D'un corpus à l'identification automatique d'erreurs d'apprenants

Marie-Paule Jacques
Lidilem, Université Grenoble-Alpes, Bâtiment Stendhal, 38058 Grenoble, France
marie-paule.jacques@univ-grenoble-alpes.fr

RÉSUMÉ

Nous présentons ici une étude préliminaire (work in progress) à l'élaboration d'un système dédié au repérage de zones potentielles d'erreurs dans des textes d'apprenants. Ce repérage permettra d'enrichir un corpus déjà constitué (que nous nommons Corpus de Littéracie Avancé) par un balisage des erreurs. Nous exposons ici la démarche adoptée pour mettre en place ce balisage : un appui sur certains textes du corpus qui sont commentés par les enseignants-correcteurs pour accéder directement aux passages problématiques, puis l'élaboration de requêtes formalisant les écarts à la norme repérés manuellement. Un exemple-jouet illustre la démarche.

ABSTRACT

From Learner corpus to Automatic Error Annotation
I present here a work in progress aiming at designing a tool to automatically retrieve errors in a learner corpus. The corpus is already constituted and available as "Corpus de Littéracie Avancée" in different formats, including an xml TEI-compliant one. My goal is to enrich the xml files with an error tagging. The paper illustrates the way I proceeded to carry out the tagging: I relied on annotations written in the original files by the readers of the texts (the teachers who ordered the work) to directly access to errors. From them, I built relevant queries to search the corpus for similar patterns. An example shows my approach.

MOTS-CLÉS : corpus de textes d'apprenants, repérage des erreurs

KEYWORDS: learner corpus, automatic error identification

1 Introduction

Le travail que nous présentons ici constitue une étude préliminaire à l'élaboration d'un système dédié au repérage automatique de zones potentielles d'erreurs dans les textes d'apprenants que nous avons rassemblés en corpus. Notre objectif à long terme est d'enrichir ce corpus, déjà disponible librement au format xml, d'un balisage des erreurs afin de permettre, d'une part, des recherches sur les particularités de ces écrits, d'autre part, l'élaboration de cours et d'exercices ciblés pour répondre aux besoins manifestés par ces erreurs.

La particularité de notre corpus et de notre étude réside dans le fait qu'il ne s'agit pas ici d'enseignement d'une langue étrangère : les apprenants sont des étudiants massivement français natifs, qui se destinent même pour certains aux métiers de l'enseignement. Au niveau universitaire,

ils ne reçoivent donc plus d'enseignements de langue similaires à ceux que suivent des apprenants de langue étrangère. Cependant, leurs productions manifestent de fréquents écarts à la norme, sur le plan orthographique en premier lieu, mais aussi sur le plan syntaxique et sur le plan textuel. Or, à notre connaissance, il existe peu d'études systématiques à ce palier de l'enseignement.

Le corpus que nous avons bâti se veut une ressource pour de telles études. Nous renvoyons à (Jacques et Rinck, à paraitre) pour sa description complète, nous en rappellerons ici les principaux objectifs pour nous focaliser surtout sur l'enrichissement que nous visons. Nous commençons par une brève revue de travaux antérieurs en lien avec notre étude avant d'expliciter notre démarche et d'indiquer quelques résultats.

2 Corpus d'apprenants pour l'enseignement

2.1 De langue étrangère

La constitution de corpus d'apprenants n'est pas une idée novatrice dans le domaine de l'enseignement des langues étrangères[1]. Ces corpus sont en effet une précieuse source d'informations sur les difficultés et écueils rencontrés lors de l'acquisition d'une langue étrangère : les erreurs qui s'y manifestent sont autant de pistes pour les remédiations ultérieures (Granger *et al.*, 2015). La production des apprenants témoigne de leur degré d'appropriation de la langue cible et de ses particularités linguistiques, aussi bien en terme de maitrise du lexique et de son usage, qu'en terme de syntaxe. La nature de certaines erreurs rend envisageable un repérage automatique : pour nombre d'apprenants d'une langue étrangère, ce sont le choix et l'usage du « bon » item lexical ainsi que l'association des verbes avec les préposition ou types d'arguments corrects qui cristallisent les erreurs de langue. Gaillat (2013) envisage par exemple d'utiliser une annotation en POS comme révélateur de mésusages des formes *this* et *that,* en procédant par comparaison avec un corpus arboré en anglais.

2.2 De langue maternelle

En ce qui concerne les locuteurs natifs et l'apprentissage de la langue maternelle, la situation est autre, du moins en France. La langue ne fait plus l'objet d'un enseignement formel et explicite au-delà du lycée, les quelques corpus d'apprenants de langue maternelle concernent donc surtout les élèves de primaire et de collège (Elalouf *et al.*, 2007). Toutefois, l'idée de s'appuyer sur les erreurs ou fragilités repérées dans ces corpus pour construire l'enseignement est aussi défendue (Cappeau & Roubaud, 2005).

La réalité des productions d'étudiants dans l'enseignement supérieur montre la nécessité de ne pas considérer que tout serait acquis à ce niveau d'études et qu'il ne serait pas nécessaire de continuer l'outillage des apprenants, alors même qu'ils sont confrontés à de nouvelles exigences et à des genres discursifs encore peu familiers, voire inconnus (on n'écrit pas de mémoire de recherche au

[1] Le site de l'Université Catholique de Louvain en recense un certain nombre : https://www.uclouvain.be/473700.html

collège ni même au lycée). C'est pourquoi nous avons depuis plusieurs années entrepris la constitution d'un corpus de textes produits par des étudiants en réponse à une « commande » universitaire, de la licence au master. Ce corpus de « Littéracie avancée », librement accessible, est disponible notamment au format xml proposé par la TEI. L'intérêt de ce format réside dans le fait qu'il comporte diverses balises utiles pour représenter les propriétés que nous jugeons pertinentes, telles que la structure textuelle (découpage en paragraphes, en sections...), et pour indiquer des métadonnées telles que la consigne d'écriture du texte ou l'année d'étude de l'étudiant (pour plus de détails, cf. Jacques & Rinck, à paraître). Le système de balises est en outre particulièrement approprié au repérage de zones remarquables, c'est-à-dire, pour notre propos, de zones dans lesquelles se donnent à voir les besoins de progression des étudiants. Nous voulons donc à terme insérer dans les textes des balises qui repèrent les erreurs, à peu près sur le modèle suivant, où la balise « lexError » signale une erreur de choix lexical (*vêtir* au lieu de *revêtir*) :

<sent> Le langage offre un panel d'inventions infinies, capable de <lexError>vêtir</lexError> différents visages en fonction de ces associations.</sent>

Ce corpus a déjà fait l'objet d'une exploitation pour la construction d'exercices ciblés (Jacques & Rinck, 2015), que nous décrivons succinctement afin de donner corps aux enrichissements visés.

2.3 Du corpus à l'enseignement

Nous avons utilisé une partie seulement du corpus (on comprendra plus loin pourquoi) pour construire des exercices destinés à aider nos étudiants (natifs, rappelons-le) sur deux plans : une amélioration de la maitrise des normes de l'écrit (tournures de phrases, enchainements discursifs, lexique…) et l'adoption d'une posture réflexive sur la langue et sur l'écriture (certains sont de futurs enseignants, ils se doivent d'entrer dans une réflexion métalinguistique). Divers exercices sont ainsi constitués d'extraits – fautifs – du corpus et de réécritures visant un rapprochement avec la norme, à partir desquels il est demandé de sélectionner soit les énoncés incorrects, soit les énoncés corrects. La Figure 1 donne un exemple d'exercice sur l'utilisation de « comment ».

La collecte des extraits susceptibles de servir de supports pour un tel entraînement a été effectuée totalement manuellement, ce qui est une tâche particulièrement gourmande en temps, c'est pourquoi elle n'a pas porté sur la totalité des 338 textes du corpus – dont certains sont des mémoires de 30 à 40 pages – mais sur un sous-ensemble d'un même genre de textes de 2 à 3 pages chacun.

Pour une réelle exploitation du corpus, l'idéal serait un inventaire exhaustif des erreurs, associé à un balisage. Pour cette tâche, le TAL est un moyen d'éviter la lecture extensive des 338 textes et 1,2 millions de mots du corpus : nous le pensons comme une assistance pour accéder aux passages des textes potentiellement erronés. Cet objectif pose de façon assez évidente la question d'une caractérisation à priori de ce qui, du point de vue d'un système de traitement automatique, peut alors « signaler » une erreur.

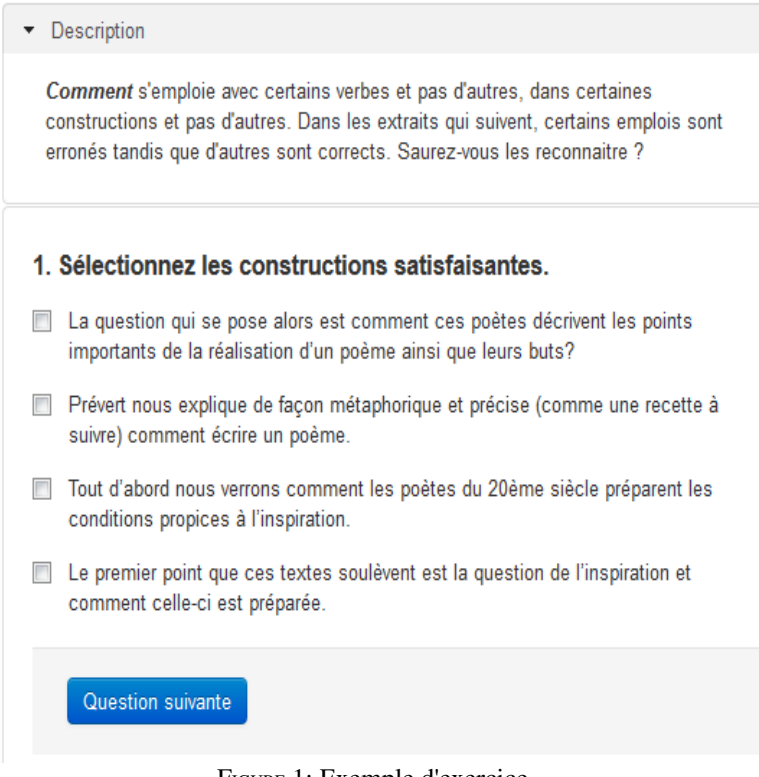

FIGURE 1: Exemple d'exercice

3 Repérer et baliser l'erreur

Reprécisons pour éviter toute ambiguïté que l'objectif ultime est d'insérer dans les textes au format xml des balises délimitant des zones d'erreur et non de proposer une correction automatique des textes. Et ajoutons que, dans la mesure où ce corpus est voué à être ensuite utilisé par des humains et non des machines, nous ne poserons pas d'exigence drastique en termes de délimitation de la zone à baliser : il suffit qu'un repère soit posé de façon à permettre un accès ciblé aux passages des textes potentiellement problématiques. Par exemple, dans l'extrait qui suit, toute la première partie de la phrase est correcte syntaxiquement, on pourrait considérer soit que l'erreur se cristallise sur le point d'interrogation qui n'a pas lieu d'être dans une interrogative indirecte, soit qu'elle tient à l'absence de ponctuation introduisant une interrogative directe, alternative qui est mise en évidence par les réécritures possibles :

Phrase incorrecte : *A partir de cette définition, se pose la question de savoir comment mettre en place cette acculturation à l'écrit en maternelle ?*

Réécriture 1 : *A partir de cette définition, se pose la question de savoir comment mettre en place cette acculturation à l'écrit en maternelle.*

Réécriture 2 : *A partir de cette définition, se pose la question : comment mettre en place cette acculturation à l'écrit en maternelle ?*

Contrairement à Lüdeling *et al* (2005), nous n'avons pas pour objectif de fournir les différentes hypothèses de réécriture – ce seront les recherches menées sur le corpus qui éventuellement inclineront à une hypothèse plutôt qu'à une autre – nous voulons juste signaler la phrase comme comportant un écart à la norme.

Pour construire notre système de traitement, nous procédons en trois temps :

- recueil manuel de passages d'erreurs ;
- classement en types d'erreurs ;
- modélisation en termes de « marqueurs » afin de construire le repérage automatique.

3.1 Des indices : les remarques des correcteurs

Dans le processus de découverte de ces marqueurs, les textes eux-mêmes recèlent des indices : 80 fichiers – fournis originellement dans un format de traitement de texte – comportent des annotations de correcteurs, insérées grâce à la fonction « commentaire » du traitement de texte. Ces annotations sont autant de pointeurs vers les passages problématiques. Elles nous facilitent donc une collecte de ces passages en vue de la typologie et de la modélisation ultérieure.

Une macro dans le traitement de textes nous permet d'extraire chaque passage avec le numéro de la page à laquelle il se trouve dans le document original et le commentaire du correcteur associé. Pour l'heure, nous avons traité 30 fichiers, ce qui nous fournit plus de 750 commentaires différents. Afin d'en extrapoler des données exploitables, il convient de les organiser, au moins grossièrement, en une typologie qui permettra ensuite d'atteindre les régularités.

3.2 Une ébauche de typologie

L'objectif ici est de ranger les passages extraits dans des classes pour spécifier des traitements automatiques différents selon les classes. Nous faisons l'hypothèse en effet que la nature des erreurs induit des procédures de repérage différentes : on ne mettra pas en œuvre la même stratégie pour identifier un problème d'orthographe et un problème de combinatoire verbe / préposition.

En l'état actuel, nos classes sont grossières mais répondent à ce besoin de sérier les extraits. Nous distinguons :

- l'orthographe : les écrits manifestent essentiellement des problèmes d'accords ;
- la formulation : sont ici concernés les passages dans lesquels le correcteur attire l'attention sur une défaillance de la formulation qui peut être liée à un choix de mot(s) inapproprié(s), une construction syntaxique contradictoire avec la combinatoire du lexique choisi, une construction syntaxique bancale qui n'assure pas la complétude de l'énoncé...
- l'agencement textuel (en lien avec la rhétorique du texte) : est ici en jeu le plan du texte et notamment la gestion des enchainements, la structuration, l'ordonnancement des idées...

Dans la mesure où le travail actuel est encore « en chantier », nous n'entrons pas davantage dans cette typologie qui doit encore s'affiner, sinon pour préciser que nous faisons le choix fort de ne pas intégrer cette typologie telle quelle dans le corpus : elle est à l'heure actuelle une heuristique pour les traitements, sa forme définitive sera étroitement liée à notre cadre théorique et aux applications envisagées, elle sera donc en ce sens éminemment contextualisée et probablement peu opératoire pour des chercheurs qui auraient d'autres visées. Elle est un temps de la démarche qui doit permettre de passer de cet ensemble d'occurrences d'erreurs déjà repérées à un système qui permette de délimiter des erreurs sur la majorité des textes du corpus qui n'est pas annotée par les enseignants-correcteurs.

Pour illustrer la démarche, nous allons montrer comment réfléchir le traitement automatique d'un problème du type « formulation », qui met plus particulièrement en jeu la gestion de contraintes syntaxiques au niveau de l'ordre des mots et de la ponctuation.

3.3 Un exemple jouet : le cas des interrogatives indirectes

3.3.1 Exposé du problème

La forme dite « interrogative indirecte » est volontiers présente dans les textes d'étudiants tels que dossiers, synthèses d'articles, mémoires, travaux d'étude et de recherche (TER), car elle y remplit la fonction souvent cruciale de l'expression de la problématique ou de la question traitée. Or elle se montre pour le scripteur d'un maniement peu aisée en raison de ses caractéristiques linguistiques : emploi d'un subordonnant exprimant l'interrogation SANS la forme syntaxique typique de l'interrogation qui elle-même repose sur l'inversion ou le redoublement du sujet et sur l'emploi d'un point d'interrogation. Une formulation normée est : « Nous verrons comment les poètes du 20ème siècle préparent les conditions propices à l'inspiration. » Mais la présence du subordonnant à valeur interrogative entraine régulièrement les étudiants dans une construction « entre-deux », c'est-à-dire qui mélange de façon erronée interrogations directe et indirecte : « On peut se demander comment les poètes de ces deux styles artistiques ont-ils défini leur travail et leur art ? ». On voit que coexistent ici des éléments linguistiques contradictoires : à la fois un subordonnant et une ponctuation interrogative. C'est cette coexistence qui « marque » potentiellement l'erreur de construction.

3.3.2 Traitement et premiers résultats

Notre objectif pour cet exemple jouet est d'identifier automatiquement les erreurs de formulation des interrogatives indirectes. Le corpus a été étiqueté morphosyntaxiquement avec Melt (Denis & Sagot, 2009), nous utilisons TXM (Heiden, 2010) pour élaborer et tester les requêtes qui doivent nous permettre d'atteindre les zones d'erreurs. Une requête représente la traduction en CQL (Corpus Query Language) des caractéristiques linguistiques identifiées manuellement, ici la discordance entre un verbe ou une tournure verbale dédiés à l'interrogation indirecte, par ex. « je me demande, on peut se demander, la question qui se pose... », et des éléments typiques de l'interrogation directe tels qu'une ponctuation finale sous forme d'un point d'interrogation ou une inversion du sujet. Voici un exemple de requête élaborée en CQL, qui correspond aux contraintes suivantes : rechercher un

pronom clitique, puis un mot dont le lemme commence par *demand* ou *question* ou *interr*, ou est *savoir*, puis une conjonction de subordination ou un mot interrogatif qui ne soit pas précédé d'un guillemet ou des deux-points, tout cela dans la limite de la phrase :

[lemma="cl."][]{0,10}[lemma="demand." | lemma="question.*" | lemma="savoir" | lemma="interr.*" | lemma="montr.*"][]{0,15}[lemma!=":" & lemma!="«"][pos="CS" | pos=".*WH"][]{0,15}[lemma="\?"] within sent*

Nous avons ainsi testé trois requêtes appuyées sur des propriétés linguistiques différentes pour le repérage des erreurs de formulation de l'interrogation indirecte. Elles donnent un score de 67 % de rappel et 40 % de précision pour une cinquantaine de contextes renvoyés. Dans la mesure où nous envisageons une validation manuelle des résultats des traitements automatiques, une précision de l'ordre de 40 % nous semble acceptable mais il sera nécessaire d'améliorer le rappel, en veillant toutefois à ne pas dégrader la précision.

4 Conclusion

Nous avons présenté la démarche que nous élaborons pour l'enrichissement d'un corpus existant, consistant en un balisage d'erreurs. Notre approche sera de fournir une ressource neutre quant à l'analyse précise des erreurs, qui est du ressort des études que ce corpus veut servir. L'étape actuelle consiste à bâtir un système à base de TAL qui repère les zones d'erreurs dans les textes et qui est vu comme un auxiliaire qui nous permet de réduire la tâche. Nous posons comme préalable une analyse des caractéristiques formelles de ces erreurs qui constitue alors le point de départ de requêtes en CQL. Même si les résultats ne montrent pas une précision élevée, notre approche permettra de cibler certains passages des textes et donc d'éviter une lecture extensive.

Remerciements

Nous remercions les relecteurs de l'atelier pour leurs remarques sur notre texte.

Références

Cappeau P., Roubaud M.-N. (2005). *Enseigner les outils de la langue avec les productions d'élèves*. Paris : Bordas.

Denis P., Sagot B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. Actes de *PACLIC 2009*.

Elalouf M.-L., Boré C. (2007). Construction et exploitation de corpus d'écrits scolaires. *Revue française de linguistique appliquée* 1/2007 (Vol. XII), 53-70.

Gaillat T. (2013). Annotation automatique d'un corpus d'apprenants d'anglais avec un jeu d'étiquettes modifié du Penn Treebank. Actes de *20e conférence sur le Traitement Automatique des Langues Naturelles*, 271-284.

GRANGER S., GILQUIN G., MEUNIER F. (2015). *The Cambridge Handbook of Learner Corpus Research.* Cambridge: Cambridge University Press.

HEIDEN S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. Actes de *24th Pacific Asia Conference on Language, Information and Computation*.

JACQUES M.-P., RINCK F. (2015). Une linguistique fondamentale et appliquée à base de corpus. *Colloque TRELA, Terrains de Recherche en Linguistique Appliquée.*

JACQUES M.-P., RINCK F. (à paraître). Un corpus de "littéracie avancée" : résultat et point de départ, *Corpus*, numéro spécial sur les corpus d'écrits scolaires.

LÜDELING A, WALTER M., KROYMANN E., ADOLPHS P. (2005). Multi-level error annotation in learner corpora. Actes de *Corpus Linguistics 2005, Birmingham*.