rôle prépondérant dans le processus de codification des langues dans différents pays du monde.

2.4 Retours d'expérience

Certaines contributions ont mis en lumière des expériences de traitement d'une LPD ou d'un groupe de LPD: la synthèse des expertises acquises par des chercheurs ayant participé à ce genre de travaux permet de nourrir une réflexion épistémologique nécessaire au développement des outils pour dépasser le cadre d'un seul type de langues. Ces expériences s'avèrent utiles pour identifier les domaines auxquels la recherche en TAL pourra contribuer. Quelle que soit la langue, un retour d'expérience est toujours précieux pour permettre d'identifier les convergences et les divergences dans les approches, constituant, de ce fait, un apport considérable pour cet atelier et pour la recherche-action qui en découlera.

L'atelier comprend sept interventions reflétant les recherches menées sur les LPD dans un contexte francophone et amazigh. Une conférence invitée (donnée par Joseph Mariani, LIMSI - CNRS) et une table ronde ont été également organisées dans le but de rendre compte des recherches sur les LPD dans le contexte scientifique actuel.