Master en Traitement Automatique des Langues
2020-2021

Proposition du projet tutoré :
Création d'une base de données
dans le cadre du projet Lex.E.M. (Lexique pour les Écoles Maternelles)
https://projetlexem.wixsite.com/website

1 Encadrement

Encadrant: PETITJEAN Etienne
Equipe et laboratoire: ATILF, équipe *Soutien Technique à la Recherche*
Contact: etienne.petitjean@atilf.fr
Responsable du Service Informatique

Co-encadrant: KNITTEL Marie Laurence
Equipe et laboratoire: ATILF, équipe *Lexique*
Contact : marie-laurence.knittel@univ-lorraine.fr

Co-encadrant: RUVOLETTO Samantha
Equipe et laboratoire: ATILF, équipe *Discours*
Contact : samantha.ruvoletto@univ-lorraine.fr

2 Motivation et contexte

Le projet Lex.E.M (Lexique pour les Écoles Maternelles) s'inscrit dans le cadre d'une collaboration avec des écoles maternelles de REP (Réseau d'Éducation Prioritaire) et concerne les enfants de 2 à 3 ans (classes de Tout-Petits et Petits).
Les enseignants ont repéré chez ces enfants (francophones ou allophones) des lacunes importantes au niveau lexical, y compris concernant le vocabulaire basique utilisé à l'école. Ces lacunes se situent à la fois en compréhension et en production.
Afin d'élaborer des outils de remédiation favorisant l'accès au lexique des enfants de 2-3 ans, il est nécessaire de connaître avec précision le vocabulaire basique employé par les enfants de milieu ordinaire à l'école maternelle. Or, il n'existe à ce jour aucun corpus de productions langagières d'enfants en milieu scolaire. La seule base de données existante est la base Manulex [LSC04], constituée à partir du lexique des manuels scolaires de l'école primaire (CP à

CM2). Il s'agit donc d'un lexique écrit, destiné à un public différent, et potentiellement éloigné des productions réelles des enfants.

De ce fait, le projet Lex.E.M (Lexique pour les Ecoles Maternelles) a pour objectifs :

1. de créer un nouveau corpus, répertoriant le vocabulaire basique utilisé à l'école maternelle,
2. d'élaborer une base de données, fondée sur les productions réelles des enfants et des interactions adultes/enfants.
3. de créer une interface consultable sur le web.
4. de développer des outils pour la remédiation, spécifiquement destinés à un public de 2-3 ans.

=> Le projet tutoré consistera en l'élaboration de la base de données (point 2). Il intégrera une réflexion linguistique à des compétences en traitement automatique des langues.

3 Objectifs

Ce projet a pour but la création de la base de données LEX.E.M. Cette création prévoit plusieurs étapes qui seront le sujet du projet tutoré.

1. Familiarisation avec les corpus existants et homogénéisation des formats

La base de données sera créée à partir de 3 corpus :

- Child Language Data Exchange System (CHILDES) [MCW00] : https://childes.talkbank.org/.
- Communication Langagière chez le Jeune Enfant (Colaje) : http://colaje.scicog.fr/index.php/corpus
- Corpus Jeunes Enfants en Milieu Scolaire (JEMS), élaboré à partir d'enregistrements inédits en milieu scolaire et transcrit par des stagiaires de Sciences du Langage pendant l'année universitaire 2019/2020.

Tous ces corpus présentent la production de l'adulte (nommé avec le même sigle si même participant, par exemple si « maman » toujours « MOT ») et de l'enfant (nommé toujours « CHI ») et sont disponibles en format XML (voir en annexe un exemple de corpus transcrit).

Corpus CHILDES :

- Certains corpus sont téléchargeables à partir de *Phonbank* en version XML (https://phonbank.talkbank.org/). Ces corpus présentent la transcription orthographique et phonétique des productions de l'enfant et de l'adulte. La transcription phonétique utilisant des caractères API se présentent en deux lignes :
    - - *Tier IPA Target* qui présente la cible phonétique (production standard attendue dans la langue de référence, ex. *parce que* \paʁs.kə\) ;
    - *Tier IPA Actual* qui montre la véritable réalisation phonétique (ex. pour *parce que* -> \pa.kə\).

- D'autres corpus sont disponibles sur CHILDES à partir de *talkbank*. Ces corpus sont en version « chat » (.*cha*) et présentent parfois la situation de communication (ligne %act), mais jamais la transcription phonétique. Ils doivent être transformés en version XML (https://talkbank.org/software/chatter.html).

Corpus COLAJE : les corpus sont disponibles en version XML mais ne comportent pas d'informations phonétiques.

Corpus JEMS : il comporte toutes les informations présentes dans le corpus CHILDES, disponibles sur *Phonbank*.

Après s'être familiarisé avec les corpus disponibles, le stagiaire aura pour tâche d'extraire les fichiers XML ou de transformer le fichier .cha en XML. Le but de cette étape et de disposer d'une base commune et homogène nécessaire pour ensuite procéder à l'extraction des formes lexicales (=lemmes) via un étiquetage.

2. Récupération des formes lexicales et étiquetage morphosyntaxique

Pour constituer la base de données, il faudra d'abord récupérer les lemmes produits par les adultes et les enfants à partir de la forme orthographique attendue sur les 3 corpus *CHILDES*, *COLAJE, JEMS.*

Pour faire cela, un étiquetage en morphosyntaxe sera nécessaire. Nous disposons des données issues de corpus oraux. L'utilisation d'étiqueteurs automatiques élaborés pour et à partir de données écrites n'est pas une solution optimale étant données les particularités des corpus oraux par rapport à l'écrit [BFS12]. **Nous proposons donc une pré-annotation automatique suivie d'un contrôle/annotation manuelle.**

Pour les pré-annotations, le stagiaire utilisera le système TreeTagger [Sch97]**, qui fournit pour chaque token d'entrée une étiquette morphosyntaxique et un lemme.**

3. Ajout de données manquantes et calcul des indices de fréquence

Des informations linguistiques seront associées aux lemmes répertoriés. Certaines informations sont disponibles à partir des 3 corpus. Le stagiaire associera à chaque production lexicale les informations suivantes :

a. Qui a produit le mot dans les corpus (adultes vs enfants)
b. L'âge du participant qui a produit le mot
c. La forme phonétique attendue (*Actual*) et produite (*Target*, si présent)
d. La situation de communication si présente (ligne %act).
e. Le tour de parole qui précède et qui suit ainsi que le participant qui a produit ces séquences

f. La catégorie morphosyntaxique de chaque item (étiquetage)
g. Le nom du corpus spécifique (fichier XML)
h. Le nom du corpus général (CHILDES ou COLAJE ou JEMS)

Pour compléter la base de données, il faudra intégrer des informations linguistiques complémentaires qui ne sont pas toujours disponibles dans les corpus. Pour faire cela, le stagiaire utilisera d'autres bases de données linguistiques existante et consultable en ligne. Il s'agira de :

i. Ajouter la forme phonétique des productions lexicales à partir de http://www.lexique.org/ [NPFM01] si elle n'est pas présente.

j. Ajouter la structure syllabique à partir de http://www.lexique.org/.

Dans cette base de données, plusieurs indices de fréquence seront associés à chaque lemme. Un travail de comptage, standardisation et une analyse de distribution seront nécessaires pour déterminer la valeur de ces indices.

Nous envisagions d'associer à chaque lemme :

k. La fréquence absolue dans le corpus spécifique d'appartenance (fichier XML)
l. La fréquence absolue dans le corpus général (CHILDES ou COLAJE ou JEMS)
m. La fréquence absolue dans les 3 corpus (CHILDES, COLAJE et JEMS)
n. La fréquence dans la base de données de l'écrit MANULEX (http://www.manulex.org/fr/home.html)
o. La fréquence dans la base de données LEXIQUE.org http://www.lexique.org/.

4. Rédaction d'un tableau de consultation rapide de la base

La création de la base de données précède la constitution d'une interface d'utilisation simple et intuitive pour la recherche en ligne sur la base (voir par exemple le modèle de http://www.lexique.org/). L'utilisateur pourra taper le mot qu'il souhaite rechercher et il obtiendra un tableau comportant toutes les informations détaillées dans le paragraphe précèdent.

Dans un premier temps il sera souhaitable d'accéder aux lemmes et aux informations linguistiques associées présentes dans la base en l'état actuel. Pour cette raison, nous proposons la création d'un tableau générale de présentation qui pourra être téléchargé en format .xls ou .csv.

En collaboration avec un étudiant stagiaire en Sciences du Langage, le stagiaire rédigera aussi un petit guide qui permettra d'expliciter les informations résultant de la recherche sur le modèle de http://lexique.org/_documentation/Manuel_Lexique.3.2.pdf

4 Références

[BFS12] Benzitoun C., Fort K. & Sagot B. (2012), TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe, in *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, volume 2: TALN, pages 99–112, Grenoble.

[LSC04] Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods*, 3*6 (1)*, 156-166.

[MCW00] MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

[NPFM01] New, B*.,* Pallier, C., Ferrand L. & Matos, R*. (*2001*). Une base de données lexicales* du français contemporain sur internet: LEXIQUE. *L'Année Psychologique*, 10, 447-462.

[Sch97] Schmid, H. (1997). New Methods in Language Processing, Studies in Computational Linguistics, édité par D. Jones et H. Somers, chapitre *Probabilistic part-of-speech tagging using decisiontrees*, pages 154–164. UCL Press, Londres.

5 Annexes

Olga, 2 ;03

48 *MOT: pourquoi tu veux pas de glace chérie?
49 *MOT: tu veux de la vache qui rit avec du [//] des petits gâteaux avec la
50 vache qui rit?
51 *MOT: +^ attends t'as +...
52 %act: MOT ajuste la serviette que CHI a autour du cou.
53 *CHI: t'es tout mouillée!
54 %pho: t tu muje
55 %act: CHI s'essuie la joue
56 %int: c'est/t'es
57 *MOT: beh oui excuse-moi.
58 *MOT: tu veux [>] <euh> [/] tu veux de la vache qui rit avec les petits
59 gâteaux là?
60 *CHI: [>] <non>
61 %pho: nɔ̃
62 *CHI: y.
63 %pho: ã
64 *MOT: oui?
65 *CHI: 0
66 %act: CHI hoche la tête.
67 %act: MOT se lève pour aller chercher la vache qui rit.
68 *CHI: xx tiens!
69 %pho: $$ tjɛ̃!
70 %act: CHI montre l'assiette sur la table.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<session xmlns="http://phon.ling.mun.ca/ns/phonbank" id="40000" corpus="GoadRose" version="PB1.2">
    <header>
        <date>1998-11-26</date>
        <language>fra</language>
        <media>40000.wav</media>
    </header>
    <participants>
        <participant id="THE">
            <role>Target Child</role>
            <name>Théo</name>
            <sex>male</sex>
            <age>P4Y0M0DT0H0M0S</age>
            <language>fra</language>
        </participant>
    </participants>
    <transcribers/>
    <userTiers>
        <userTier tierName="mod" grouped="false"/>
        <userTier tierName="pho" grouped="false"/>
        <userTier tierName="segtyp" grouped="false"/>
    </userTiers>
    <tierOrder>
        <tier tierName="Orthography" visible="true" locked="false" font="default"/>
        <tier tierName="IPA Target" visible="true" locked="false" font="default"/>
        <tier tierName="IPA Actual" visible="true" locked="false" font="default"/>
        <tier tierName="Notes" visible="true" locked="false" font="default"/>
        <tier tierName="Segment" visible="true" locked="false" font="default"/>
        <tier tierName="mod" visible="true" locked="false" font="default"/>
        <tier tierName="pho" visible="true" locked="false" font="default"/>
        <tier tierName="segtyp" visible="true" locked="false" font="default"/>
    </tierOrder>
    <transcript>
        <comment type="Code">pid 11312/c-00028254-1</comment>
```