Informatisation d'un corpus de dictées :

40 années de pratique orthographique (1967-2008)

Cédrick Fairon* (Université catholique de Louvain, CENTAL)
Anne Catherine Simon* (Université catholique de Louvain, VALIBEL)

Introduction

La question de l'orthographe reste au centre des débats. La maitrise de la norme du français écrit n'est pas que l'affaire des linguistes : elle est un marqueur social et un lieu d'intervention des politiques de la langue. Cet article apporte une contribution aux études sur la maitrise de l'orthographe chez les jeunes. Nous y présentons la méthode selon laquelle un vaste corpus, représentant 40 années de dictées passées par des étudiants de première et deuxième années à l'université, a été informatisé. La transformation de ces archives en un corpus informatisé permet d'envisager des analyses quantitatives aussi bien que qualitatives. Dans la dernière partie de l'article, nous présentons les premiers résultats d'une analyse d'une partie du corpus.

I. Les dictées et la passion de l'orthographe

A. *« La langue française en crise… »*

« Le français m'a tuer »[1]
« Enseignement de la langue. Crise, tension ? »[2]
« L'écriture électronique, une menace pour la maitrise de la langue ? »[3]
« MoliR, rev1 vit… il son 2vnu foo ! »[4]

Ces quelques titres d'ouvrages et d'articles récents donnent le ton : la question de l'orthographe et de la maitrise de la langue reste un sujet d'actualité, souvent traité de manière passionnelle (voir Paveau et Rosier 2008 : 241 *sq.*). L'avènement des technologies de l'information et de la communication (SMS, chat, forum, blog, etc.) a promu d'une manière accrue la communication écrite (« on n'a jamais tant écrit », disent les enthousiastes) tout en suscitant l'apparition de nouvelles pratiques qui donnent des sueurs froides aux enseignants confrontés à des jeunes « qui ne savent plus écrire », un poncif remis au gout du jour à propos du « langage SMS » et de l'internet.

La nouvelle orthographe, théoriquement plus simple, devrait faciliter l'apprentissage du système orthographique par les jeunes. Mais force est de constater que, de 1990 à nos jours, sa pénétration dans la société a été limitée. L'apparition de cette orthographe réformée dans les logiciels de correction orthographique, la mention officielle de la nouvelle orthographe dans les programmes d'enseignement en France (depuis 2008) et, enfin, les décrets signés en octobre 2008 par les différents ministres belges en charge de l'enseignement pour recommander l'usage prioritaire de la nouvelle orthographe sont des signes tangibles qu'un changement est à l'œuvre.

Bien entendu, les inquiétudes au sujet de la maitrise de la langue par les jeunes ne sont pas nouvelles : Hopper (1975, cité par Gagné 1979) cherchant des traces de ce débat a retrouvé des textes allant de 1689 à 1933. En 1730, Pierre Restaut écrivait que « les jeunes gens sortent des collèges aussi ignorans [de leur langue maternelle] que s'ils avoient esté élevez chez des étrangers ». L'évolution et les débats autour de l'orthographe montrent que cette question n'est pas seulement linguistique ; elle est aussi (et surtout) une question politique et sociale[5].

Du point de vue du linguiste, deux approches sont envisageables. D'une part, on peut étudier l'évolution des relations entre langue orale (code phonique) et écrite (code graphique), qui fondent les conventions

---

* Les auteurs de cet article l'ont rédigé sur la base d'une expérience menée avec les étudiants du cours CLIG2250 *Méthodologie de l'analyse de corpus en linguistique* durant l'année 2007-2008 à l'UCL.

1 Didier et al., 2006.

2 *Le Français aujourd'hui* nº 156 (1/2007).

3 David et Goncalves (2007).

4 Jalabert (2006).

5 Marie-Anne Paveau et Laurence Rosier (2008) analysent les relations complexes qui se tissent, sur la question de l'orthographe, autour de ces trois niveaux.

orthographiques. D'autre part, on peut chercher à mesurer et à comprendre les relations des scripteurs à l'orthographe, par des études sur des corpus de productions écrites.

B. *Études sur la maitrise de l'orthographe*

Est-il possible d'objectiver un phénomène tel que l'évolution de la maitrise de norme de l'écrit ? Plusieurs auteurs s'y sont essayés. Dans la Communauté Wallonie-Bruxelles, on peut relever l'étude de Théo HACHEZ et Bernadette WYNANTS (1991). Dans le cadre d'une enquête subventionnée par le Service de la Langue française, les auteurs ont analysé un corpus de 1200 textes produits par des élèves du secondaire et ont analysé les « écarts » concernant deux grands types de comportement : la formulation (il s'agit de textes produits par les élèves, et non pas de dictées) et la transcription (c'est-à-dire la conformité orthographique à la norme du français). Cette étude constitue donc une première référence objective[6] pour des textes produits par des élèves.

En France, André CHERVEL (1989) a lui aussi publié une enquête visant à comparer les performances orthographiques d'élèves français, à l'échelle d'un siècle. A notre connaissance, aucun corpus de dictées n'est accessible en Communauté française de Belgique. Le « Corpus Lenoble-Pinson » dont il est question dans cet article constitue une ressource exceptionnelle permettant de suivre sur 40 années l'évolution de la maitrise de l'orthographe au sein d'une population particulière.

Durant toute sa carrière, Michèle Lenole-Pinson a enseigné la grammaire aux étudiants des Facultés universitaires Saint-Louis et évalué leur maitrise de l'orthographe par des dictées. Soucieuse de pouvoir un jour exploiter ce qui allait un jour devenir une mine de données, elle a conservé méticuleusement l'ensemble des dictées d'examen rédigées par les étudiants entre 1968 et 2008. La méthodologie adoptée pour la correction n'a presque pas changé au cours du temps (des critères supplémentaires ont été introduits après l'entrée en vigueur des rectifications orthographiques), ce qui donne un ensemble de dictées corrigées et annotées de manière cohérente. Les amis et collègues qui connaissent bien Michèle Lenoble-Pinson savent qu'elle a en la matière la minutie et l'application de l'orfèvre.

Ce corpus a déjà été étudié de manière manuelle pour les participes passés[7]. Pour susciter de nouvelles exploitations et études, le corpus a été donné au Centre de traitement automatique du langage de l'UCL (CENTAL). Bien entendu, l'exploitation manuelle d'un tel corpus est extrêmement difficile : il faut manipuler les copies, lire (déchiffrer) les textes, relever les phénomènes pertinents, etc. La question de son informatisation s'est donc naturellement posée et nous rendons compte ci-après d'une première expérience réalisée à l'UCL dans le cadre d'un cours consacré à la linguistique de corpus.

II. Transformer des archives en un corpus exploitable

Des textes, aussi régulière et rigoureuse qu'ait pu être leur collecte, ne peuvent faire l'objet d'une analyse linguistique sans traitement préalable. Le premier critère qui distingue une archive (« collection aléatoire de textes ») d'un corpus est que ce dernier « is designed to represent a particular language or language variety » (Mc Enery, Xiao et Tono 2006: 13). Plus précisément, un corpus spécialisé est représentatif d'un genre particulier de textes, ou d'un domaine spécifique.

Le deuxième critère qui distingue une archive d'un corpus est que les textes d'un corpus sont documentés. Les métadonnées sont des informations sur la source du texte (qui en est l'auteur?), la date, et la situation dans laquelle il a été produit. Ces informations permettent d'interpréter, en relation à des facteurs extralinguistiques, les particularités langagières observées dans les textes du corpus.

Troisièmement, les textes d'un corpus doivent être exploitables informatiquement. Les premiers linguistes à travailler sur des données attestées, recueillies sur le terrain, se sont contentés de fiches rassemblant leurs observations. L'analyse de grandes masses de données n'autorise plus un traitement manuel et passe par

6 Les copies ont été produites par environ 600 élèves de cinquième et sixième années du secondaire, issus de dix-sept institutions scolaires. Par copie (2000 signes environ), on dénombre une moyenne de 22 écarts. La répartition des écarts est très inégale et, toujours selon les auteurs, « quatre cinquièmes des écarts enregistrés sont des écarts de transcription. Mieux encore : près de la moitié des écarts observés sont explicables par homophonie. Ce qui frappe également, c'est que l'essentiel des différences de score entre les élèves est due à la transcription. En d'autres termes, la conformité de la formulation semble plus stable […]. » (Hachez et Wynants 1991)

7 À la demande de la Commission de l'enseignement du Conseil supérieur de la langue française de la Communauté française de Belgique, et en vue de la participation de ladite Commission à l'Observatoire francophone du français contemporain, Jonas BENA (2004) a réalisé une enquête sur l'accord du participe passé dans 500 dictées sélectionnées entre 1995 et 2003.

l'acquisition numérique des données qui ne sont pas sur ce type de support. C'est *a foritori* le cas pour des dictées faites par des étudiants à l'occasion de sessions d'examen.

Un corpus peut ne contenir que des textes (pour autant qu'ils soient sur support informatique). Le plus souvent cependant, on enrichit les textes par des annotations. Les annotations sont le résultat d'une analyse et reposent sur une théorie linguistique. Par exemple, on peut annoter les « parties du discours » (*part of speech*) en indiquant pour chaque occurrence du texte sa catégorie grammaticale (nom, verbe, pronom, etc.). L'annotation est généralement automatique, avec une vérification manuelle. Elle présuppose une liste des catégories à identifier. Un corpus enrichi par des annotations peut faire l'objet d'analyses plus poussées et plus fines.

Avec les étudiants de première année de master[8], nous avons formé le projet de transformer les archives de dictées léguées au Cental par Michèle Lenoble-Pinson en un corpus en bonne et due forme, permettant de conduire des analyses sur la maitrise de l'orthographe par les étudiants de première année à l'université.

A. *Encodage des métadonnées et identification des textes*

Chaque dictée comporte des informations sur son auteur, en l'occurrence l'étudiant qui l'a écrite. Outre le nom et le prénom, on connait la date de la dictée, la faculté à laquelle appartient son auteur et son niveau d'étude (première ou deuxième année à l'université). Le prénom permet généralement, mais pas toujours, de déduire le genre (masculin ou féminin) de l'auteur.

Ces informations sont importantes si l'on souhaite, par exemple, étudier l'évolution dans le temps des résultats obtenus par les étudiants, ou si l'on cherche à savoir si, au cours d'une même année académique, les étudiants obtiennent de meilleurs résultats à la session de juin qu'à celle de janvier. Afin d'être en mesure d'interroger les métadonnées, nous avons créé une base de données utilisant le logiciel Microsoft Access pour encoder ces informations[9]. Pour chaque texte, on a ainsi encodé :

- **le type de texte** : outre des dictées, les archives de Michèle Lenoble-Pinson contiennent des rédactions, que nous n'avons pas encore traitées ;
- **l'année** : d'après la date à laquelle la dictée a été rédigée (pour regrouper les dictées d'une même année académique, il faut tenir compte du fait que les sessions peuvent tomber sur deux années civiles);
- **la session d'examen** : la dictée a pu être passée à la rentrée académique, en janvier, en juin ou en septembre ;
- **un numéro unique** : ce numéro identifie chaque étudiant d'une pile de dictées et est attribué selon l'ordre alphabétique ;
- **le sexe** (masculin, féminin ou inconnu, dans les cas où le prénom est unisexe) ;
- **le type de diplôme** auquel l'étudiant est inscrit (par exemple la Philologie romane) ;
- **l'année dans le programme** : les dictées ont été faites par des étudiants de première ou de seconde candidature.

D'un point de vue sociolinguistique, les informations sur l'auteur d'une dictée se limitent au sexe et au type de formation suivie. Aucune information n'est fournie sur la langue maternelle de l'étudiant (dans le cas où ce n'est pas le français), ni sur son parcours scolaire antérieur. Un chercheur qui voudrait prendre ces informations en compte pour son analyse devrait les obtenir auprès de l'université elle-même.

L'encodage des métadonnées dans une fiche du logiciel Access génère un code d'identification unique (par exemple, D-1993-3-21-1-ROM-1). Cet identifiant est manuellement reporté sur l'original de la dictée. Cette technique présente deux avantages. Elle garantit une forme d'anonymisation dans la dénomination des textes lors de leur informatisation[10] et elle permet d'accéder à des informations minimales à partir du nom du fichier en vue de faciliter l'archivage raisonné des textes.

Dans la base de données Access, on encode également le nombre total de fautes faites dans chaque dictée (sur la base des commentaires et du décompte faits par l'enseignante et mentionnés manuscritement sur la

8 Les étudiants du cours de *Méthodologie de l'analyse de corpus en linguistique* en 2007-2008 provenaient principalement du Master en Langues et littératures françaises et romanes et du Master en linguistique de l'Université catholique de Louvain.

9 La base de données a été programmée par un étudiant.

10 Les étudiants auteurs des dictées n'ont pas été explicitement informés que leurs productions pourraient être rassemblées dans un corpus. Le caractère anonyme des données est donc capital et doit être préservé (pour une réflexion sur cet aspect éthique de la consitutition de corpus, voir Baude (2006 : 119 et suivantes).

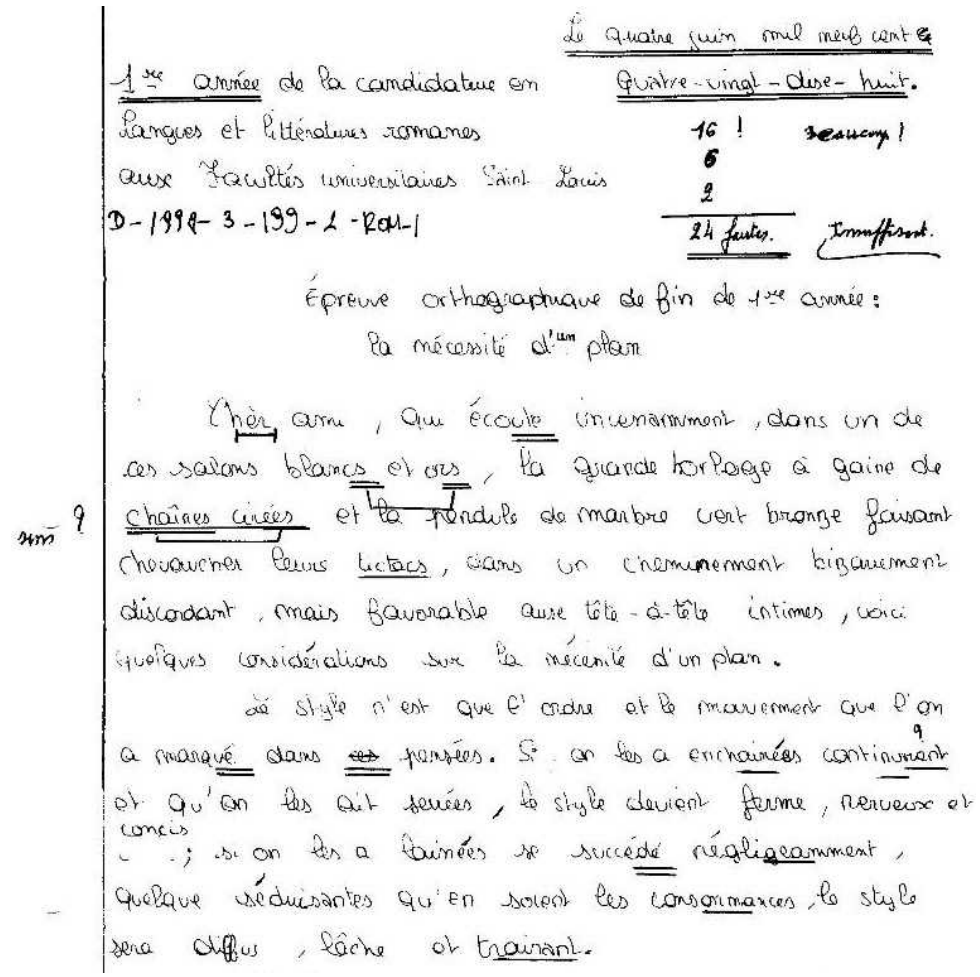
copie). Pour chaque pile de dictées, on encode également le nom de l'étudiant responsable du traitement du corpus[11]. Une fois le traitement des métadonnées effectué, ce sont les textes eux-mêmes qu'il s'agit de formater.

B. *Informatisation des textes manuscrits*

A chaque session d'examen, Michèle Lenoble-Pinson a soigneusement sélectionné (et parfois modifié) un texte qu'elle a dicté aux étudiants. Chaque copie d'étudiant, même si elle devrait en principe contenir le même texte, est unique, en fonction les variantes ou les erreurs qui ont été portées au texte original (nous appelons *texte matrice* le texte qui a été dicté). C'est donc chaque dictée originale qui doit être encodée sur support informatique.

Lorsqu'il s'agit de textes tapuscrits (comme un article de journal ou un livre), on peut procéder à un encodage semi-automatique, en utilisant un logiciel de reconnaissance optique de caractères qui opère à partir d'une image scannée du texte imprimé. Pour des textes manuscrits, cette approche n'est pas envisageable, aucun logiciel ne pouvant, sans entrainement préalable, reconnaitre automatiquement des caractères tracés à la main.

Les textes des dictées sont donc encodés manuellement sur support informatique. Afin de faciliter ce travail, nous avons procédé en deux étapes. Premièrement, le texte matrice a été reconstitué, sur la base de différents exemplaires de dictées et éventuellement d'un retour à l'édition du texte[12]. Cette reconstitution du texte dicté comprend également la segmentation du texte en paragraphes et sa ponctuation.


Le quatre juin mil neuf cent quatre-vingt-dix-huit.

1re année de la candidature en Langues et Littératures romaines aux Facultés universitaires Saint-Louis

D-1998-3-199-1-RAL-1

16 ! beaucoup !
6
2
24 fautes. insuffisant.

Épreuve orthographique de fin de 1re année :
La nécessité d'un plan

Cher ami, qui écoute tictaquement, dans un de ces salons blancs et ors, la grande horloge à gaine de chênes cirés et la pendule de marbre vert bronze faisant chevaucher leurs tictacs, dans un cheminement bigaurement discordant, mais favorable aux tête-à-tête intimes, voici quelques considérations sur la nécessité d'un plan.

Le style n'est que l'ordre et le mouvement que l'on a marqué dans ses pensées. Si on les a enchaînées continuement et qu'on les ait servies, le style devient ferme, nerveux et concis ; si on les a traînées se succédé négligeamment, quelque séduisantes qu'en soient les consonances, le style sera diffus, lâche et traînant.

Figure 1 : Exemple de dictée manuscrite[13]

11 Comme on l'explique ci-dessous, le travail d'encodage sur support informatique et d'annotation des textes d'un corpus est un travail délicat et qui requiert une expertise. Outre l'auteur de chaque dictée, il y a donc, à un second niveau, des « auteurs du corpus », responsables de sa mise en forme pour l'exploitation linguistique.

12 Une partie des textes dictés sont repris au livre de M. Grevisse (1982) : *300 dictées progressives commentées.* Certains textes ont été utilisés plusieurs fois à quelques années d'intervalle.

13 Les mots *tictacs*, *enchainées* et *trainant* sont soulignés une fois en vert, ce qui signale la présence d'une forme respectant les

Ce texte matrice est ensuite dupliqué en autant d'exemplaires qu'il y a de dictées d'étudiants. Chaque fichier est nommé en utilisant l'identifiant unique (décrit au paragraphe précédent). Une étape cruciale peut alors commencer : on encode manuellement les informations écrites par l'étudiant dans l'entête de sa dictée[14] ainsi que toutes les différences (par rapport au texte matrice) apportées par l'étudiant dans sa copie, qu'il s'agisse ou non de *fautes*. Cet encodage se fait sur la base de la copie papier, avec l'aide des corrections apportées par l'enseignante.

Une partie des informations disponibles sur la copie papier ne fait pas l'objet d'un encodage informatique. En effet, Michèle Lenoble-Pinson a corrigé ces dictées en suivant un protocole de correction constant, et en attribuant un symbole ou une couleur particulier à chaque type d'erreur. Ces corrections manuelles se font sous la forme de soulignements ou de commentaires (cf. Figure 1) :

- lettres soulignées deux fois en rouge : une faute de grammaire qui aurait pu être évitée en appliquant une règle ;
- lettres soulignées une fois en rouge : une faute d'usage dont, en principe, la consultation d'un dictionnaire de langue permet de faire la correction ;
- barre verticale rouge dans le texte : une faute de ponctuation ;
- lettres ou mots soulignés une fois en vert : des graphies rectifiées (d'après les recommandations de l'Académie française de 1990). Ces graphies étant enseignées et acceptées dans les dictées, elles peuvent cohabiter avec les anciennes.

Il est malaisé d'encoder les corrections des écarts de l'étudiant dans un format de texte simple. Nous avons donc décidé d'en tenir compte lors de l'étape d'annotation des textes du corpus.

Une relecture est indispensable lors de l'encodage des textes sur support informatique. En effet, la version informatique doit être parfaitement équivalente à la version manuscrite, et il faut éviter d'introduire toute forme de variation lors de l'encodage. La conséquence serait un manque de fiabilité du corpus constitué. Il faut en outre savoir que, une fois accomplie, on ne peut pas revenir sur cette étape, car les textes encodés sur support informatique sont importés dans un logiciel pour être annotés. Après cette importation, aucune modification ne peut leur être apportée.

C. *Annotation des textes*

L'informatisation des textes manuscrits, si elle est indispensable à leur analyse, n'est cependant pas suffisante. Elle permet de constituer une collection de textes « bruts » et documentés. C'est l'enrichissement de ces textes au moyen d'annotations qui va permettre leur exploitation.

L'annotation de corpus est une pratique désormais répandue en linguistique de corpus et en traitement automatique des langues. Elle consiste à « expliciter des informations linguistiques jusqu'alors implicites dans le matériau, en y ajoutant des données métalinguistiques » (Salmon-Alt, Romary et Pierrel, 2004). En procédant à une analyse manuelle ou automatique du texte, on cherche à révéler certaines caractéristiques qui ne sont pas accessibles directement en surface et on les enregistre de telle manière qu'elles puissent servir ultérieurement dans l'exploitation informatique du corpus (l'annotation peut viser différents niveaux d'analyse : les parties du discours, les structures syntaxiques, les relations anaphoriques, etc.). Ces annotations apportent donc une valeur ajoutée au corpus, car elles permettent d'accéder à d'autres niveaux d'analyse et surtout d'automatiser les recherches et dénombrements de phénomènes particuliers.

En ce qui nous concerne, l'annotation des textes répond à deux types d'objectifs. L'annotation vise premièrement à rendre le texte « lisible » et exploitable par le logiciel d'analyse[15]. Le balisage d'un texte vise à fournir des informations générales à propos du texte (type de texte, date de production, etc.), ainsi que des informations sur la structure du texte (où commencent et où finissent l'entête, le titre, chaque paragraphe dans le texte ?). Ce sont des informations dont le linguiste aura besoin pour interpréter les résultats de son analyse.

---

recommandations de la nouvelle orthographe.

14 Outre son nom et la date, l'étudiant est tenu d'écrire le nom et l'adresse de la faculté, ainsi que de mentionner (à partir de 1993) s'il fait ou non usage de la nouvelle orthographe. Ces informations sont donc différentes sur chaque dictée et sont localisées dans l'entête.

15 Les logiciels les plus couramment utilisés permettent de dénombrer les occurrences ou les types dans un corpus, de fournir des concordances, etc. Il s'agit donc de dénombrements, de tris, de classements, sur des formes lemmatisées ou non.

A un second niveau, on souhaite ajouter des informations issues d'une préanalyse du texte. Dans le cas qui nous occupe, l'annotation la plus importante concerne les types de variations introduites par les étudiants dans les dictées. Par *variations* nous entendons à la fois les erreurs de grammaire ou d'orthographe d'usage, mais aussi la segmentation des mots graphiques due aux passages à la ligne, les formes orthographiées ou non selon la nouvelle orthographe, les modifications de casse, etc.

Nous avons donc établi une typologie pour annoter ces variations (essentiellement des erreurs), en nous basant sur la littérature existante (Catach, Duprez et Legris 1980 ; Granger 2003a). On attribue à chaque item concerné une étiquette indiquant s'il s'agit une erreur grammaticale (accord d'un verbe, d'un nom ou d'un participe passé, etc.) ou d'orthographe d'usage (erreur dans l'orthographe d'un mot, confusion entre deux homonymes, etc.). On annote également les modifications de la ponctuation et les formes omises ou ajoutées de manière indue par les étudiants. Enfin, on signale systématiquement si la forme est concernée par la réforme de l'orthographe de 1990.

Dans une certaine mesure, notre travail se rapproche de l'étiquetage des erreurs dans les corpus d'apprenants : comme dans le projet *French Interlanguage Database* (FRIDA : Granger 2003a) ou dans le projet *International Corpus of Learner English* (ICLE : Granger 2003b). À la différence des données à notre disposition, il s'agit de corpus de productions en langue seconde[16] et ayant généralement la forme de textes libres (*vs* dictés). Il s'agit de rédactions dans lesquelles peuvent apparaitre des problèmes de construction syntaxique, de structuration du discours, de choix lexical, de niveau de langue, etc. Dans le cas des dictées, la variété des erreurs est nécessairement plus limitée puisque, par définition, la dictée suppose la conformité au texte de référence qui est lu. On trouvera donc principalement dans nos textes des erreurs d'orthographe d'usage ou d'orthographe grammaticale mais aussi, en quantité moindre, des phrases inachevées, des substitutions de mots, etc. Le jeu d'étiquettes que nous utiliserons sera donc plus limité que celui proposé par Granger (2003a) pour l'étiquetage de FRIDA.

D. *Logiciel et catégories d'annotation*

L'annotation des textes a été réalisée avec le logiciel CorpusTool[17] (O'Donnel 2007, 2008). Pour chaque texte importé dans le logiciel, trois schémas d'annotation, créés par nos soins (cf. Figure 2), sont appliqués. On réalise une annotation des métadonnées (Meta.xml), de la structure du texte (Structure.xml) et des variations apportées par l'étudiant au texte matrice (Corrections.xml).

16 Même si la grande majorité des étudiants représentés dans notre corpus sont des locuteurs francophones natifs, signalons tout de même la présence d'un petit nombre d'allophones. On en reconnait certains à la nature (et au nombre) des fautes commises ou aux commentaires du correcteur. Mais cette information n'est pas disponible de manière systématique.

17 A la recherche d'un logiciel d'annotation, nous avons testé plusieurs applications parmi lesquelles Callisto, Dexter Coder, Sacodeyl, etc. Nous avons retenu le logiciel CorpusTool parce qu'il est gratuit, facile d'utilisation (la définition des schémas d'annotation se fait de manière interactive et visuelle – plus pratique qu'une DTD XML) et qu'il dispose d'outils de recherche et de consultation permettant de passer directement à l'exploitation du corpus une fois celui-ci étiqueté. Il offre également la possibilité d'obtenir des informations statistiques sur l'ensemble du corpus ou sur certaines catégories d'annotation. Enfin, les annotations de chaque schéma sont enregistrées dans des fichiers séparés en adoptant un principe de l'annotation « débarquée » (ou *stand-off annotation*). CorpusTool peut être téléchargé à l'adresse http://www.wagsoft.com/CorpusTool/.

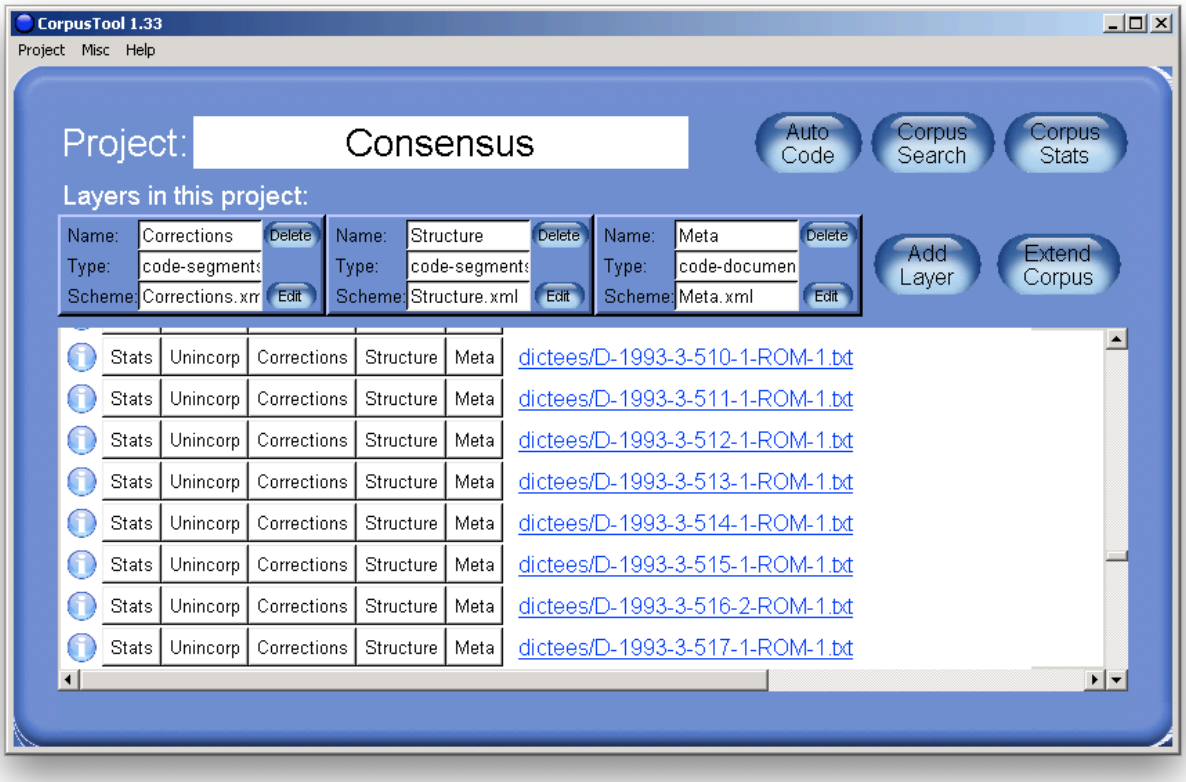


Figure 2 : Logiciel CorpusTool (liste des textes annotés)

La couche d'annotation des métadonnées (cf. Figure 3) est redondante avec l'information encodée dans la base de données Access (cf. description ci-dessus). Nous avons cependant décidé d'ajouter ce niveau d'annotation pour exploiter les métadonnées au sein même du logiciel CorpusTool. Son moteur de recherche permet en effet de faire des recherches en croisant les annotations provenant de différents niveaux. Dans notre cas, on peut donc combiner automatiquement les critères linguistiques et sociolinguistiques (pour obtenir par exemple, un décompte par type d'erreur et par sexe). Notons au passage que les catégories du schéma d'annotation Meta.xml sont affectées à l'ensemble du texte et non à des segments de texte, comme c'est le cas avec les deux schémas suivants.

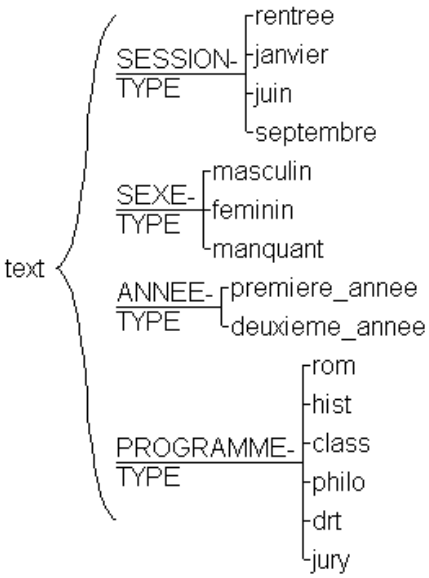


Figure 3 : Schéma d'annotation des métadonnées

Le deuxième schéma d'annotation (Figure 4) permet de baliser la structure du texte et d'identifier les différentes zones de la dictée : dans ce que nous appelons l'entête, on trouve la date, l'année d'étude, la

session, l'adresse de l'université et, après 1990, une information précisant si l'étudiant applique ou non la nouvelle orthographe (pour des raisons évidentes de confidentialité, nous n'avons pas encodé le nom de l'étudiant). Dans le corps du texte, on balise le titre ainsi que chaque paragraphe séparément. Cette approche permet d'étudier de manière contrastive les différentes parties de la dictée afin de vérifier, par exemple, s'il existe des zones particulières qui concentrent plus d'erreurs.

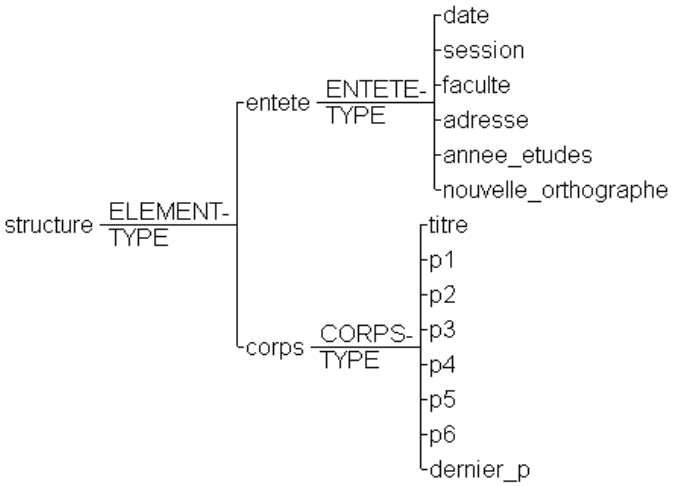


Figure 4 : Schéma d'annotation de la structure du texte

L'interface d'annotation est illustrée à la Figure 5. On y voit que le titre de la dictée, « Épreuve orthographique de fin de 1ere année : la nécessité d'un plan », a reçu l'étiquette *structure/corps/titre*. Pour affecter ces annotations, l'utilisateur surligne la portion de texte à annoter, puis il sélectionne dans les menus du bas de la fenêtre les étiquettes qui conviennent. Le logiciel enregistre les annotations de chaque schéma dans un fichier XML indépendant (cf. Figure 6).

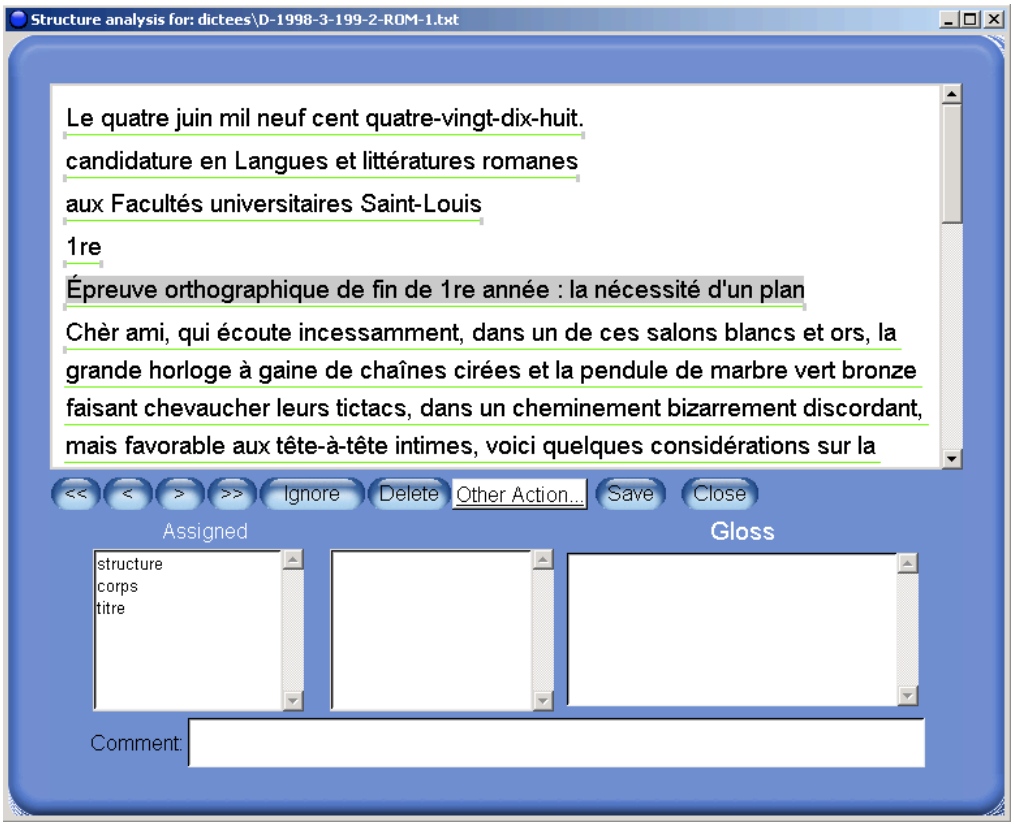
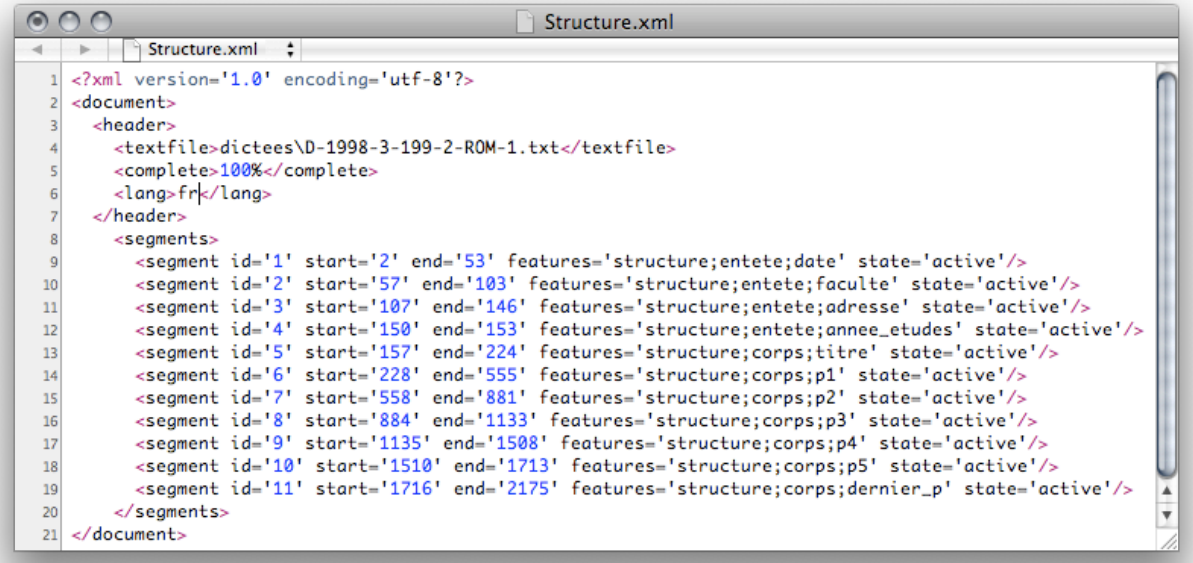


Figure 5 : Logiciel CorpusTool : interface pour l'annotation de la structure


Structure.xml
Structure.xml
<?xml version='1.0' encoding='utf-8'?>
<document>
<header>
<textfile>dictees\D-1998-3-199-2-ROM-1.txt</textfile>
<complete>100%</complete>
<lang>fr</lang>
</header>
<segments>

</segments>
</document>

Figure 6 : Fichier d'annotation en XML

Le schéma d'annotation le plus important et le plus complexe est bien entendu celui des corrections (Figure 7). On y retrouve assez naturellement des catégories qui respectent l'esprit du système de correction utilisé par M. Lenoble-Pinson (cf. II.B. *supra*).

Notre classification repose sur des choix qui pourraient être discutés : par exemple, les problèmes de casse ont été rangés dans la catégorie *orthographe d'usage* alors que ce phénomène est régi par un certain nombre de règles[18]. Autre exemple, nous avons choisi de regrouper ensemble toutes les erreurs portant sur l'accord du verbe, à l'exception du participe passé. Ce choix relève naturellement d'une stratégie pragmatique : plutôt que de viser un étiquetage très fin (qui reposerait exclusivement sur l'expertise et la concentration des encodeurs), on préfère utiliser des catégories d'encodage relativement larges et faciles à distinguer les unes des autres. Il reviendra à l'utilisateur final du corpus de sous-catégoriser les erreurs si cette finesse de description est justifiée par son étude. L'identification d'une catégorie particulière pour le participe passé se justifie par la difficulté que constituent ses règles d'accord en français.

18 On distingue deux problèmes liés à la casse : l'usage abusif d'une majuscule (casse_us) et l'absence d'une majuscule pourtant nécessaire (casse_uj).

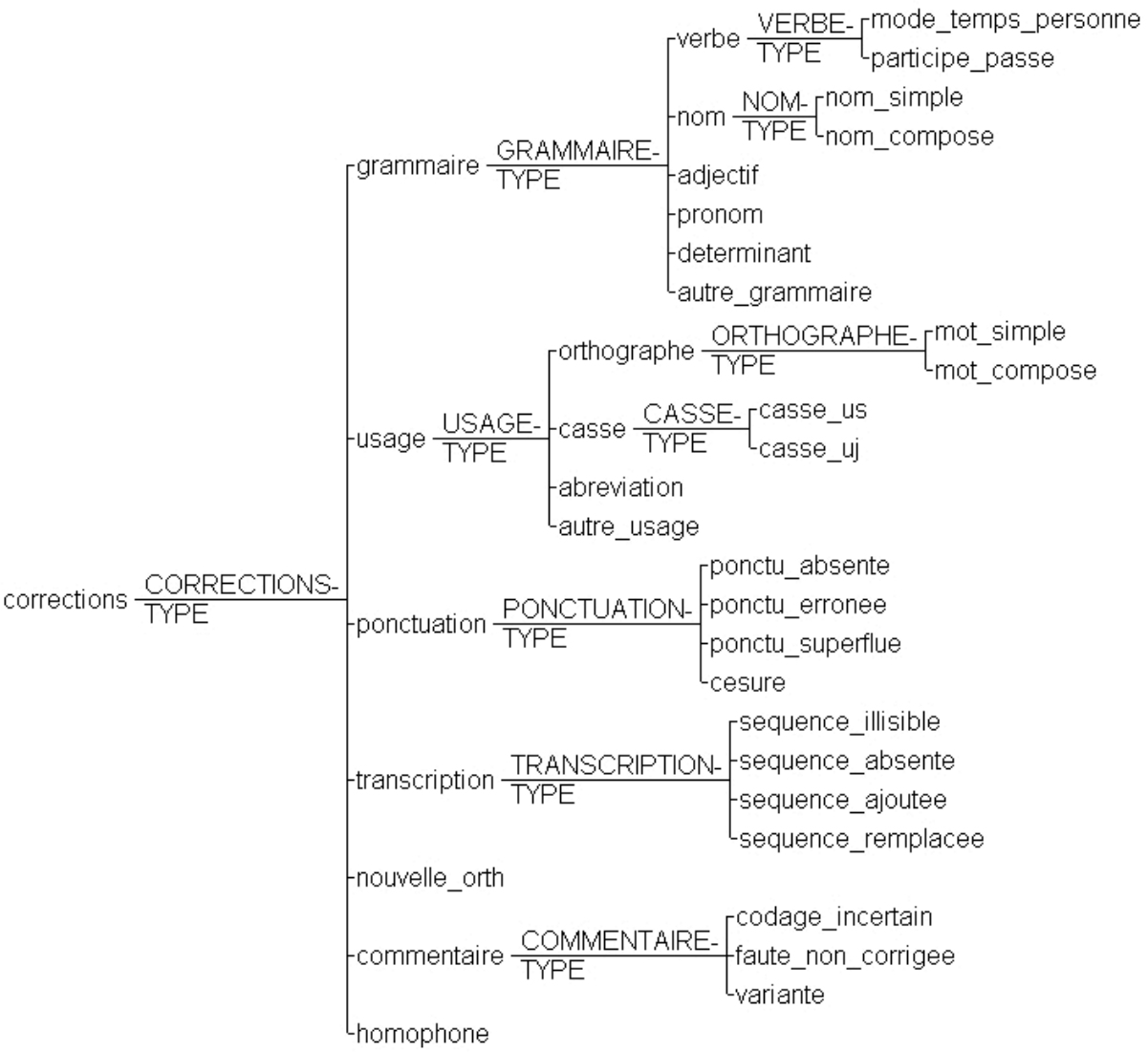


Figure 4 : Schéma d'annotation des corrections

La catégorie « transcription » couvre les cas dans lesquels le texte saisi par l'étudiant ne correspond pas au texte lu par le professeur ou est illisible. La catégorie « commentaire » permet de signaler des erreurs qu'on ne sait pas comment coder (*codage_incertain*). Elle permet également de signaler des erreurs non corrigées (dans certains cas, au-delà d'un certain nombre d'erreurs, la dictée n'est plus systématiquement corrigée) ou des variantes qui ont été tolérées par le professeur (par exemple, un pluriel facultatif). Enfin, la catégorie *homophone* permet de repérer des erreurs commises pour des raisons (présumées) d'homophonie (que l'erreur concerne l'usage ou la grammaire) : *verre / vers* ; *quelque / quelle que / quels que*, etc. Cette annotation est ajoutée à l'annotation déterminant le type d'erreur (le logiciel autorisant la superposition de plusieurs annotations).

E. *Description du corpus annoté*

Les archives transmises au Cental par Mme Lenoble-Pinson contiennent les dictées et les rédactions réalisées par les étudiants du premier cycle universitaire des Facultés universitaires Saint-Louis durant la période 1967-2008. Elles concernent principalement des étudiants en Langues et littératures romanes et classiques (pour un nombre restreint d'années) suivant le cours de *Grammaire du français moderne*[19].

19 Ce cours avait un volume de 60h jusqu'en 200X et de 30h ensuite. En première année, il était évalué de manière principale par une dictée et de manière secondaire par une rédaction (destinée à montrer que l'étudiant est capable d'écrire sans faute quand le texte n'est pas imposé). En deuxième année, l'évaluation principale était la rédaction et l'évaluation secondaire, la dictée. Celle-ci avait cette fois pour but de vérifier que l'étudiant conservait bien sa maitrise de l'orthographe et de la matière vue en première année.

Jusqu'à présent, nous avons exclusivement travaillé sur les dictées : un corpus de 605 textes provenant de 22 sessions d'examens[20] ont ainsi été annotés. Cela représente un total d'environ 200 000 mots[21]. Au sein de ce corpus, 13 255 étiquettes signalant des variations par rapport au texte matrice ont été apposées.

Afin de garantir un équilibre du corpus et une comparabilité entre les années, nous avons essayé de couvrir l'ensemble de la période (1967-2006[22]) en choisissant systématiquement les dictées de la session de juin.

III. Premiers résultats de recherche

Que nous dit ce corpus sur la maitrise de la langue française écrite par les étudiants de première année universitaire entre 1969 et 2006 ?

A. *Observations générales*

Le corpus annoté représente 605 textes. Les erreurs ont été classées selon quatre grands types : grammaire (erreurs d'accord et de flexion, des noms, verbes, adjectifs, etc.) ; usage (orthographe d'usage, y compris les majuscules et les abréviations); ponctuation (erronée, absente, etc.) et les erreurs de transcription (séquence illisible, absente ou ajoutée). En outre, on annote les formes qui appliquent les recommandations de la réforme orthographique de 1990[23].

Tableau 1. Répartition des types d'erreurs annotées

Type d'erreur	**Nombre d'occurrences**	**Pourcentage**
Grammaire	4025	32,7%
Usage	5190	42,2%
Ponctuation	2749	22,4%
Transcription	328	2,7%
Total	**12292**	**100%**

Parmi les erreurs de grammaire, 42,9 % concernent l'accord des verbes (dont un peu plus de la moitié pour les seuls participes passés), 21,6 % concernent l'accord des noms et le même nombre (21,7 %) concerne l'accord des adjectifs. Les erreurs concernant les pronoms, les déterminants et les autres formes se partagent les 14,8 % restants. Pour ce qui est des erreurs d'usage, la majorité concernent l'orthographe et un pourcentage réduit, la casse (15,3%).

Une rapide approximation permet de voir que, en moyenne, chaque texte du corpus comporte 20,3 formes annotées (y compris les erreurs de ponctuation et de transcription). La question se pose de savoir si, au fil du temps, on observe une évolution (augmentation ou diminution) du nombre d'erreurs par dictée.

B. *Évolution dans le temps du nombre d'erreurs par dictée*

Une première réponse est apportée par Séverine Jaspard (2008) qui analyse le décompte des erreurs effectué par l'enseignante et encodé dans la base de données Access. Sur l'intervalle de 30 années couvert par le corpus, Séverine Jaspard observe un premier pic dans les années 1986-1987 et un second, plus important, dans les années 2001-2006 (faut-il voir un lien avec la diminution du volume d'heure du cours de grammaire qui est passé de 60 à 30h ?).

Cette mesure se base sur la moyenne d'erreurs corrigées par l'enseignante dans chaque dictée, pour chaque année encodée. Cette moyenne varie de 6 à 12 erreurs par dictée entre 1969 et 1985. En 1986 et 1987, la moyenne monte à une vingtaine d'erreurs par dictée. Elle monte légèrement entre 1988 et 1999, puis subit une brusque augmentation en 2001 (36 erreurs par dictée), qui se maintient globalement par la suite.

---

20 Années 1969 à 1972, 1975, 1976, 1979, 1985 à 1993, 1998, 1999, 2001, 2002 et 2004 à 2006.

21 Ce chiffre dépend naturellement de la méthode d'identification des mots (*tokenisation*) : le logiciel WordSmith Tools dénombre 205 022 mots (*tokens*) et le programme Unix *wc* en dénombre 191 013. Fait remarquable qui s'explique parfaitement par la nature de ce corpus, il y a seulement 4454 formes différentes (*types*), selon WST. Chaque texte se répétant autant de fois qu'il y a d'étudiant à une session d'examen, le vocabulaire croît bien plus lentement que dans d'autres types de corpus.

22 Quand le travail dont nous rendons compte a commencé, les années 2007 et 2008 n'étaient pas encore en notre possession.

23 680 annotions désignant des formes écrites par les étudiants dans la nouvelle orthographe ont été enregistrées. Il s'agit uniquement des formes identifiées par le professeur dans son code de correction (un soulignement vert).

Cette analyse est complétée par une mesure de l'écart-type : plus l'écart-type est important, moins la classe d'étudiants est homogène. L'écart-type augmente de manière régulière jusqu'en 1984 (il passe de 3 à 9) ; il redescend ensuite pour se stabiliser autour de 5, mais on observe un pic en 1998 (16) et un maintien à une valeur élevée. D'une manière globale, l'hétérogénéité entre les étudiants ne cesse d'augmenter. À partir de la fin des années nonante, les très bons résultats côtoient les mauvais. Les données sociolinguistiques (manquantes) ne nous permettent pas d'observer si cette hétérogénéité croissante va de pair avec une diversification du public étudiant.

C. *Comment mesurer la difficulté d'une dictée*

Le corpus n'a pas été annoté morphosyntaxiquement. Il n'est donc pas possible, actuellement, de comparer le degré de difficulté des dictées entre elles, en se basant par exemple sur le nombre de participes passés ou de mots composés qu'elles contiennent, et qui sont réputés particulièrement difficiles. Cependant, Nathalie Dehaut (2008) montre qu'« il y a un rapport entre la fréquence d'un mot et le fait que celui-ci est mal orthographié ». Elle établit une liste des 110 lemmes les plus fréquemment mal orthographiés dans le corpus et mesure ensuite la probabilité d'apparition de chaque forme grâce à la base de données Lexique3[24] (en ne retenant que le corpus de sous-titres de films, supposé comporter des mots relativement généraux). Sur les 110 lemmes dont l'orthographe pose problème aux étudiants, 54 (soit 49 %) font partie des emplois très rares, 32 (soit 29%) ont une fréquence moyenne et 24 (soit 22 %) sont très fréquents[25]. Plus de la moitié des mots mal orthographiés font ainsi partie du groupe des mots rares.

Conclusion

Cet article ne vise pas à répondre à la question posée régulièrement du niveau de maitrise de la langue écrite par les jeunes. D'une part, le corpus ne permettrait pas de répondre à cette question, car « les jeunes » qui sont représentés constituent une population très spécifique de jeunes adultes entamant des études universitaires de lettres ou de sciences humaines dans une université bruxelloise. D'autre part, notre propos visait davantage à expliciter les étapes qui rendent une collection de textes exploitable pour l'analyse linguistique.

Il n'en reste pas moins que l'on perçoit aisément l'intérêt de la ressource constituée : elle permet de jeter un regard longitudinal sur la maitrise de l'orthographe au sein d'une sous-population particulière (il faut cependant rester prudent dans la mesure où la composition de cette population est susceptible d'évoluer au cours du temps, par exemple en raison de la démocratisation des études). En quelques clicks, on peut très aisément rassembler un grand nombre d'occurrences de problèmes orthographiques spécifiques, ce qui manuellement aurait nécessité un temps de dépouillement très long. De nombreuses études pourront donc être réalisées à partir de ce corpus.

Soulignons également l'intérêt de ce corpus pour les applications de traitement automatique du langage : un corpus étiqueté tel que le nôtre doit permettre à des programmes informatiques de faire un « apprentissage » et de construire un « modèle d'erreurs » qui enregistre les fautes les plus fréquentes ou les plus typiques en fonction d'un contexte particulier. Un tel modèle pourrait être utile pour tester ou améliorer le fonctionnement d'un logiciel de correction orthographique ou concevoir des logiciels de génération d'exercices orthographiques.

24 http://www.lexique.org/

25 Parmi les mots rares, on peut citer *rhéteur, ochracé, diaconnesse, mellifère, rasséréner*; parmi les mots à fréquence moyenne, *précipitamment, abysse, bizarrerie, airelle*; parmi les mots fréquents: *soi-disant, phénomène, apercevoir, quelquefois, bouleverser, atmosphère, mépriser*, etc.

Références bibliographiques

BAUDE (Olivier), (éd.), *Corpus oraux. Guide des bonnes pratiques*, CNRS Éditions, Paris.

BENA, (Jonas Makamina), *Les déficits en matière de français-langue maternelle : diagnostic et base de remédiation*, L'Harmattan, Paris-Budapest-Turin, 2004, p. 128-160.

CATACH (Nina), DUPREZ (Daniel) et LEGRIS (Michel), *L'enseignement de l'orthographe*, Nathan, Paris, 1980.

CHERVEL (André), *La dictée: les Français et l'orthographe*, Institut national de recherche pédagogique, Paris, 1989.

DAVID (Jacques) et GONCALVES (Harmony), « L'écriture électronique, une menace pour la maitrise de la langue ? », *Le Français aujourd'hui*, nº 156, Armand Colin, Paris, 2007.

DEHAUT (Nathalie), « Analyse d'un corpus de dictées: influence de la fréquence d'un mot dans la langue sur l'orthographe d'usage ». Travail réalisé dans le cadre du cours CLIG2250, Manuscrit non publié, Université catholique de Louvain, 2008.

DIDIER (Jean-Jacques), HAMBURSIN (Olivier), MOREAU (Philippe) et SERON (Michel), *Le français m'a tuer*, coll. Cahiers du Cental n°1, Presses universitaires de Louvain, Louvain-la-Neuve, 2007.

GAGNÉ (Gilles), « Pédagogie de la langue ou pédagogie de la parole ? », in Gagné (Léo), (éd.), *La qualité de la langue après la loi 101. Actes du Colloque, Québec 30 septembre - 3 octobre 1979*, Editeur officiel du Québec. [accessible en ligne sur : http://www.cslf.gouv.qc.ca/publications/PubD103/D103-IIa.html, visité le 16/10/2008]

GRANGER (Sylviane), « Error-tagged Learner Corpora and CALL: A Promising Synergy », *Calico*, vol. 20, n ° 3, 2003a, [accessible en ligne sur : http://www.calico.org/, visité le 16/10/2008].

GRANGER (Sylviane), « The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research », *TESOL Quarterly*, vol. 37, n°3, 2003b, p. 538-546.

GREVISSE (Maurice), *300 dictées progressives commentées*, Duculot, Louvain-la-Neuve, 1982.

HACHEZ (Théo) et WYNANTS (Bernadette), *Les élèves du secondaire et la norme du français écrit*, coll. « Français & Société », n° 3, Service de la langue française, Direction générale de la culture et de la communication, Bruxelles, 1991.

HOPPER, (Christophe), *Depuis quand date la crise de l'enseignement du français ?*, Montréal : PPMF élémentaire, Université de Montréal, 1975.

JALABERT (Romain), « MoliR, rev1 vit… il son 2vnu foo ! », In *L'orthographe*, Dossier des Cahiers pédagogiques, n° 440, Cercle de recherche et d'action pédagogiques, Paris, 2006.

JASPARD (Séverine) « Évolution du nombre d'erreurs commises dans le corpus de dictées ». Travail réalisé dans le cadre du cours CLIG2250, Manuscrit non publié, Université catholique de Louvain, 2008.

MC ENERY (Tony), XIAO (Richard) et TONO (Yukio), *Corpus-Based Language Studies. An advanced resource book*, coll. Routledge Applied Linguistics, Routledge, London, New York, 2006.

O'DONNELL (Mick), *UAM CorpusTool*, Version 1.2 User Manual (December 2007) [accessible en ligne sur : http://www.wagsoft.com/CorpusTool/documentation.html, visité le 5/09/2008].

O'DONNELL (Mick), « Demonstration of the UAM CorpusTool for text and image annotation », Proceedings of the ACL-08:HLT Demo Session (CompanionVolume), Columbus, Ohio, June 2008, Association for Computational Linguistics, 2008, p. 13-16.

PAVEAU (Marie-Anne) et ROSIER (Laurence), *La langue française. Passions et polémiques*, Vuibert, Paris, 2008.

SALMONT-ALT (Suzanne), ROMARY (Laurent) et PIERREL, (Jean-Marie), « Un modèle générique d'organisation de corpus en ligne : application à la *FReeBank TAL* », *T.A.L.*, vol. 45, n°3, 2004, p. 145-169.