Description et annotation des erreurs : le cas des francophones s'exprimant en anglais

Camille Albert, Marie Garnier, Arnaud Rykner
Université de Toulouse Le Mirail France

Patrick Saint-Dizier
Institut de Recherche en Informatique de Toulouse – CNRS France

Résumé

Dans ce document, nous présentons quelques défis soulevés par l'analyse et la correction des erreurs commises par des rédacteurs francophones lorsqu'ils s'expriment en anglais. Nous présentons en particulier notre méthode de constitution d'un corpus d'erreurs et notre approche pour catégoriser ces erreurs. Un ensemble d'annotations est ensuite proposé pour marquer les erreurs, donner leurs caractéristiques et intégrer une ou plusieurs corrections.

1. Introduction

Les rédacteurs francophones qui doivent écrire en anglais sont souvent confrontés à des problèmes de forme (lexique et syntaxe) et de style, qui pénalisent parfois lourdement l'intelligibilité de leurs textes. Notre projet a pour objectif (1) de développer des procédures de correction sur des classes d'erreurs qui ne sont pas traitées à ce jour par les éditeurs de textes, même les plus avancés, (2) de modéliser et de réaliser un assistant intelligent qui accompagnerait des apprenants dans leur démarche de perfectionnement de leur compétence en anglais écrit, où interviennent de nombreuses considérations (style, niveau de langue, usages du domaine de spécialité, contexte de la phrase, public, etc.) (Han et al. 2005), (Lee et al. 2006).

Dans le but d'élaborer un tel assistant, nous analysons des productions spontanées, grand-public, peu contrôlées (courriels, forums) ainsi que des productions soignées, davantage professionnelles (publications, rapports). La première tâche est de repérer les erreurs puis de les catégoriser dans une perspective opérationnelle. Cette catégorisation se fait sur la base de traits communs à plusieurs erreurs. Nous considérons des erreurs d'origine grammaticale (par exemple, structure du GV, place de l'adverbe, etc.), lexicale, de style (par exemple, l'usage maladroit du passif) et de calque (à partir des structures françaises). Nous nous attachons ensuite à annoter les erreurs dans les textes en faisant apparaitre plusieurs caractéristiques, dont : la gravité de l'erreur,

l'intelligibilité du segment concerné, la ou les corrections (avec des segments de texte corrigés souvent différents selon la correction), la certitude de l'annotateur-correcteur.

L'un des buts majeurs du projet est d'explorer, puis de modéliser les stratégies cognitives utilisées en phase de détection et de correction d'erreurs habituellement déployées par des enseignants de langues ou des traducteurs et de les inclure dans l'assistant intelligent. En effet, la correction de nombre d'erreurs se base sur un processus d'analyse et de décision complexe que nous allons explorer. D'un point de vue opérationnel, nous proposons une analyse de la détection de l'erreur et de sa correction en faisant ressortir les arguments (et les objets abstraits qu'ils manipulent) pour et contre la correction et sa nature, ainsi que le modèle de décision associé. En complément de la correction, l'assistant devra s'adapter dynamiquement au niveau de l'apprenant et à ses besoins, en lui permettant, par exemple, de poser des questions sur la grammaire (règles, exceptions, structures courantes, etc.).

Ce travail comprend en particulier les composantes fondamentales et applicatives suivantes :

(1) Les aspects linguistiques : le fonctionnement de la grammaire en général, et en particulier en anglais, les liens entre le lexique et la grammaire, les différents paramètres du style selon le type de document (courriels, forums, rapports et publications), la prise en compte de l'utilisateur-apprenant,

(2) Les aspects cognitifs : les stratégies déployées par les experts pour identifier et analyser les erreurs et proposer une ou plusieurs corrections, les formes d'argumentation employées ainsi que les processus de décision menant à la correction effective,

(3) Les aspects didactiques : les types de dialogues entre l'assistant et l'utilisateur visant à faciliter l'acquisition de compétences en langue, dans le cadre d'une théorisation de l'explication,

(4) Les aspects de modélisation : le développement de modèles adaptés en particulier en traitement automatique des langues, ainsi qu'en argumentation et théorie de la décision (pour la correction) et en question-réponses (pour l'assistant intelligent et coopératif),

(5) La réalisation d'un prototype pouvant se greffer sur un éditeur de textes: ce prototype fera l'objet d'une évaluation détaillée par des étudiants et des enseignants en enseignement des langues.

Dans ce document, nous présentons tout d'abord la méthode de constitution du corpus, c'est-à-dire les différents paramètres étudiés et comment ceux-ci sont réalisés dans le corpus. Ensuite,

nous abordons, à travers l'analyse de travaux déjà réalisés ou en cours, une méthode de catégorisation des erreurs réalisée à la fois sur une base linguistique et opérationnelle. Enfin, nous proposons un ensemble d'annotations pour marquer ces erreurs et proposer des corrections. Ces annotations feront l'objet d'une analyse quant à leur intelligibilité, adéquation au problème à traiter et utilisation par des annotateurs. Ces annotations sont utilisées dans un cadre inductif pour induire des règles de correction (Albert et ali. 2009).

2. Constitution et paramètres du corpus

La constitution du corpus est une étape majeure pour l'analyse des erreurs. En effet, la qualité du corpus, au niveau des paramètres pris en compte, va déterminer en bonne partie la nature des erreurs relevées ainsi que la catégorisation qui sera effectuée. Une des difficultés principales est de déterminer les types de documents qui entrent dans la constitution du corpus : celui-ci doit en effet contenir les erreurs les plus fréquemment rencontrées dans les productions écrites par des francophones s'exprimant en anglais. La diversité des documents, des auteurs et des domaines, la taille des documents et leurs sources (professionnelles vs. privées) sont autant de critères qui doivent être pris en considération pour notre objectif.

2.1 Méthodologie générale de la constitution du corpus

Nous présentons ci-dessous les principaux paramètres qui nous sont apparus comme fondamentaux pour construire un corpus valide par rapport au problème que nous étudions. Ces paramètres ont été établis suite à une première analyse de quelques documents, dont le but était d'identifier les facteurs de variation des types d'erreurs en fonction des types de documents. Selon les documents et les auteurs, on observe une grande disparité dans les types de fautes.

Pour garantir une diversité et une bonne représentativité des erreurs, il convient de prendre en compte les paramètres suivants :

- les registres de langue,
- les domaines abordés (ceux du voyages et tourisme, de la médecine, de la technologie et des sciences, etc.),
- les sources des documents (professionnelles vs. privées, avec leurs sous-classes),
- les auteurs (79 auteurs au total) et leur niveau de performance linguistique,
- le lecteur ou le public visé,

- le niveau de contrôle présumé des documents.

2.2 Les paramètres de la constitution du corpus

Les différents paramètres pris en compte lors de la constitution de notre corpus sont présentés ci-dessous :

- **Diversité des documents et des niveaux de contrôle** : de façon à être représentatif des erreurs textuelles les plus courantes, notre corpus comporte environ 140 pages de textes, 90 pages pour les documents avec un niveau de contrôle peu élevé, et 50 pages pour les documents d'un niveau de contrôle élevé. Nous avons tenté de prendre en considération un large éventail de documents relativement faciles à obtenir: emails, blogs, forums, publications, rapports, etc. Les emails et les notes de forums, de blogs, et de pages web personnelles sont associés à un niveau de contrôle faible. Les publications, les rapports (domaines universitaires et professionnels) et les pages web professionnelles sont associés à un niveau de contrôle élevé. Pour toutes ces catégories, les situations sont très variables et nous observons l'existence d'un continuum, allant d'un niveau de contrôle très élevé à un niveau de contrôle très faible. Le registre de langue est associé au type de document.
- **Diversité des auteurs** : nous avons considéré environ 35 auteurs pour les documents d'un niveau de contrôle faible, et 25 auteurs pour les documents d'un niveau de contrôlé élevé, provenant de domaines divers et faisant preuve de différents niveaux de compétence en anglais.
- **Diversité des domaines** : La diversité des documents permet de toucher à différents domaines comme ceux du service public hospitalier, de l'entreprise (public et privé), du tourisme, de l'administration publique, de la gastronomie, etc. Cette diversité permet la prise en compte des savoir-faire, des pratiques linguistiques et des modes d'expressions propres à certaines communautés.

Les caractéristiques de notre corpus actuel, défini pour une étude de faisabilité, sont résumées dans les tableaux ci-dessous :

Le niveau de contrôle :

Type de documents/ Niveaux de contrôle		Très contrôlé	Contrôle suffisant	Peu contrôlé
Publications		×		
Sites web	Perso			×
	Office de tourisme		×	
	Gastronomiques		×	
	Universitaires	×		
	Hôteliers		×	
	Ministère des aff. Etrangères	×		
Emails	Universitaires		×	
	Informatique			×
	Aéronautique			×
	Médical			×

Diversité des auteurs :

Type de documents	Taille (nombre de mots)	Diversité des auteurs
Publications	33564	25
Sites web	7694	19 (estimation)
Emails	9331	35

Diversité des domaines :

Type de documents	Domaines	Taille (en nombre de mots)
Publication	Universitaire	33564
Site web	Personnel	457
	Offices de tourisme	2194
	Gastronomique	1836
	Universitaire	2380
	Hôtelier	601
	Ministère des aff. Etrangères	226
Email	Universitaire	2097
	Informatique	1753
	Aéronautique	3018
	Médical	2282
	familial	181

3. Une méthode de catégorisation des erreurs

A l'interface de la didactique et de la linguistique, l'apprentissage des langues étrangères est un domaine d'étude qui suscite de nombreuses recherches, notamment en ce qui concerne

l'apprentissage de l'anglais. L'analyse des erreurs produites par les apprenants dans la langue cible en est une composante importante. L'hypothèse selon laquelle les erreurs seraient la conséquence des interférences dues aux « habitudes » des apprenants dans leur langue maternelle (ou langue source) a donné naissance au domaine de l'analyse contrastive (*Contrastive Analysis*). Cette hypothèse a été remise en question dans les années 1970, avec le développement du domaine de l'analyse de l'erreur (*Error Analysis*), qui offrait une méthodologie pour l'exploration de la « langue » des apprenants (*learner language*). Cette méthodologie inclut les phases de constitution d'un corpus, de relevé des erreurs, de leur description et de leur explication. La dernière étape explore les raisons expliquant la production d'erreurs (surgénéralisation, mauvaise connaissance d'une règle de grammaire etc.), et relève notamment des domaines de la psycholinguistique et des sciences de l'éducation. La création de catégories d'erreurs est un élément important de la phase de description des erreurs relevées.

3.1 Explicitation de la méthode et des paramètres pris en considération

Précisons dès à présent qu'il s'agit bien de catégoriser les erreurs selon leur nature propre et non en fonction de la correction qui pourrait être apportée. Selon (Ellis, 1994, p. 54-56), les erreurs sont le plus souvent classées selon deux méthodes différentes. La première consiste à utiliser ce qu'il appelle des « catégories linguistiques », c'est-à-dire les parties du discours, les structures ou systèmes précis (système des auxiliaires, formation des propositions, etc.), ou encore des niveaux très généraux tels que la morphologie, la syntaxe et le lexique. (Izumi et ali., 2005, p.76) proposent par exemple une classification des erreurs d'apprenants Japonais en anglais sur la base des parties du discours (nom, verbe, modal, adjectif, adverbe, etc.). (Cain et ali., 1989, p. 25) fondent leur catégories d'erreurs d'apprenants francophones de l'anglais sur des systèmes tels que la détermination, la deixis, ou encore la transitivité. La seconde méthode consiste à observer les « stratégies de surface » (*surface strategies*) telles que l'omission, l'ajout ou le choix erroné d'un élément. (Cerda et ali. 1999, p.87) présentent un système de catégories qui utilise ces deux méthodes pour classifier les erreurs d'hispanophones s'exprimant en anglais : les catégories relèvent de niveaux linguistiques généraux mais des distinctions plus fines sont établies à l'intérieur de ces dernières pour refléter les stratégies de surface.

Dans certains cas, les erreurs sont regroupées dans quelques catégories ad hoc, c'est-à-dire valables uniquement pour la paire de langues à l'étude, et reflétant des points de friction précis

entre langue source et langue cible. Par exemple, (Chan, 2004, p. 56) choisit d'étudier et de classifier seulement cinq types d'erreurs représentatifs de la production en anglais de locuteurs du chinois de Hong-Kong.

La création de catégories ad hoc ne nous semblait pas convenir à une étude visant la description générale des erreurs produites par des francophones s'exprimant en anglais. La méthode de constitution des catégories basée sur l'observation des stratégies de surface ne nous semble pas permettre, quant à elle, de décrire finement les erreurs produites, et offre peu d'informations concernant leur source.

La catégorisation présentée dans cette étude repose donc sur des critères linguistiques. Nous avons choisi de regrouper les erreurs selon le groupe syntaxique (ou syntagme) dont elles relèvent le plus, ce qui garantit d'une part la reconnaissance de ces groupes par le plus grand nombre d'utilisateurs possible, et d'autre part permet de donner une visibilité aux erreurs de syntaxe à l'intérieur de ces groupes. De plus, les catégories ainsi formées sont de même niveau linguistique, ce qui confère une cohérence interne à ce système de classification. Ce choix de catégorisation est également intéressant car le système ainsi obtenu peut être transféré à d'autres paires de langues. La catégorisation présentée ici inclut aussi des catégories internes plus fines, et qui ont été formées suite à l'observation des types précis d'erreurs relevées.

3.2 Présentation de la catégorisation

Les catégories sont illustrées d'exemples, suivis de leur correction. Les lettres (P) et (I) font référence à la source des segments relevés (P pour « publications et rapports », I pour « emails, notes de forums, sites web (internet) »). Nous indiquons également la fréquence d'occurrence du type d'erreur à l'aide des lettres (TF) pour « très fréquent », (N) pour « normal », et (R) pour « rare », ces fréquences étant relatives à nos corpus.

GROUPE VERBAL

- Ordre des composants placés à la suite du verbe (arguments et ajouts):
  - Objet direct séparé du verbe (TF)
    - *Ontological domains include in our view objects, their properties and relations.* (P)
      → In our view, ontological domains include objects…
  - Placement de la particule adverbiale (*phrasal verbs*) (N)
    - *It does not take into account context.* (P)

→ It does not take context into account

- Réalisation dans le groupe verbal des contraintes lexicales liées au verbe
  - Choix de la préposition qui doit suivre le verbe (TF)
    - *These scores depend from the gold standard.* (P)
      → These scores depend on the gold standard
  - Transitivité (TF)
    - *I'm waiting your answer*. (I)
      → I'm waiting for your answer
- Adverbe:
  - Placement erroné après le verbe principal (TF)
    - *They exhibit nevertheless the dependency relationships observed in the source parse tree.* (P)
      → Nevertheless, they exhibit the dependency relationships…
  - Placement maladroit avant le verbe principal (R)
    - *The value of N is experimentally elaborated*. (P)
      → The value of N is elaborated experimentally.
  - Utilisation des adverbes de négation (N)
    - *They are not only constrained to the author's point of view any more.* (P)
      → They are not constrained only to the author's point of view any more.
- Choix de l'aspect (continu, perfectif) (TF)
  - *This summer, our association organizes a trip.* (I)
    → This summer, our association is organizing a trip.
- Choix des auxiliaires modaux (N)
  - *It appears that patients who suffer from FRS would be unable to correctly monitor their actions.* (P)
    → It appears that patients who suffer from FRS are unable to…
- Morphologie:
  - Formation des temps composés (*present perfect*) (N)
    - *You do not have takes action* (I)
      → You have not taken action

GROUPE NOMINAL

- Adjectif:
  - Placement de l'adjectif par rapport au nom (R)

- *The carrying of weapons is permitted in fifty states different.* (I)
  → The carrying of weapons is permitted in fifty different states

➢ Apposition abusive de l'adjectif (TF)

- *A subset of classes useful to describe our model* (P)
  → A subset of classes which is useful to describe our model

➢ Ordre des adjectifs dans une liste (N)

- *European academic and industrial partners* (P)
  → academic and industrial European partners

➢ Placement de l'adverbe modifiant l'adjectif (TF)

- *A quite detailed analysis* (P)
  → quite a detailed analysis

- Détermination:
  - ➢ Choix de l'article (TF)
    - *A Merovingian necropolis was built on exact site of the villa.* (I)
      → A Merovingian necropolis was built on the exact site of the villa
- Construction de la subordonnée complétive du nom (R)
  - […] *all the feelings that another is controlling the patient's thoughts* (P)
    → the feeling that another is controlling…
- Construction NØN:
  - ➢ Construction abusive pouvant générer des problèmes de compréhension (TF)
    - *Security object granularity* (P)
      → the granularity of security objects
  - ➢ Construction erronée, agrammaticale (TF)
    - *The objects properties* (P)
      → The properties of the objects, the objects' properties
- Morphologie:
  - ➢ Accord déterminant/nom (N)
    - *I didn't order this goods* (I)
      → I didn't order these goods
  - ➢ Accord agrammatical adjectif/nom (R)
    - *News clothes* (I)
      → New clothes

GROUPE PREPOSITIONNEL (y compris faisant partie d'un GN)

- Choix de la préposition à utiliser en fonction du cotexte (N)

- *They are exchanged and read <u>on</u> their electronic form* (P)

  → They are exchanged and read in their electronic form

PHRASE ET PROPOSITION

- Construction des propositions interrogatives directes (TF)
  - *<u>It is</u> possible to receive the parcel by the end of August?* (I)

    → Is it possible to receive the parcel by the end of August?
- Construction des subordonnées:
  - Interrogative indirecte (TF)
    - *It is necessary to know what <u>is their role</u> in the action expressed by the predicate*. (P)

      → It is necessary to know what their role is in the action…
  - Subordonnée à verbe non-fini (N)
    - *They read annotations <u>for evaluating</u> them.* (P)

      → They read annotations to evaluate them
- Placement des ajouts (TF)
  - *Goals and subgoals are most of the time realized by means of titles*. (P)

    → Most of the time, goals and subgoals are realized…
- Construction du comparatif (N)
  - […] *<u>as many as possible of incorrect analyses</u>* (P)

    → a many incorrect analyses as possible
- Gestion de l'information/micro-planification (N)
  - *August 15 in France, <u>it is a holiday</u>.* (I)

    → In France, August 15 is a holiday.
- Morphologie
  - Accord sujet/verbe (N)
    - *The first written mention of Issigeac <u>date</u> from 1008*. (I)

      → The first written mention of Issigeac dates from 1008.

LEXIQUE ET DIVERS

- Connaissance des caractéristiques lexicales d'un terme:
  - Nom vs. Verbe (N)
    - *You will give my <u>apologize</u> to her.* (I)

      → You will give my apologies to her
  - Nom dénombrable vs. Nom indénombrable (TF)

- *A valuable information* (P)
  → valuable information

➢ Adverbe vs. Conjonction (N)

- *Although, such an excess of mental effort should be reduced at all costs* (P)
  → However, such an excess…

➢ Signification précise (TF)

- […] *to remind a document*
  → to remember a document

- Utilisation de collocations et d'expressions idiomatiques (TF)
  - *Well cordially* (I)
    → Yours sincerely
- Orthographe (N)
  - *A shedulle* (I)
    → a schedule
- Ponctuation (TF)
  - *The purchase price will be validated by you and me, for the year*. (I)
    → The purchase price will be validated by you and me for the year.

3.3 Difficultés liées à la catégorisation des erreurs

La complexité des erreurs peut rendre la tâche de catégorisation relativement difficile. Pour commencer, une même erreur peut parfois être classée dans deux catégories. Par exemple, dans le segment **Our system is able to derive automatically information*, le placement erroné d'un adverbe après un verbe admettant un complément d'objet direct provoque également la séparation du verbe et de son objet : le segment erroné peut donc trouver sa place dans ces deux catégories.

De plus, certains segments erronés présentent plusieurs erreurs juxtaposées ou hiérarchisées, comme dans l'exemple suivant : **Nobody have answer me*. Ici, le segment présente deux erreurs de morphosyntaxe, l'une concernant l'accord du sujet et du verbe (Phrase et proposition – Morphologie), et l'autre la formation d'un temps composé (Groupe Verbal – Morphologie). Le choix de l'utilisation de l'aspect perfectif est également une erreur (Groupe Verbal – Choix de l'aspect). Ce segment cumule donc trois erreurs relevant de trois catégories différentes.

Enfin, il est parfois malaisé d'évaluer la nature de l'erreur, comme dans le segment suivant : **It is essential to be honest on both sides so that functions*. Ici, la double étiquette de *that* (pronom démonstratif / conjonction de subordination) rend cet erreur ambiguë : il peut s'agir d'une omission du sujet, ou bien de l'omission de la seconde partie de la conjonction *so that*, *that* étant ici traité comme le sujet (ceci peut être le résultat d'un calque du français « pour que cela fonctionne »). Cette ambiguïté n'empêche pas la correction mais peut mener à des difficultés de catégorisation.

Si les erreurs observées peuvent présenter un obstacle à la catégorisation, la nature même de cette tâche se révèle souvent problématique. Pour commencer, tout système de catégorisation est dans une certaine mesure ad hoc, puisqu'il repose sur une paire de langues précise. Par exemple, on ne peut douter que les locuteurs natifs du thaï (langue sans morphologie et sans ordre imposé des composants dans la phrase) s'exprimant en anglais produisent des erreurs de nature différente de celles retrouvées dans l'expression des francophones.

De plus, il est difficile de proposer un système de catégories pertinent pour tous les types de corpus étudiés ici, car on y trouve des erreurs de natures différentes. Les productions très contrôlées telles les publications présentent des erreurs de syntaxe complexe (subordination, construction NØN etc.), mais peu d'erreurs concernant la morphologie ou le lexique, notamment car les auteurs peuvent avoir recours à des correcteurs informatiques. Les productions d'un niveau de contrôle faible incluent en revanche de nombreuses erreurs morphologiques et lexicales, tandis que les erreurs syntaxiques touchent plutôt la syntaxe simple, car les auteurs évitent de produire des énoncés très complexes et de multiplier le risque d'erreur. Le type de classification que nous avons choisi permet néanmoins de rassembler ces erreurs dans les mêmes catégories principales, puisqu'elles peuvent convenir à tout type de texte. Les distinctions se retrouvent alors au niveau des catégories internes.

Toutefois, il faut garder à l'esprit que le choix de catégories est toujours arbitraire dans une certaine mesure. En effet, notre système n'offre pas la possibilité de classer les erreurs selon qu'elles relèvent de la morphologie ou de la syntaxe, et les erreurs liées au lexique ne peuvent pas toutes être regroupées dans la dernière catégorie (i.e. Lexique et autres). Cependant, choisir ce type de classification se révèlerait également insatisfaisant, d'une part parce qu'il ne permettrait pas l'identification des erreurs selon leur groupe syntaxique, et d'autre part parce qu'il générerait tout autant de flottements (lexique et grammaire sont parfois difficiles à démêler).

Enfin, la dernière difficulté identifiée concerne la granularité, ou le degré de précision à donner au système de catégories. Si des catégories très précises permettent de décrire finement les erreurs rencontrées, elles sont un frein à l'implémentation des résultats dans un système informatique. Il faudrait donc pouvoir adapter la catégorisation aux objectifs de la recherche, tout en leur conservant leur pertinence en tant que système indépendant.

3.4 La catégorisation idéale

La solution à certaines des difficultés présentées précédemment pourrait être de créer un système de catégorisation multidimensionnel, capable de croiser des informations linguistiques générales (morphologie, syntaxe, lexique, etc.), avec les groupes linguistiques dont semblent relever les erreurs. Ce système pourrait également inclure la mention des stratégies de surface (omission, ajout, emploi erroné, etc.). De plus, il serait nécessaire d'évaluer avec le plus de justesse possible les sources de l'erreur (surgénéralisation, transfert syntaxique, ignorance d'une règle), qui pourraient également être une dimension du système, dans le but de faciliter la proposition de corrections pertinentes (cf. Suri et McCoy, 1993).

4. L'annotation des erreurs

Introduisons à présent le schéma d'annotation que nous avons développé. Celui-ci est encore au stade de validation et peut faire l'objet de précisions et de modifications. En effet, stabiliser un tel schéma complexe demande de nombreuses expérimentations en boucle. Ce schéma est une tentative pour refléter, de façon factuelle et déclarative, les différents paramètres qui ont émergé lors de l'analyse des comportements cognitifs déployés par les didacticiens et traducteurs humains lorsqu'ils détectent et corrigent une erreur.

Nous employons un formalisme très standard, XML, enrichi d'attributs dont les valeurs associées ont une granularité élaborée par des didacticiens de notre groupe. La structure des attributs a été conçue de façon à ce qu'ils puissent être utilisés dans un cadre d'argumentation (Albert et ali. 2009).

Nous présentons ci-dessous les différentes étapes de l'annotation, en partant de la détection de l'erreur pour aller jusqu'à sa correction. Les étiquettes sont en anglais, de façon à pouvoir par la suite avoir un métalangage utilisable dans de nombreux contextes.

4.1 Délimitation et caractérisation de l'erreur

<error-zone> étiquette le groupe de mots impliqués dans l'erreur. Cette zone doit être aussi réduite que possible. Cette étiquette a plusieurs attributs :

Comprehension : avec des valeurs de 0 à 4 (0 étant le plus mauvais), indique si le segment erroné est malgré tout compréhensible ou non (0 : incompréhensible, 1 : compréhensible avec beaucoup d'efforts, 2 : compréhensible avec un peu d'effort, 3 : compréhensible mais peu naturel, 4 : bon).

Agrammaticatlity : avec des valeurs de 0 à 2 indique le niveau de gravité de l'erreur (0 : agrammatical, 1 : doute quant à la grammaticalité du segment, 2 : grammatical).

Categ : indique la catégorie principale de l'erreur : lexicale, syntaxique, stylistique, sémantique, textuelle. Si une erreur couvre plusieurs champs, ce qui est assez fréquent, il est possible de spécifier plusieurs valeurs. Le choix d'une catégorie peut être aussi parfois quelque peu ambigu.

Source : indique, globalement, l'origine de l'erreur, par exemple : calque (lexical ou grammatical), surgénération, manque d'effort, etc.

4.2 Délimitation et caractérisation de la correction

<correction-zone> identifie le fragment de texte sur lequel se fera la correction. Il est égal ou plus grand que <error-zone>. Nous abordons ici le point central. Chaque correction possible de l'erreur est introduite par le tag <correction> avec ses attributs associés. Nous donnons ici chaque attribut avec ses valeurs. En général, un système simple à 3 valeurs a été adopté. Celui-ci peut bien entendu évoluer si la faisabilité est possible. La ou les valeurs positives sont soulignées :

Surface : caractérise la taille du texte affectée par la correction : <u>minimal</u>, average, maximal,

Grammar : indique, lorsque c'est approprié, le statut de la correction : <u>by default</u>, alternative, improbable,

Meaning : indique si, par la correction proposée, le sens a été altéré ou non : yes, <u>no</u>, somewhat.

Change : indique la nature de la correction proposée : syntactic, lexical, stylistic, sémantic, textual,

Comp : indique si la correction proposée est un fragment de texte facile à comprendre : <u>yes</u>, average, no.

Fix : indique si la correction correspond à une situation locale, idiosyncratique, qui ne peut être étendue à aucune autre situation similaire : yes, no.

Qualif : indique le niveau de compétence et de certitude que l'annotateur a de sa correction : low, average, high.
Correct : indique le segment de texte corrigé.

A titre d'illustration, voici le cas de la construction NØN incorrecte, comme dans *the meaning utterance* (ou *the goal failure)*, où deux corrections sont possibles :

<error-zone comprehension="2" agrammaticality="1" categ="syntax" source="calque">

the meaning utterance

<correction qualif="high" grammar="by-default"
surface= "minimal" meaning= "not altered" Var-size="+2"
change="synt" comp="yes"
correct= "**the meaning of the utterance**">
</correction>

<correction qualif="high" grammar="unlikely"
surface= "minimal" meaning= "somewhat" Var-size="0"
change="lexical+synt" comp="average"
correct= "**the meaningful utterance**">
</correction>
</error-zone> </correction-zone> without a context.

5. Conclusion

Dans cet article, nous avons présenté certaines des étapes préliminaires à la réalisation d'un assistant informatique permettant d'améliorer les compétences et les performances de francophones rédigeant des textes en anglais. Ce projet implique d'établir des procédures et d'étudier les stratégies de détection et de correction des erreurs. Notre corpus exploratoire a été constitué en prenant en considération plusieurs paramètres, dont les différents types des documents et de niveaux de contrôle, ainsi que la diversité des auteurs et des domaines. L'objectif de la prise en compte de ces paramètres est de rassembler et de décrire des erreurs représentatives des productions de francophones dans différents contextes. Les documents qui constituent notre corpus exploratoire proviennent donc de sources très diverses (université, commerce, forums de discussion) et représentent des niveaux de contrôle allant de très fort (ex. publications scientifiques) à très faible (ex. emails personnels).

L'étape suivante que nous avons abordée est la méthode de catégorisation des erreurs. Nous avons choisi de classer les erreurs principalement selon leur groupe syntaxique le plus

probable, en créant également des catégories plus fines à l'intérieur de ces groupes généraux. Ce choix permet à notre système de conserver une cohérence interne et d'être transférable à d'autres langues de structure relativement similaire. De plus, le degré de granularité choisi semble être en accord avec notre objectif de création d'un système informatisé. La catégorisation pourrait être améliorée par la mise en place d'un système multidimensionnel permettant d'inclure et de croiser d'autres types d'informations (morphologie, syntaxe, lexique, etc.). Il s'avèrera nécessaire d'analyser les sources psycholinguistiques et cognitives des erreurs afin d'entrer dans la phase d'explication des erreurs et ainsi à terme de proposer à l'utilisateur des corrections aussi pertinentes et utiles que possible.

La proposition d'un schéma d'annotation des erreurs relevées permet de refléter les paramètres qui émergent lors de la phase de détection et de correction des erreurs par des correcteurs humains. Le schéma proposé utilise un formalisme XML complété par des balises spécifiques, et inclut les étapes de délimitation et de caractérisation de l'erreur, et de délimitation et caractérisation de la correction. Nous prévoyons d'utiliser ce schéma afin de produire des règles de correction. Les règles de correction seront produites par induction à partir des exemples annotés. Il s'agit d'une forme d'apprentissage rudimentaire, mais bien adaptée à notre problématique. Un des difficultés sera de dégager les bons niveaux d'abstraction tout en gérant de façon fiable des ensembles d'exceptions ou particularismes.

Bibliographie

Albert, C., L. Buscail, M. Garnier, A. Rykner, et P. Saint-Dizier, (2009). « Annotating language errors in texts : investigating argumentation and decision schemas » ACL-LAWIII, Singapour.

Cain, A. (dir.) (1989). L'analyse d'erreurs, accès aux stratégies d'apprentissage : une étude inter-langues, INRP.

Cerda, E., C. Valero, C. Flys et G. Mancho (1999). « Error analysis and the teaching of English in tourisme studies », III Congres Internacional sobre Llengues per a Finalitats Especifiques, Universitat de Barcelona, p. 86-89.

Chan, A. Y. W. (2004). « Syntactic transfer: evidence from the interlanguage of Hong Kong Chinese ESL learners », The Modern Language Journal, 88, I, p. 56-71.

Ellis, R. (1994). The study of second language acquisition, Oxford University Press.

Han et al. (2005). Detecting Errors in English Article Usage by Non-native Speakers, NLE.

Izumi, E., K. Uchimoto et H. Isahara (2005). « Error annotation for corpus of Japanese learner English », Proceedings of the 6th International Workshop on Linguistically Annotated Corpora, p. 71-80.

Lee, H., et A. Seneff (2006). Automatic Grammar Correction for Second-language Learners, Proc of InterSpeech.

Camille Albert est titulaire d'un Master 2 Recherche en Etudes Anglophones (spécialité : linguistique), délivré par l'Université de Toulouse le Mirail (UTM). Ses recherches concernent l'apprentissage de l'anglais au collège en France. Elle travaille sur le traitement de l'erreur et la correction automatique au sein de l'équipe ILPL (Informatique Linguistique et Programmation Logique) de l'IRIT depuis octobre 2008.
calbert433@aol.com, 05 61 55 62 44

Marie Garnier est agrégée d'anglais et titulaire d'un Master 2 Recherche en Etudes Anglophones (spécialité : linguistique), délivré par l'UTM. Ses recherches concernent notamment l'importance cognitive du corps dans la langue. Elle travaille sur le traitement de l'erreur et la correction automatique au sein de l'équipe ILPL depuis octobre 2008, et est actuellement en préparation de thèse en collaboration avec l'équipe de P. Saint-Dizier et l'UTM (projet CAPRA).
mhl.garnier@gmail.com, 05 61 55 62 44

Arnaud Rykner est Professeur à l'UTM et membre de l'Institut Universitaire de France. Il est également directeur du laboratoire Lettres, Langages et Arts, et est spécialisé dans les études théâtrales et l'esthétique de la représentation. Il est une des personnes à l'origine du projet d'amélioration des corrections/traductions automatiques de texte (projet Transtyler), et travaille en collaboration avec l'équipe ILPL depuis 2007.
arnaud.rykner1@neuf.fr, 05 61 50 35 48

Patrick Saint-Dizier est Directeur de recherches au CNRS en poste à l'IRIT. Il dirige l'équipe ILPL depuis plus de vingt ans. Ses principaux axes de recherche sont le traitement automatique

des langues naturelles, la sémantique textuelle et les systèmes de question-réponse. Il contribue au projet actuel de création d'un système informatique d'amélioration des productions des francophones en anglais.

stdizier@irit.fr, 05 61 55 62 44