Projets Professionnels - Master IdL 2
2023-2024

Enseignant - coordinateur : Thomas Lebarbé
Stendhal, Université Grenoble Alpes

Table des matières

1 Préambule

Dans les pages qui suivent vous trouverez une liste de projets proposés par des enseignants, des chercheurs ou des partenaires en lien avec le traitement automatique des langues. Ils revêtiront le rôle de commanditaires et ne sont pas là pour vous aider à mener à bien votre projet.

Quand vous aurez sélectionné le projet qui vous convient (thématique, compétence linguistique, didactique et technologique), vous prendrez contact avec votre commanditaire afin de définir avec plus de précision ses attentes. Vous lui remettrez un cahier des charges qu'il validera (ou pas). A partir de la validation, le commanditaire vous aura fourni tout le matériau dont vous aurez besoin et il ne pourra en aucun cas revenir sur le cahier des charges.

Sauf exception négociée avec le responsable de cours (cas de nombre impair d'étudiants dans le cours), les projets sont réalisés en binôme. Vous vous assurerez d'une répartition équitable du travail. La composition des binômes ainsi que les projets choisis seront indiqués par mail à l'enseignant coordinateur au plus tard à la date de choix des projets. Aucun changement ne sera possible au delà de cette date.

2 Calendrier

12/10/2023	Présentation des projets
19/10/2023	Choix des projets par les étudiants.
14/12/2023	Soumission des cahiers des charges.
21/12/2023	Acceptation des cahiers des charges par les commanditaires.
23/02/2024	Rendu des projets, documentations et manuels.
29/02/2024	Soutenance publique.

Les dates pour le rendu et la soutenance sont sujettes à modification selon les départs en stage (théoriquement au 1er mars). En cas de force majeure, vous pourrez éventuellement négocier des changements de calendrier avec les commanditaires (en informant le responsable de cours).

3 Évaluation

Le travail de projet professionnel est évalué par équipe en deux temps :

— Semestre 1 : travail avec le(s) commanditaire et cahier des charges,
-> 3 ECTS

— Semestre 2 : production ($\frac{1}{3}$), documents ($\frac{1}{3}$), soutenance ($\frac{1}{3}$),
-> 10 ECTS

4 Projets

4.1 Le Lexique de Makki en ligne

Commanditaire :	Mme Salam Diab Duranton, LIDILEM
	`salam.diab-duranton@univ-grenoble-alpes.fr`

Le lexique de Makki est une compilation lexicale conséquente (3200 pages dactylographiées environ). Elle porte sur le parler libanais d'une manière générale, mais plus spécifiquement celui du grand Sud. Elle a été collectée par un ingénieur libanais, Hassan Makki, originaire de Erkay, dans le sud-est de la ville de Saïda, sur une quinzaine d'année à partir de la fin des années 1990. A ce jour il prend la forme d'un fichier de traitement de texte en 4 colonnes : forme arabe classique, forme arabe libanaise, forme française et transcription phonétique.

Dans une perspective de publication en ligne, de patrimonialisation d'un fond documentaire inégalé, et en tenant compte de l'évolution permanente de cette ressource, nous souhaitons faire développer une plateforme en ligne.

L'objectif du projet est :

- développer un dispositif numérique définissant une structure de base de données ainsi que des interfaces de consultation, ajout et modification
- transformer la donner du document de traitement de texte en donnée SQL intégrée à la base de
- concevoir une procédure d'export de la donnée dans un format interopérable (TEI P5, chapitre Dictionaries)
- dans une philosophie Science Ouverte, le projet devra anticiper des évolutions aussi bien du point de vue de la structure de donnée que de l'interface, notamment dans des perspectives de recherche en linguistique et sociolinguistique.

4.2 Data CNIL

Commanditaire :	Mme Emilie Masson, DPO Adjointe INRIA emilie.masson@inria.fr

La CNIL (Commission Nationale Informatique et Libertés) a l'obligation, depuis la loi République Numérique, de publier le résultat de ses délibérations en ce qui concerne l'application de la loi Informatique et Libertés et du RGPD. Ces données sont disponibles sur les serveurs de l'état [1], mais difficilement manipulable.

Un premier prototype [2] a été développé. Outre la nécessité de revoir l'ensemble du programme pour construire un outil solide, les points suivants sont, notamment, à prendre en charge dans le cadre de cette commande :

- Intégrer une mise à jour automatisée de la base de données en fonction des dernières publications sur les serveurs ;
- Intégrer une correction automatique des erreurs de balisage (dans les fichiers originaux)
- Intégrer un moteur de recherche performant ;
- Permettre la recherche d'expression complète par exemple « entrepôt de données de santé »
- Prévoir la de sauvegarde et l'export des résultats
- Proposer et fournir des outils TAL et des représentations dynamiques permettant d'explorer l'ensemble des délibérations (à ce jour, environ 25.000) entre distant reading et close reading ;

1. https://www.data.gouv.fr/fr/datasets/les-deliberations-de-la-cnil/
2. http://i3l.univ-grenoble-alpes.fr/~lebarbet/DataCNIL

4.3 Évaluation des outils de traitement automatique de la parole utilisés dans le projet DyLNet

Commanditaire :	Mme Aurélie Nardy, LIDILEM aurelie.nardy@univ-grenoble-alpes.fr Mme Isabelle Rousset, LIDILEM isabelle.rousset@univ-grenoble-alpes.fr Mme Solange Rossato, LIG solange.rossato@univ-grenoble-alpes.fr Mme Solène Evain, LIG solene.evain@univ-grenoble-alpes.fr

Contexte :

Dans le cadre du projet DyLNet (Dynamiques langagières, apprentissages linguistiques et sociabilité à l'école maternelle : apport des capteurs de proximité pour le recueil de données massives [3] , coordonné par Aurélie Nardy (Laboratoire Lidilem, UGA) et financé par l'Agence Nationale de la Recherche (ANR-16-CE28-0013), les enfants et adultes d'une école maternelle ont été équipés, une semaine par mois, pendant 2 ans et demi, de capteurs enregistrant à la fois les proximités entre individus et leur voix.

La finalité de ce projet est d'examiner comment les relations qu'entretiennent les enfants à l'école maternelle influencent le développement de leur langage oral (comment les enfants s'influencent les uns les autres et comment les adultes présents dans l'école ' enseignants et ATSEM ' influencent le langage enfantin).

Environ 45 000 heures d'enregistrements audio ont été recueillis (fichiers audios d'une heure maximum). Parmi elles, 800 heures ont été segmentées, annotées et transcrites avec le logiciel ELAN [4].

La mise en place de ce projet nécessite de pouvoir exploiter de manière automatique cette importante masse d'enregistrements. Le projet pro proposé contribue à cette automatisation des traitements en évaluant deux outils de traitement automatique de la parole : la segmentation et la transcription automatique de parole.

Point d'attention :

Le protocole de cette étude a été soumis au Comité Opérationnel d'Évaluation des Risques Légaux et Éthiques de l'INRIA et de la Commission Nationale de l'Informatique et des Libertés qui l'ont validé. Dans ce cadre,

3. https://dylnet.univ-grenoble-alpes.fr
4. https://tla.mpi.nl/tools/tla-tools/elan/download/

des mesures strictes garantissant la confidentialité et l'anonymat des participants ont été prises. À ce titre, toute personne en contact avec des données issues de l'étude s'engage à respecter ces mesures et signera un engagement de confidentialité.

Objectif :

L'objectif de ce projet pro est de procéder à l'évaluation des outils automatiques de traitement de la parole mis en place jusqu'ici dans le cadre du projet DyLNet. Il se décline en 3 modules :

1. Module segmentation : évaluation de la détection automatique de la voix du porteur du micro
   Un système de détection automatique de la parole du porteur du capteur a été mis en place. À partir d'un corpus de référence d'environ 800 heures d'enregistrements segmentées par des humains (temps de parole/silence), il s'agira de procéder à une évaluation de l'outil de segmentation automatique à partir d'une ou plusieurs mesures pertinentes. Il s'agira ainsi dans un premier temps de déterminer quel type de métrique peut être utilisé avant d'implémenter sa mise en 'uvre et de proposer une interprétation de cette métrique.
2. Module RAP : évaluation d'un système de RAP
   Dans le cadre d'un travail mené en Reconnaissance Automatique de la Parole (RAP) qui a consisté à utiliser un système end-to-end sur des représentations issues de modèles faiblement supervisés de type wav2vec et adapté sur les données de DylNet, une transcription automatique des fichiers audios a été proposée. Les WER (Word Error Rates) sont très variés. Il s'agira ici d'établir une procédure permettant une évaluation ' en fonction du locuteur, de la situation (classe, récréation, sport, ') ' et de cibler les cas posant difficulté à l'outil.
3. Évaluation de la chaîne de traitement complète
   Il s'agira d'implémenter la chaîne de traitement sur les fichiers dont nous avons la transcription manuelle pour évaluer la chaîne complète de traitement. L'objectif est de savoir dans quelle mesure cette chaîne de traitement peut être étendue aux 45 000 h d'enregistrements.

4.4 JADE distant & close reading

Commanditaire : M. Romain Rambaud, CRJ
romain.rambaud@univ-grenoble-alpes.fr

Le projet proposé s'inscrit dans le cadre du projet de recherche IDEX JADE. Le développement des recherches fondamentales concernant l'intelligence artificielle et ses champs d'application, dans une perspective interdisciplinaire, constitue une priorité scientifique. Du point de vue des sciences juridiques, l'un des axes prioritaires de ces recherches concerne l'utilisation de l'IA en matière de justice. Le projet interdisciplinaire « Justice algorithmique des élections » (JADE), porté par le Centre de recherches juridiques (CRJ), le Laboratoire Jean Kuntzmann (LJK), le laboratoire d'informatique de Grenoble (LIG), avec le soutien de PACTE et de la chaire de société algorithmique du MIAI (Multidisciplinary Institute in Artificial intelligence), a pour objet d'utiliser des méthodes de mathématiques appliquées et d'IA à un objet juridique pour lequel elles n'ont jamais été utilisées et présentent un intérêt particulier, le contentieux des élections politiques. L'hypothèse retenue par le projet JADE est que la justice algorithmique pourrait être pertinente en contentieux électoral, en tant d'une part qu'elle pourrait améliorer la compréhension de sa rationalité (objectif de connaissance) et d'autre part faciliter son application par les citoyens, les avocats et les juges (objectif de prévision), par exemple par l'intermédiaire d'un logiciel. Le point focal de la problématique en contentieux électoral est l'analyse de la sincérité du scrutin, c'est-à-dire la question de savoir si des irrégularités ont pu avoir un effet sur le résultat de l'élection, et en pratique l'un des principaux critères utilisés pour le déterminer est l'écart de voix entre les candidats ou les listes. Si le contentieux électoral ne veut pas reposer sur la seule intuition, il faudrait donc qu'il existe une détermination scientifique de l'écart de voix pertinent en fonction des irrégularités. Pourtant, les travaux juridiques qui ont cherché à systématiser la question sont peu nombreux et parfois contradictoires.

La méthodologie utilisée est pluridisciplinaire, de type humanités numériques, à titre principal juridique et informatique, visant les usages davantage que la régulation de l'IA en droit. La construction d'algorithmes se décompose en général en deux étapes : la constitution d'une base de données et leur analyse par des logiciels. D'ici décembre 2023, un corpus de décisions du constitutionnel sera annoté par des étudiants de droits afin de mettre en évidence les éléments juridiques pertinents sur lesquels s'appuieront les outils d'IA.

De ce point de vue, le projet a besoin de la compétence d'étudiants de Master 2 - Industries de la Langue : il s'agit en effet d'offrir une perspective différente des inférences sur les données sources et de proposer des modalités de représentation du corpus afin de faciliter une démarche hypothético-déductive par close et distant reading. Les responsables du projets attendent donc :

- une interface cartographique sur le corpus (par circonscription)
- des systèmes de filtrage des décisions par temporalité (date des décisions), thématiques (données par le commanditaire et annotées dans le corpus)
- une représentation graphique par proximité des décisions.

4.5 BD-E-Calm

Commanditaire :	M. Claude Ponton, LIDILEM `claude.ponton@univ-grenoble-alpes.fr`

Dans le cadre du projet E-Calm, on dispose d'un corpus de transcription de textes d'élèves en XML-TEI. Ces transcriptions répondent à une DTD spécifique.

L'objectif du projet est :

— D'alimenter automatiquement une base de données à partir d'indicateurs calculés sur les fichiers XML (ex. longueur du texte, nombre de ratures, présence ou non de commentaires enseignants, etc.)

— De proposer une interface Web permettant d'interroger cette base de données et d'afficher la liste des fichiers XML correspondants à la requête