Avant d'aborder les écrits des apprenants chinois, il nous semble pertinent de nous expliquer d'abord sur les choix méthodologiques. En vue d'atteindre les objectifs décrits dans l'introduction de cette étude, il a été indispensable de créer un corpus écrit permettant d'observer les pratiques orthographiques des apprenants chinois du français L2. Étant donné que la création d'un corpus relève de maints paramètres pratiques et méthodologiques, nous entamons, dans ce chapitre, une discussion sur les considérations méthodologiques, ainsi que sur les choix des méthodes pour la constitution du corpus et les analyses de données. Cette discussion nous semble tout à fait importante pour la compréhension de ce travail et pour l'interprétation des données.

3.1. Considérations méthodologiques

Les données sont fondamentales dans les recherches en science du langage. Au cours de notre travail sur les données nous avons dû prendre position par rapport aux différents aspects méthodologiques que nous discutons brièvement ici. Premièrement, le recueil des données (3.1.1.) : il s'agit d'adopter une méthode adéquate aux objectifs de notre recherche tout en tenant compte des contraintes et problématiques propres aux apprenants du français L2. Nous évoquerons ensuite les problèmes méthodologiques qui s'imposent à l'analyse des données (3.1.2.) et les choix théoriques ayant permis leur description et la mise au jour des résultats.

Un dispositif pour observer les performances au cours de la production écrite

La production langagière est « une habileté spécifiquement humaine » (Segui & Ferrand, 2000). Si la production orale est extrêmement rapide et efficace (Fayol, 2002),

l'écriture demande plus d'effort à des scripteurs que celle à l'oral, et elle « s'avère laborieuse et parfois avec un succès qui n'est pas garanti » (Bonin, 2002). Notons que la production verbale sous la modalité écrite n'a fait, comme l'écrit Fayol (2007), que très récemment l'objet de recherches scientifiques, plus récemment encore quand il s'agit de son acquisition, contrairement à l'intérêt porté à la production verbale orale ou à la lecture. Malgré ce relatif retard, les études consacrées à la production verbale écrite ont augmenté de façon significative au cours des trois dernières décennies. Dans le cadre spécifique du français, le groupe de recherche GDR CNRS 2567[19] se situe dans ce contexte, en fédérant les différents chercheurs nationaux impliqués dans l'étude de la Production Verbale Écrite. Toutefois, dans le cadre spécifique du français L2, les recherches conduites sur la production écrite sont en effet relativement rares quand le système graphique de la langue source est très éloigné de celui du français, tels que le chinois et le français, que rien ne semble rapprocher.

Comment les locuteurs de chinois font-ils pour apprendre à orthographier en français ? Comment gèrent-ils les différences entre deux systèmes graphiques si opposés ? Comment s'approprient-ils le français dans « un milieu scolaire (universitaire) complètement coupé du contexte social de la langue cible » (Trevisiol, 2003) ? Ce fut le point de départ qui a motivé notre travail. Nous souhaitons ainsi créer un dispositif pour observer les performances orthographiques des apprenants chinois étudiant le français en Chine.

3.1.1.1. Processus de mise en texte

Depuis le modèle fondateur de Hayes et Flower (1980), il est généralement admis que « la production verbale écrite met en œuvre des processus récursifs qui sollicitent une activité langagière complexe et de haut niveau » (Plane, Rondelli & Venerin, 2013). Parmi d'abondants travaux sur la production écrite, on voit un intérêt particulier porté sur l'un des processus d'écriture, la révision, surtout dans le champ de la psychologie

[19] http://www.gdr-pve.fr/

cognitive, qui a pour objectif d' « affiner les formalisations de l'activité scripturale en s'appuyant sur des expérimentations menées dans un cadre très contrôlé » (Plane, 2015). Pourtant, l'autre grand processus, la textualisation est rarement abordé ou du moins partiellement étudiés par les chercheurs. On voit notamment les recherches menées par les linguistes à travers les analyses génétiques des brouillons d'écrivains (Lumbroso, 2004 ; Hamon, 2009) ou des textes d'élèves (Fenoglio & Boucheron-Pétillon, 2002 ; Doquet-Lacoste, 2003). En essayant de reconstituer la genèse d'un texte, ces études se sont notamment attachées à caractériser les singularités des scripteurs, et il nous semble donc particulièrement intéressant d'enrichir les connaissances portant sur les variations dans les processus de mise en texte par des apprenants-scripteurs, pour identifier plus précisément les différentes manières dont ils procèdent dans ce processus et des types de difficultés rencontrées, afin, dans une perspective didactique, de construire des dispositifs d'enseignement différenciés et appropriés (Plane, Rondelli & Venerin, 2013).

3.1.1.2. Interlangue

Si l'étude porte sur les variations dans les processus de mise en texte par des apprenants-scripteurs d'une langue étrangère, le travail serait enrichi par un autre élément très important, l'interlangue, qui est, selon Selinker (1972), « un système linguistique productif par référence aux deux autres systèmes linguistiques que constituent la langue maternelle de l'apprenant et la langue étrangère » (Rosen & Porquier, 2003). De ce fait, pour mettre en lumière les processus impliqués dans la production langagière chez un apprenant allophone, il faut prendre en compte les types et la quantité de l'input, ainsi que la perception de l'apprenant de l'input à travers ses connaissances linguistiques préalables (Perdue, 1984 :26), qui illustre la complexité des facteurs à considérer quand il s'agit d'expliquer l'acquisition d'une langue étrangère. Comme l'indique Benazzo (2005), « la prise en compte d'un seul type de facteur, qu'ils soient d'ordre purement linguistique ou d'ordre purement cognitif », ne permet pas de mettre en lumière l'acquisition d'un élément de la langue étrangère. C'est l'interaction de ces facteurs, variable suivant le stade atteint par l'apprenant, qui apporte des éclaircissements sur le

processus acquisitionnel.

3.1.1.3. Deux contraintes mises en jeu dans la production de texte

Dans la production écrite, il existe deux niveaux de difficultés auxquelles les apprenants-scripteurs d'une langue étrangère seraient confrontés : d'une part celles que pose le système linguistique, venant de l'input, ainsi que de leurs connaissances linguistiques préalables ; d'une part celles de la psycholinguistique, venant de la tâche d'écriture, d'autre part. Ces difficultés sont antagonistes, non hiérarchisables entre elles, et offrent des degrés de liberté différents (Plane, Olive & Alamargot, 2010).

3.1.1.4. Bilan

Étant donné les deux niveaux de contraintes mises en jeux dans la production écrite, il nous semble important de proposer un dispositif empruntant aux deux champs, pour observer les variations orthographiques produite par les apprenants chinois du français L2 dans une production écrite réelle. En vue d'éliminer, dans la mesure que possible des variables interférentes, un certain nombre de facteurs doivent être pris en compte lors de la création du dispositif pour recueillir les données :

- La langue maternelle de tous les apprenants est la même et ils ont approximativement le même âge.
- Les apprenants ont un profil linguistique similaire en ce qui concerne les langues étrangères qu'ils ont maîtrisées.
- Les apprenants choisis dans la recherche étudient le français dans un même établissement.
- A partir d'une perspective développementale, il nous semble important de recueillir les données auprès les apprenants se trouvant à des niveaux langagiers différents.
- Les textes écrits recueillis doivent être issus de la même tâche d'écriture.
- Pour faciliter une analyse interindividuelle, il nous semble intéressant que les

apprenants fassent une tâche d'écriture demi-spontanée à partir d'un même matériau écrit.

- Tous les textes doivent être écrits dans une salle de classe, avec la présence du professeur des apprenants.
- L'expérimentateur doit être le même.
- Pour diminuer le stress des apprenants devant une tâche difficile, l'accent doit être mis sur l'importance du contenu et non de la forme doit être accentuée.
- Il faut préciser que la tâche à accomplir n'est pas évaluée. Le résultat ne comptant pas pour la note du français.
- Le temps pour accomplir la tâche doit être le même pour les différents groupes d'apprenants.

En examinant leurs textes écrits, ainsi que leurs performances orthographiques, nous souhaitons identifier les traitements orthographiques mis en œuvre par les apprenants chinois lors d'une production écrite en français, et caractériser l'interlangue de ces apprenants dans la production écrite par sa variabilité et sa systématicité.

3.1.2. Réflexions sur l'analyse des données en interlangue

Dans le domaine de l'acquisition et de l'apprentissage des langues secondes, les différents types d'analyse des données sont utilisés par les chercheurs afin de mieux comprendre le processus d'acquisition. Si les théories adoptées par les études qualitatives constituent souvent un biais lié à la subjectivité du chercheur, les méthodes, quantitatives et le recours aux statistiques tentent de réduire le biais du traitement des données. Toutefois, ces méthodes ont elles-mêmes leurs limites pour interpréter les données. De fait, l'articulation entre les données et la méthodologie ne s'avère pas toujours aller de soi. Cela nous oblige à questionner sans cesse la méthodologie à travers les objectifs de la recherche. Telle est l'orientation que nous entendons donner à la partie suivante, en nous attachant particulièrement à une présentation des méthodes différentes utilisées pour le traitement des données dans le domaine de l'acquisition de

langues secondes.

3.1.2.1. Analyse comparative

La linguistique contrastive est née dans les années 1950. En considérant l'interférence causée par les différences structurelles entre la langue maternelle de l'apprenant et une langue étrangère comme un des principaux obstacles à l'apprentissage de cette langue étrangère, les premières études contrastives sont suscitées dans une perspective d'application. Fries (1945: 9) indique que « the most effecient materials are those that are based upon a scientific description of the language to be learned, carefully compared with a parallel description of the native language of the learner ».

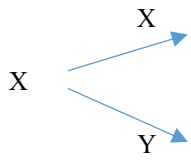
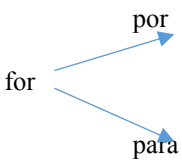
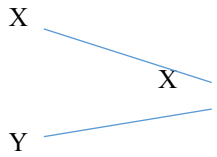
Notons que la linguistique contrastive est étroitement liée à son utilité. Comme l'indique Debyser (1970), les études contrastives « peuvent rencontrer en chemin des problèmes théoriques intéressants, permettre la vérification d'hypothèses, et susciter des descriptions qui manquaient mais ne se justifient véritablement que par les services qu'elles peuvent rendre ». Malgré que l'accent des études contrastives soit mis sur la diversité des langues, la théorie est une et permet, en outre, non seulement la description mais aussi la comparaison. En d'autres termes, ces études disposent d' « un instrument théorique et métalinguistique assez général et assez unitaire » pour rendre comparable des langues différentes (Debyser, 1970).

La linguistique contrastive suppose au départ que l'apprentissage d'une langue étrangère ne pose pas les mêmes problèmes que l'apprentissage de la langue maternelle. C'est dans le cadre de cette hypothèse psycholinguistique fondamentale que les recherches contrastives sont entreprises. Afin de répondre aux besoins d'une pédagogie spécifique des langues étrangères, la linguistique contrastive s'appuie principalement sur les études de la nature et du rôle des erreurs dans l'apprentissage d'une langue étrangère. Autrement dit, elle a pour l'objectif « de prévoir, de décrire et d'expliquer les erreurs et les difficultés » rencontrées par les apprenants d'une langue étrangère dues à l'influence de leur langue maternelle. En partant de la conviction que les différences linguistiques pourraient être utilisées pour prévoir les difficultés

d'apprenants, Weinreich (1953 : 1) affirme que plus la distance entre langue cible et langue maternelle est grande, plus les formes et les modèles s'excluant mutuellement dans les deux systèmes sont nombreux, plus les problèmes d'apprentissage de la langue cible s'élèvent. Autrement dit, dans le domaine où les deux langues sont similaires, un transfert positif se produirait, alors qu'un transfert négatif ou une interférence aurait lieu dans le domaine où les deux systèmes se distingue l'un de l'autre.

Cependant, il faut remarquer que les analyses contrastives ne sont pas que des listes de prédictions binaires sous la forme suivante : similitude/différence = facilité/difficulté (Larsen-Freeman & H.Long, 1990 : 53). Le tableau suivant présente une version simplifiée de la hiérarchie de difficulté (Stockwell, Bowen & Matin, 1965a), illustrant une analyse contrastive sophistiquée :

Tableau 3. Hiérarchie de difficulté (Larsen-Freeman & H.Long, 1990 : 54)

Type de difficulté	L1 : anglais ; L2 : espagnol	Exemple
Séparé		
Nouveau	⦸ ---------- X	Faire l'accord du genre grammatical
Absent	X ----------⦸	*Do* comme une marque de temps
Combiné		His/her est réalisé comme une seule forme *su*
Correspondance	X ---------- X	*-ing = -ndo*

Les chercheurs nous montrent, à partir de ce tableau, que le point linguistique le plus facile à maîtriser serait celui où la langue cible correspond à la langue maternelle au niveau de structure et de sens. Progressivement plus difficiles sont les points

linguistiques conflués, où plusieurs formes de L1 convergent en une en L2, la forme qui se présente en L1 mais est absente en L2, ainsi que le point linguistique tout nouveau en L2. Le domaine le plus difficile à maîtriser relève de la division, où une forme en L1 se traduit par deux ou plusieurs formes en L2.

Ajoutons également que dans le cadre de la linguistique contrastive, le travail de recherche comporte systématiquement trois étapes : la première réside à analyser d'une manière précise des structures de chaque langue. La seconde étape à réaliser est d'observer le comportement des deux langues, dans l'exploitation de leur matériel. Précisons ici que « l'investigation doit porter en priorité sur les traductions, c'est-à-dire sur les équivalences d'énoncés » (Delen Karragaç, 2012). La troisième étape est portée sur les prévisions des erreurs éventuelles. A la suite d'une interprétation des erreurs étudiées, certaines conclusions seront proposées.

Néanmoins, lorsque les prédictions provenant des analyses contrastives ont été finalement soumises à des tests empiriques, de graves lacunes ont été constatées (voir, par exemple, Chamot, 1978 ; Arabski, 1979). Comme l'indiquent Long et Sato (1984), « the most fatal flaw of the CAH was the dubious assumption that one could depend solely upon an analysis of a linguistic product to yield meaningful insight into a psycholinguistic process, i.e. second language learning » (cité par Larsen-Freeman & H.Long, 1990 : 56). Malgré les critiques, les analyses contrastives continuent à être menées dans le domaine de l'acquisition de langues étrangère, afin d'identifier les transferts positifs ou négatifs provenant de la langue maternelle.

3.1.2.2. Analyse d'erreurs

En partant des résultats décevants des enquêtes empiriques qui cherchent à vérifier les hypothèses provenant de l'analyse contrastive, Wardhaugh (1970) propose une distinction entre une version forte et une version faible pour examiner ces hypothèses. Comme l'indiquent Larsen-Freeman et H.Long (1990 : 57), pour la version forte, la prédiction d'erreurs dans l'apprentissage d'une langue étrangère est basée sur une analyse contrastive a priori de la langue maternelle et de la langue cible. Et comme nous

l'avons mentionné plus haut, ce type de prédiction n'est pas toujours vérifiée. En revanche, dans la version faible, les chercheurs commencent par les erreurs de l'apprenant et expliquent un sous-ensemble de ces erreurs en faisant appel aux similitudes et aux différences exsistant entre langue maternelle et langue cible. Selon Larsen-Freeman et H.Long, les analyses contrastives pourraient ne pas être utiles a priori, mais elles posséderaient encore un pouvoir explicatif a posteriori. De ce fait, les chercheurs pourraient s'en servir dans une approche plus large d'interpréter les erreurs commises par l'apprenant, à savoir l'analyse des erreurs.

3.1.2.2.1. Erreurs interlinguales *versus* erreurs intralinguales

Notons que l'apprenant L2 commet toujours les erreurs dues à l'impact de sa langue maternelle, et Richards (1971) les appelle erreur interlinguale. Néanmoins, il faut remarquer qu'un grand nombre d'erreurs similaires sont commises par les apprenants L2, indépendamment de leur langue maternelle. Il s'agit donc des erreurs intralinguales, provenant des caractéristiques intrinsèques d'une langue étrangère.

Corder (1967) soutient que les erreurs des apprenants sont de grand intérêt pour l'étude du processus de l'apprentissage de langues. En classant les erreurs commises par les apprenants, les chercheurs pourraient en déduire les stratégies d'apprentissage adoptées par eux. Ainsi sont motivées de maintes études sur les typologies d'erreurs [20] :

- Erreurs relevant de la surgénéralisation (Richards, 1971), où l'apprenant omet de respecter les limites d'une règle. Par exemple : * *I wonder where are you going*. Ici, l'apprenant a probablement surgénéralisé la règle de l'inversion sujet-auxiliaire, et l'a appliquée incorrectement dans une relative.
- Erreurs relevant de la simplification (George, 1972), telles que l'omission de la marque du pluriel pour un nom se trouvant dans un contexte pluriel. Par exemple : * *I study English for two year*.
- Erreurs relevant de la communication (Selinker, 1972). Ce type d'erreur se produit

[20] Nous avons repris les exemples utilisés dans le travail de Larsen-Freeman et H.Long (1990 : 59) pour illustrer les différents types d'erreur.

lorsque l'apprenant a recours à des stratégies communicatives. Par exemple, l'apprenant utilise la forme **airball* pour le mot *balloon*. Dans ce cas, l'apprenant étiquette incorrectement l'objet « ballon », mais il communique avec succès le concept souhaité.

- Erreurs relevant d'une induction (Stenson, 1974). Souvent se produit ce type d'erreur lorsque deux éléments linguistiques sont présentés par l'enseignant d'une manière qui crée une confusion. Par exemple, **She cries as if the baby* (pour *She cries like a baby*). Cette erreur s'explique probablement par le fait que l'expression *as if* est défini comme *like* par l'enseignant, et que l'enseignant n'a pas expliqué la modification structurelle nécessaire que demande l'expression *as if*.

3.1.2.2.2. Faute *versus* erreur

Les tentatives d'identifier et de classer les erreurs commises par les apprenants sont d'une importance significative pour les recherches dans le domaine de l'acquisition langagière, puisqu'elles attribuent un nouveau statut aux erreurs (Larsen-Freeman et H.Long, 1990 : 59).

Corder (1967) fait une distinction entre la notion d'erreur et celle de faute. Les fautes sont liées à la performance, et causées par la fatigue, l'excitation, etc. On a par conséquent la possibilité de faire soi-même la correction parce qu'on connait les règles. En revanche, les erreurs sont liées à la compétence, et relèvent d'une méconnaissance de règles. Autrement dit, l'erreur est souvent un produit involontaire, et l'apprenant ne peut pas la corriger lui-même. Selon Larsen-Freeman et H.Long, les erreurs sont un reflet du stade actuel du développement langagier de l'apprenant (1990 : 59). Ainsi, il est de mise de penser que l'appellation « erreur » renvoie à un intérêt pour le processus d'apprentissage langagier propre à chaque apprenant, et que « faute » est, en revanche, utilisé pour décrire une déviation à la norme.

3.1.2.2.3. Perspective différente de l'analyse contrastive

A la différence de l'analyse contrastive, l'analyse des erreurs permet de réinterpréter le rôle de l'apprenant qui est confronté aux difficultés à apprendre une langue étrangère.

Les erreurs ne résultent davantage d'une intrusion des habitudes L1 sur lesquelles l'apprenant n'a aucun contrôle. Au lieu d'être receveur passif des inputs de la langue cible, l'apprenant est activement engagé dans le processus d'apprentissage : il traite les données, génère les hypothèses, puis les teste et les affine (Larsen-Freeman et H.Long, 1990 : 61). Et l'analyse des erreurs se base donc sur la description puis l'exploitation des traits observés concernant des erreurs produites par l'apprenant, afin d'expliquer le système intermédiaire propre à cet apprenant.

Néanmoins, il faut indiquer qu'en décrivant le processus d'apprentissage de langues étrangère, l'analyse des erreurs présente également les insuffisances. En se concentrant uniquement sur les erreurs, les chercheurs n'ont plus une vue d'ensemble : ils étudient les erreurs commises par l'apprenants, mais pas ses réussites. En outre, il est toujours difficile, voire impossible d'identifier la source unitaire d'une erreur. Etant donné les insuffisances que présente la perspective de l'analyse des erreurs, on constate son incorporation dans l'analyse de performances. Cette dernière vise à décrire les performances interlangagières de l'apprenant et elle ne se limite pas à analyser seulement les erreurs.

3.1.2.3. Analyse de performance et études d'interlangue

Pour un apprenant L2, la performance concerne la mise en œuvre de ses connaissances linguistiques dans les activités communicatives. Il s'agit à la fois de l'émission, où l'apprenant fait des phrases, et de la réception, où l'apprenant comprend des phrases. Dans l'apprentissage de langues étrangère, la performance d'un apprenant est souvent considérée comme une indication indirecte de sa compétence (Richards, 2002). Si l'analyse des erreurs se concentre uniquement sur les produits langagiers de l'apprenant et constitue souvent un biais lié à la subjectivité du chercheur, l'analyse des performances, en tant qu'approche globale, vise à mettre au jour le processus d'apprentissage d'une langue étrangère.

Depuis les années 1970, les tentatives ont été faites pour effectuer une analyse linguistique de la langue de l'apprenant. Nous constatons notamment les études (voir,

par exemple, Dulay & Burt, 1973 ; 1974) sur l'ordre d'acquisition de morphèmes de l'anglais, qui mettent un accent fort sur le produit linguistique. A partir d'une approche plus élucidant, les chercheurs commencent à découvrir comment l'apprenant traite les données langagières afin de les acquérir. Ainsi le glissement général de l'intérêt vers le relevé des processus sous-jacents de l'apprentissage langagier fait-il une transition naturelle vers les études d'interlangue (Ringbom, 1994).

Dans les années 1990 les études d'interlangue connaissent un fort développement. Les sujets de recherche sont : le développement du langage de l'apprenant, les stratégies utilisées par l'apprenant, les variations systématiques de la langue de l'apprenant, et l'analyse du transfert (Ringbom, 1994).

3.1.2.4. Bilan

En retraçant le développement des différents modes de recherche en acquisition d'une langue seconde, nous constatons que chaque nouveau type d'analyse élargit notre perspective et apporte sa propre contribution. Ainsi, comme l'indiquent Larsen-Freeman et H.Long (1990 : 73), il serait injuste de dire que chaque type d'analyse remplace son prédécesseur. Au contraire, on estime que chaque type d'analyse subsume ce qui est venu avant.

Dans notre travail, nous souhaitons exploiter les données orthographiques des apprenants chinois à l'aide du logiciel *Le Trameur*. Les analyses seront concentrées sur les variétés diachroniques et synchroniques produites par les apprenants de niveaux scolaires différents. A partir d'une analyse quantitative, nous mènerons une enquête sur les performances des apprenants en fonction des différentes catégories d'orthographe, à savoir l'orthographe lexicale et l'orthographe grammaticale. En outre, une analyse qualitative des données nous permet d'identifier les contraintes linguistiques et cognitives possibles qui influencent les apprenants chinois. En faisant appel à des méthodes d'analyse différentes, nous souhaitons mettre en lumière le processus de la construction de connaissances orthographiques chez les apprenants chinois du français L2.

3.2. Corpus

En décrivant le corpus écrit analysé dans ce travail, il convient en premier lieu de présenter les apprenants chinois du français L2 participant dans l'étude ainsi que la tâche d'écriture utilisée pour recueillir des textes en français. Ensuite, nous aborderons les opérations à effectuer pour constituer le corpus, telles que la transcription et l'annotation des textes, la création de base pour le logiciel *Le Trameur*. Nous noterons également dans cette partie les difficultés que nous avons rencontrées lors de la préparation de base à l'aide du Trameur.

3.2.1. Description de la population

Au total, 99 étudiants chinois ont participé à l'expérimentation : 45 d'entre eux sont au début de leur deuxième année d'étude du français, 24 sont à la 3e année, et 30 sont à la 4e année.

Ces étudiants chinois sont tous dans le département de langue et littérature françaises de l'Université des Langues étrangères N° 2 de Pékin. La plupart d'entre eux ont commencé à apprendre l'anglais dès la première année de leur scolarité. Rappelons que le recrutement des élèves en Chine dans un lycée se fait après neuf années de scolarité obligatoire. Les élèves ayant terminé leur cursus au lycée peuvent accéder à l'université après un concours d'entrée très sélectif organisé par le Conseil Suprême de l'Enseignement Supérieur. Après avoir passé le concours, les élèves arrivent en général à l'université à l'âge de 18 ou 19 ans sans connaître un mot français. Ils commencent donc à apprendre de façon intensive le français.

Notons que ces étudiants ayant passé un concours national d'entrée à l'université viennent de différentes régions chinoises. Jusque vers les années 1990, la grande partie de la Chine connait une situation diglossique : le chinois standard (*putonghua*), « joue le rôle de variété haute » ; et « la langue locale est variété basse » (Saillard & Boutet, 2008). Après une politique de véhicularisation de *putonghua* de près de cinquante ans, on constate actuellement en Chine un phénomène de « vernacularisation ». Saillard et

Boutet (2008) explique ce phénomène de façon suivante :

> « On constate cependant depuis peu en Chine (Saillard, 2002 ; 2004) des phénomènes d'appropriation du *putonghua* par les locuteurs d'autres variétés du chinois. Cela se traduit par des mutations tant fonctionnelles (utilisation du *putonghua* pour des fonctions relevant de la vie privée) que formelles (emprunts morphosyntaxiques et lexicaux aux autres langues chinoises qui donnent lieu à des variétés locales de chinois standard). La diffusion et la vernacularisation du putonghua, langue véhiculaire à l'échelle de la Chine, sont tributaires non seulement d'un facteur spontané, la mobilité des populations … mais aussi d'une planification gouvernementale visant à imposer le standard dans toutes les situations de la vie publique et, ce, notamment dans le domaine de l'éducation et des médias ».

Avec la promotion par intermédiaire des médias et de l'enseignement, le *putonghua* est considéré comme « la seule langue officielle de la scolarisation », ainsi que « le seul vecteur d'apprentissage des compétences écrites » (Saillard & Boutet, 2008). Par conséquent, on constate que les enfants chinois maîtrisent de plus en plus tôt le *putonghua*.

3.2.2. Réalisation d'une tâche d'écriture

3.2.2.1. Protocole expérimental

L'expérimentation dont nous allons présenter les premiers éléments a été mise en place dans une perspective didactique, qui a pour visée de caractériser les modes d'écriture usités par les apprenants chinois au milieu scolaire. L'organisation de cette expérimentation a donc obéi à deux contraintes :

L'expérimentation s'est située dans une perspective écologique. Autrement dit,

elle s'est déroulée dans le cadre ordinaire des classes des apprenants, et avec la collaboration de leurs enseignants. Ce protocole interdit ainsi le recours à des matériels sophistiqués tels que les dispositifs de suivi oculaire, afin d'apporter le moins de perturbation possible, et conduit à proposer des activités proches de celles auxquelles les apprenants sont habitués (Plane, 2011).

L'expérimentation avait pour objectif d'obtenir les informations sur les différentes manières dont les scripteurs perçoivent et analysent les inputs, en utilisant leurs connaissances linguistiques préalables afin de produire leur propre texte. Cette visée interdit donc que les opérations des scripteurs interviennent sur des substrats hétérogènes, ce qui aurait été le cas « si la consigne d'écriture avait lassé aux scripteurs la liberté d'inventer de toutes pièces le texte qu'ils allaient écrire » (Plane, 2011). De ce fait, nous avons proposé aux apprenants participant à l'expérimentation un même substrat initial comme point de départ, à savoir un conte, dont nous avons décidé les caractéristiques.

3.2.2.2. Texte de source

Le texte source utilisé pour cette expérimentation est issu de la légende, *Les chevaliers de la table ronde* (2010). L'extrait que nous avons choisi, *Perceval et le chevalier Vermeil*, a l'intérêt d'offrir aux apprenants chinois un univers exotique, qui ne fait pas partie de leur répertoire culture classique : Perceval le Gallois, le type du chevalier « aventureux », autour du roi Arthur et de la Table Ronde, est toujours prêt à se battre pour défendre les dames en difficulté. A travers tous les événements que ce conte narre, les apprenants chinois pourraient se plonger dans un monde étrange et remonter à des temps anciens. En revanche, il s'agit d'une version adaptée pour les jeunes locuteurs d'aujourd'hui, où la plupart des difficultés lexicales et syntaxiques sont déjà lissées.

Ce texte montre une grande diversité au niveau de temps verbaux. Dans la narration, les temps dominants sont le passé simple et l'imparfait, alors que dans le discours direct des personnages, nous trouvons le présent, le passé composé, ainsi que le future simple. Outre le mode de l'indicatif, on trouve des formes verbales au

subjonctif et au conditionnel. Il faut indiquer que nous avons élaboré deux versions de texte : l'une, destinée au groupe de débutants, dans laquelle nous avons remplacé systématiquement le passé simple utilisé dans la narration par le passé composé, en tenant compte que ce groupe des apprenants, ayant une année d'études de français, n'a pas encore appris le passé simple ; et l'autre, destinée à deux autres groupes de niveau scolaire plus haut, dans laquelle tous les temps verbaux sont gardés.

En outre, le texte est riche de propos des personnages, qui sont tous rapportés au discours direct, ce qui devrait nous permettre d'observer comment les apprenants font parler les personnages dans leur propre texte, qui est une tâche rédactionnelle à la fois cruciale et problématique.

Le dernier élément à noter concerne la variété lexicale. Dans le texte existent les mots se référant aux objets du moyen âge, tels qu'*épée* et *armure*, qui pourraient constituer un problème pour les apprenants chinois. Cependant, il nous semble intéressant de voir comment les apprenants récupèrent le sens de ces mots inconnus à partir du contexte, comment ils gèrent ces difficultés dans leur production de texte, qui pourraient transformer probablement les mots anciens aux objets quotidiens.

3.2.2.3. Déroulement

Le dispositif comporte deux phases :

1. Lecture par les étudiants du texte du conte intitulé *Perceval et le chevalier Vermeil*
   Le texte a été diffusé aux étudiants, avec une feuille blanche, dans laquelle les étudiants pourraient prendre les notes pendant la lecture.
   Durée de la lecture : 20 minutes
2. Restitution par écrit du texte lu, immédiatement après la lecture
   Les étudiants avaient reçu la consigne de lire attentivement le conte, de noter tous les éléments qui leur sont parus importants dans la feuille blanche, et de s'efforcer de mémoriser le conte car il leur serait demandé de le mettre par écrit sitôt après la lecture.

Durée de l'écriture : 30 minutes

3.2.2.4. Consignes à donner

Les consignes suivantes sont données en français aux apprenants chinois lors de la tâche d'écriture :

Vous disposez de 20 minutes à lire un texte. Il s'agit d'une histoire qui se passe aux temps anciens. Puis, vous aurez à produire un texte en vous appuyant sur cette histoire. Vous pouvez prendre les notes sur cette feuille blanche pendant votre lecture. On va reprendre le texte au bout de 20 minutes, mais vous pouvez garder vos notes. Puis, vous disposez de 30 minutes à produire votre propre texte.

3.2.2.5. Recueil de copies

Après la tâche d'écriture, nous avons collecté au total 92 textes, 40 textes venant des étudiants de la 2e année de licence (23 du premier groupe, 17 du deuxième groupe), 24 textes de la 3e année, et 28 textes de la 4e année. En raison du non-aboutissement de la tâche d'écriture, les textes produits par sept étudiants ne sont pas compris dans notre corpus final[21].

3.2.3. Constitution du corpus

3.2.3.1. Procédures de transcription et d'annotation

Ces textes écrits sont scannés et dénommés pour faciliter la suite de la recherche. Le nom doit indiquer :

[21] Parmi ces sept étudiants, 4 étudiants n'ont pas rendu leur texte produit après 30 minutes d'écriture ; un étudiant n'a produit qu'une phrase de dizaine de mots, sans commettre les erreurs orthographiques ; deux étudiants ont produit un texte en expliquant pourquoi ils ne voulaient pas faire cette tâche d'écriture.

- Le niveau de groupe : L2/L3/L4
- L'identification du groupe : 01/02
- Le numéro de l'étudiant dans le groupe : 001, 002…

Ce système de dénomination nous permet de spécifier l'origine de chaque écrit. Et les données ainsi numérisées constituent la première partie de notre corpus (V1).

Les données primaires sont par la suite transcrites via un éditeur de texte brut (Sublime Texte pour Mac) sous une forme au plus près de la copie scannée. Comme nous voudrions établir des transcriptions qui puissent être exploitées ensuite par *le Trameur*, en privilégiant la linéarisation du texte, certaines opérations faites par les étudiants, telles que suppression, ajout, remplacement, nécessitent un codage particulier. Ici, nous prenons la convention de transcription établie par le projet d'*Ecriscol*. Et les données ainsi informatisées constituent la deuxième partie de notre corpus (V2).

Tableau 4: Convention de transcription

Élément à transcrire	**Proposition de codage**
Suppression	[caractères supprimés]
Ajout	~~caractères ajoutés~~
Remplacement de texte	+R[met]+R+//se_mis//+R
Segment illisible	#xxx#
Retour à la ligne	§

Les données primaires donnent lieu également à l'établissement de la troisième partie de notre corpus (V3) : les annotations. Il s'agit ici d'annoter les éléments qui comportent les erreurs orthographiques. Les données primaires sont ainsi converties en données linguistiquement « normées », qui nous permettent à construire des bases textométriques pour *le Trameur*.

Pour annoter un mot ou une suite de mots dont on souhaite rectifier l'orthographe, on procède de la manière suivante :

- Isoler le segment à annoter entre deux signes { }
- Placer ensuite le tiret bas _
- Écrire le segment normé, entre les deux signes { }

La proposition de codage pour toutes les ponctuations dans les annotations (virgule, point, apostrophe etc.) :

- {s'a}_{s'est} : {s_QUOTE_a}_{s_QUOTE_est}

3.2.3.2. Préparation de bases pour *le Trameur*

Avant d'importer les données secondaires dans *le Trameur*, il faut tout d'abord concaténer les 92 fichiers en un seul. Le fichier résultant est ensuite soumis à différentes procédures de nettoyage, formatage etc. Il nous faudrait également insérer la balise de partie sous la forme suivante :

<partie="L2">
<p="L2-01-001-B">

La base ainsi obtenue a un format compatible avec celui d'une base textométrique importable dans *le Trameur*. Comme la figure 1 l'illustre, chaque item dans le corpus est composé de 4 couches d'annotation pour le moment :

- Annotation n°1 : forme initiale. Dans notre corpus, la première couche d'annotation concerne la forme produite par les apprenants.
- Annotation n°2 : soit le lemme construit via *Treetagger*[22], soit l'opération de

[22] Le programme Treetagger : « système d'étiquetage automatique des catégories grammaticales des mots avec lemmatisation. Il permet aussi de générer et de gérer des annotations multiples sur les unités du texte (et de traiter les niveaux d'annotations visés) » (http://www.tal.univ-paris3.fr/trameur/).

transformation réalisée sur la forme initiale. Par exemple, si l'apprenant a mis dans le texte *chevaliers*, ce sera donc le lemme *chevalier* mis dans la deuxième couche d'annotation.

- Annotation n°3 : catégorie grammaticale construite par *Treetagger*. Si nous reprenons l'exemple *chevaliers*, il s'agit donc de la catégorie du nom dans la troisième couche d'annotation.

- Annotation n°4 : forme annotée. Si la forme initiale produite par l'apprenant relève d'une erreur orthographique, cette forme erronée sera corrigée. Ainsi, ce sera la forme corrigée qui se trouve dans la quatrième couche d'annotation. En revanche, si la forme produite par l'apprenant est correcte, la forme de la première couche coïncide avec celle de la quatrième couche.

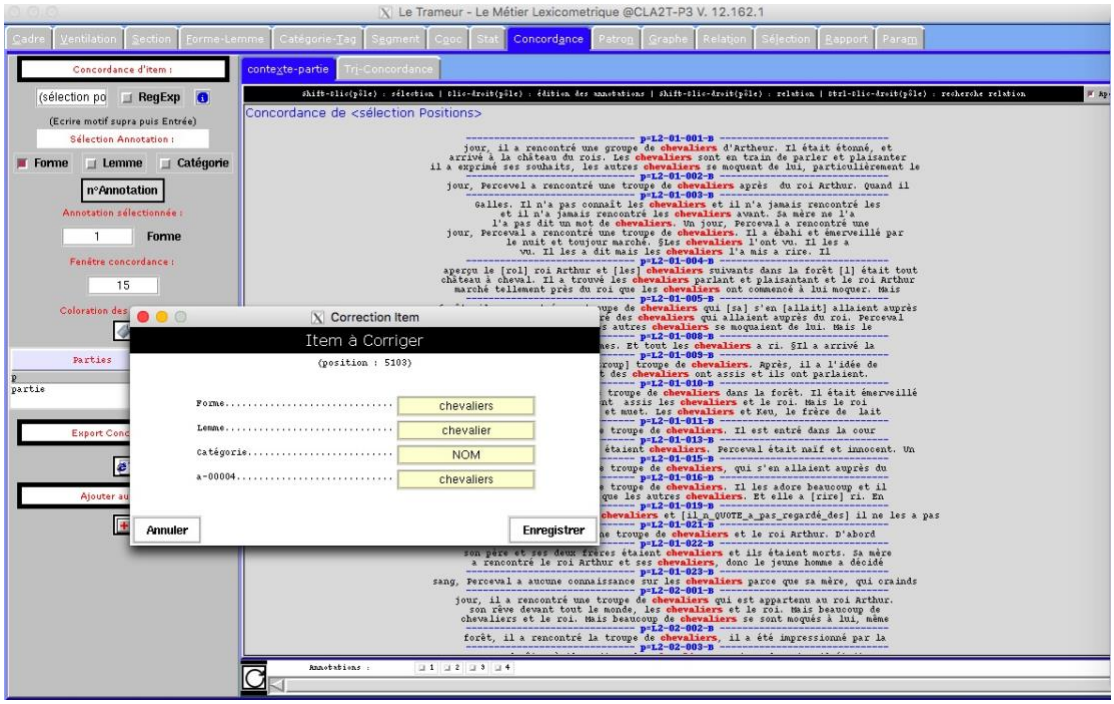


Fig. 1 : Quatre couches d'annotation pour l'item

Notons que la troisième couche d'annotation de notre base concerne les catégories grammaticales de mots, qui joue un rôle important dans l'analyse suivante. Cependant, pour un certain nombre de mots, leur catégorie ne peut pas être reconnue automatiquement par *le Trameur* pour les raisons suivantes :

- Majuscule : *le Trameur* groupe tous les mots commençant par une lettre majuscule dans la catégorie de nom propre. Cependant, la plupart des cas concerne la lettre initiale d'une phrase.

- Élision : en français, on observe la suppression de la voyelle finale d'un mot devant un mot commençant par une voyelle ou un *h* muet. Cette élision se marque par l'apostrophe. En raison de la suppression de la voyelle finale et l'ajout de l'apostrophe, *le Trameur* ne peut plus reconnaître la catégorie de mot. Par exemple :

(1) *On dit que la fille n'allait pas rire…*

Avec la suppression de la voyelle finale *e* et l'ajout de l'apostrophe, *le Trameur* ne peut plus reconnaître la catégorie du mot *ne*.

- Auxiliaire : pour le classement des catégories, nous observons une tendance à confondre l'auxiliaire qui sert à former les temps composés et le verbe principal de la phrase, surtout quand les deux éléments sont séparés. Par exemple :

(2) **Keu n'a pas voulu accepté le fait.*

Au lieu d'être classé comme auxiliaire, *a* est considéré comme verbe à l'indicatif présent, en raison de l'insertion de l'adverbe *pas*.

- Autres : si la catégorie d'un certain nombre de mots est mal reconnue, c'est parce que la phrase comporte une structure complexe. Par exemple :

(3) * *Mais Perceval sauvé les autres avec les armés que le roi lui a donné.*

Dans cette phrase, le mot *que* introduit une relative subordonnée, et il s'agit donc d'un pronom relatif. Cependant, ce mot *que* est classé dans la catégorie de conjonction.

Le nettoyage de ces catégories mal reconnues est nécessaire quand on prépare les bases pour une analyse textuelle. Pour certaines erreurs qui se présentent d'une manière systématique, nous pouvons la corriger automatiquement dans *le Trameur*. Par exemple : le mot *chevalier* est systématiquement classé dans la catégorie d'adjectif. Avec *le Trameur*, on peut afficher une liste de toutes les occurrences du mot *chevalier*, et la correction de cette erreur sera appliquée automatiquement à toutes les occurrences.

Cependant, pour le problème posé par, par exemple, la majuscule de la lettre initiale de phrase, on est obligé d'intervenir à chaque occurrence. Au lieu de mettre la

lettre initiale en majuscule pour marquer le commencement d'une phrase, il nous semble pertinent de proposer un autre codage dans la suite de recherche, afin d'éviter un travail de nettoyage qui est assez coûteux en temps.

Le travail de nettoyage concerne également les mots qui pourraient appartenir à deux catégories grammaticales, notamment en ce qui concerne les mots outils, tels que *le, la, les*. Se situant avant le nom, ces mots monosyllabiques sont déterminants, tandis qu'ils sont classés dans la catégorie du pronom, en collant au verbe. Ainsi, chaque occurrence demande une analyse particulière du contexte syntaxique. Et si la catégorie grammaticale est mal reconnue, une intervention manuelle est obligatoire. Par exemple, dans la phrase *il s'est décidé de* ***les*** *joindre et d'être un chevalier*, le mot *les* doit être classé dans la catégorie de pronom :

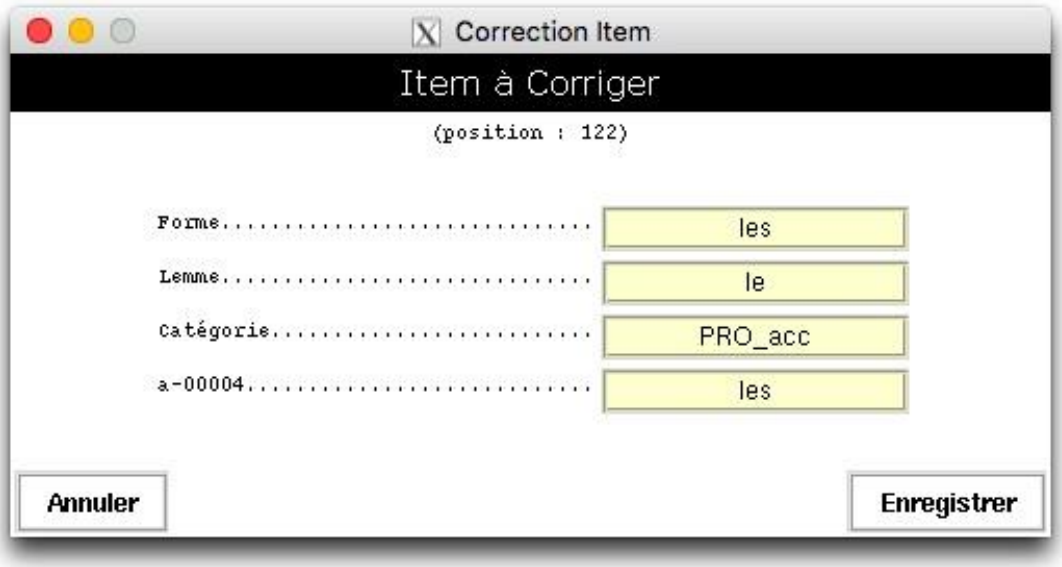


Fig. 2 : Correction des annotations.

3.3. Analyse des données

Dans la démarche d'analyse, nous lisons chacune des copies successivement et soulignons les mots qui comportent une erreur. Les formes erronées sont ensuite caractérisées. Pour dessiner le profil orthographique des apprenants de notre corpus, il nous paraît nécessaire de comprendre la nature des erreurs qu'ils ont commises et de

porter un regard aussi objectif que possible sur eux. Autrement dit, il faut donner un critère descriptif qui qualifie la nature de l'erreur. Nous établissons donc notre typologie des erreurs, en référence à la recherche de Manesse et *al.* (2007), qui s'appuie elle-même sur la typologie établie par Catach (2003). Le classement des erreurs de Manesse et *al.* est le suivant :

Tableau 5. Typologie des erreurs d'orthographe (Manesse & Cogis, 2007 : 88)

Types d'erreur	Définition	Exemples d'erreur
1	Mauvais découpage du mot ; mot sauté ou tronqué	Sans fonce
2	Aberration dans la représentation des sons	S'onvonce
3	Substitution de mot	Sont font
4	Cumul de fautes grammaticale et lexicale	S'enffons
5	Faute d'orthographe grammaticale ; la catégorie grammaticale n'est pas représentée ou est mal représentée	S'enfonce S'enfonces
6	Orthographe lexicale	S'enfonssent
7	Orthographe lexicale : forme approchante	S'enfonsent S'anfonsent
8	Signes orthographiques et majuscules	S'enfonçent

Notre corpus et celui de Manesse et *al.* présentent certaines similitudes : il s'agit des corpus à peu près homogènes. En d'autres termes, les deux corpus concernent des ensembles de textes qui ont répondu à la même consigne et dont les scripteurs sont proches d'un point de vue scolaire (Masseron & Luste-Chaa, 2008). Ce type de corpus présente donc un vocabulaire prédictible, y compris du point de vue de l'environnement morphosyntaxique. Néanmoins, nous apportons certaines modifications au classement d'erreurs conçu par Manessse et al. (2007), en nous basant sur les caractéristiques de notre corpus :

- Les textes recueillis dans notre corpus sont issus de la même tâche d'écriture semi-spontanée : après une lecture d'un conte intitulé « Perceval et le chevalier Vermeil » (20 minutes), les étudiants ont été invités à reconstituer par écrit le texte lu (30 minutes). A la différence du corpus de Manesse et *al.*, dont les textes ont été collectés après une dictée, les erreurs de type « mots sautés »[23] ne sont pas trouvées dans notre corpus.
- Selon le classement ci-dessus, les erreurs de type 2 regroupent les graphies dont la représentation phonétique est aberrante. Selon Manesse et al. (2007 : 73), ce type d'erreur concerne le mot qui est visiblement inconnu de l'élève et l'élève tente de reconstituer les sons de ce mot qu'il semble percevoir pour la première fois. Les erreurs de ce type sont également observées dans notre corpus. Cependant, étant donné la nature de la tâche d'écriture adopté par notre étude, ces graphies erronées ne pourraient pas être interprété par la méconnaissance de mots.
- Rappelons que l'apprenant chinois apprend simultanément le français oral et écrit, tandis que l'enfant français aborde l'écrit avec une maîtrise précoce du français oral. Ainsi, les erreurs orthographiques altérant la valeur phonétique de mots sont inévitables pour les apprenants chinois qui n'ont pas encore maîtrisé le système phonologique du français. Par conséquent, nous souhaitons apporter un affinement de la description des erreurs orthographiques en mettant l'accent sur cette spécificité des apprenants en français L2.
- Les erreurs de grammaire relèvent de deux domaines : les accords (en nombre et en genre) et la conjugaison du verbe. Le premier présente une grande homogénéité, car il s'agit de l'application de règles régulières, enseignées dès le début de l'enseignement du français. Les erreurs relevant de la conjugaison verbale sont, en revanche, plus hétérogènes et notre corpus présente une diversité importante de temps et de modes. Ainsi nous traiterons séparément les deux domaines de l'orthographe grammaticale.

[23] Il s'agit d'omission des mots dictés, qui touche généralement l'identification de ces mots dans la chaîne parlée.

- Selon la typologie ci-dessus, le dernier type d'erreur concerne les signes orthographiques (l'apostrophe, le trait d'union) et majuscules. Ces erreurs ne seront pas traitées dans notre corpus et les erreurs de l'accent sont classées dans la catégorie de l'orthographe lexicale.

Nous retenons au total quatre catégories, sept types d'erreurs. Nous proposerons un tableau récapitulatif dans la partie suivante.

3.3.1. Classement d'erreurs

Tableau 6. Classement d'erreurs d'orthographe

Types d'erreur	Définitions
1	Mauvais découpage du mot
2	Substitution de mot
3a	Erreur orthographique à dominante phonétique
3b	Erreur phonogrammique
3c	Erreur d'orthographe lexicale sans altération de la valeur phonique
4a	Erreur d'orthographe grammaticale : la marque du nombre ou du genre n'est pas présentée ou est mal présentée
4b	Erreurs d'orthographe grammaticale : les désinences verbales

Chaque forme erronée doit être rattachée à l'un des types d'erreur. Cependant, il arrive qu'une forme erronée comporte deux erreurs en même temps. Par exemple, la forme erronée * *les consil* (pour *les conseils*) comporte une erreur d'orthographe lexicale ainsi qu'une erreur d'accord, et cette forme erronée sera donc recensée deux fois dans le relevé.

- Mauvais découpage du mot, qui touche généralement à la reconnaissance du mot comme entité dans la chaîne parlée. Les étudiants ne se rendent pas compte de l'existence indépendante d'une unité lexicale, soit qu'ils ne sachent pas la

séparer par des blancs de ses voisins de gauche ou de droite, soit qu'ils la tronçonnent en deux ou trois unités (Manesse & Cogis, 2007 : 73). Les erreurs de ce type relèvent surtout, selon Ros Dupont (2000), d'une difficulté de délimitation, et elle est « le fait des apprentis qui découvrent l'écrit ». Cela explique les formes erronées observées chez les apprenants chinois, qui commencent juste de découvrir l'écrit de la langue française. Le plus souvent, les erreurs surviennent au moment où les étudiants sont confrontés à une production d'écrit réelle : « la transcription d'un énoncé totalement pensé et proposé par lui, dont les éléments constitutifs ne leur sont pas forcément tous connus » (Ros Dupont 2000). Par exemple, L'erreur suivante relève de l'hypersegmentation : le verbe *embrassa* a été coupé en deux segments.

(4) * *Sa mère {l_QUOTE_en_brassa}_{l_QUOTE_embrassa} et lui dit que {le}_{la} mort honorable*. (L3-004)

- Substitution de mot, qui concerne la forme qui présente, correctement orthographiée, un autre mot que le mot attendu. Notons que ce n'est pas à proprement parler une erreur d'orthographe (Manesse & Cogis, 2007:74). S'agit-il toujours d'une mauvaise compréhension du sens du mot, ou une faute d'inattention assimilable à un lapsus. D'une manière générale, il est impossible de le dire. Par exemple, dans notre corpus, il existe des erreurs concernant les homophones *cours* et *cour*.

(5) * *Finalement Perceval a quitté {le}_{la} {cours}_{cour} de roi.* (L2-02-013)

Les deux mots ont la même prononciation mais les sens différents. Au niveau morphosyntaxique, ils sont tous des noms, mais ont les genres différents. L'orthographe de ce couple homophone lexical peut être déterminée par l'utilisation de mots de la même famille : la consonne finale *s* de *cours* peut être

déduite des mots comme *course*, *coursier*, *courser*, etc. Cependant, le stockage lexical des scripteurs de notre corpus vient d'être établi, soit ils ne connaissent pas encore les mots dérivés, soit ils ne sont pas capables de faire un lien entre eux. Dans ces cas, seul le recours à la mémoire leur permet de choisir la forme adéquate, mais chez le scripteur ci-dessus, cette mémoire est déficitaire.

Dans la catégorie des erreurs d'orthographe lexicale, il existe trois types d'erreur :

- Erreur orthographique à dominante phonétique, qui concerne les formes graphiques dont la représentation phonétique est aberrante. La mauvaise transcription des sons par les scripteurs nous amène à considérer que ces formes erronées sont bien loin de la pure et simple faute lexicale. Selon Cogis (2005 : 17-18), les erreurs de ce type, où la valeur phonique est dégradée, correspondent « à une erreur dans le langage "oral" » et relève d'« un travail sur l'audition et l'articulation ». Autrement dit, elles sont dues à une mauvaise production orale. Et pour y remédier, il faut assurer l'oral, afin d'asseoir, chez les scripteurs, une connaissance précise des différents phonèmes. Par exemple, la forme erronée ci-dessous cumule deux erreurs. Pour la première, il s'agit de la confusion de voyelles orales [a] et [u] ; quant à la deuxième, il s'agit de la confusion entre [u] et [ø].

(6) * *Le frère d'Arthur est {jouleux}_{jaloux}. (L2-01- 019)*

- Erreur d'orthographe lexicale altérant la valeur phonique, qui relève du domaine de l'orthographe lexicale mais le choix des graphèmes altère la valeur phonique des mots concernés. Le fait que les scripteurs n'ont pas été en mesure de restituer correctement l'orthographe des mots s'explique par une maîtrise imparfaite du système graphique. Plus précisément, les erreurs sont issues de la méconnaissance ou de la mémorisation insuffisante des phonogrammes, des lois de position, etc. (Katoozian, 2013). Par exemple, la forme erronée ci-dessous concerne l'omission du *e* caduc où il possède une valeur auxiliaire, introduisant ainsi une modification de la valeur phonique. Plus précisément,

son absence ôte la prononciation aux consonnes qui le précèdent.

(7) * *il a rencontré une {troup}_{troupe} de chevaliers*. (L2-02-010)

- Erreur d'orthographe lexicale sans altération de la valeur phonique. Cette catégorie ressemble également les erreurs relevant de l'orthographe lexicale mais il s'agit des erreurs qui ne modifient pas la prononciation des mots. Autrement dit, la forme produite par le scripteur est une des variantes possibles graphiquement mais non retenue par la norme orthographique (R.Honvault, 1995 : 61). Par exemple, l'erreur suivante concerne l'adjonction d'une lettre finale, qui n'est pas prononcée et n'influe non plus la prononciation du phonogramme voisin.

(8) * *Elle a eu un {désire}_{désir} que Perceval peut parler avec le roi. (L2-01-017)*

Le type suivant regroupe les erreurs d'orthographe grammaticale, où la marque propre à la catégorie grammaticale n'est pas présentée ou est mal présentée. Les erreurs relèvent de deux domaines : le problème de l'accord, par exemple, **le matins* (pour *le matin*), et celui de la conjugaison verbale, par exemple : **la fille a rit* (pour *la fille a ri*).

3.3.2. Analyse quantitative à travers *le Trameur*

Notre travail a pour objectif d'étudier les comportements orthographiques des apprenants chinois et d'avoir une première vue de leurs lacunes et de leurs points forts lors de leur apprentissage. Pour ce faire, nous relevons les erreurs, les hésitations, les options manifestées chez les apprenants dans les copies recueillies. Dans la partie précédente, nous avons exposé les principes de classement des erreurs en huit types. Néanmoins, si nous souhaitons appréhender le système des erreurs, il faut observer la

répartition des erreurs en différents types et le poids relatif de chaque type. Ainsi, nous nous donnons des éléments « pour comprendre quels sont les secteurs faibles et les secteurs mieux maitrisés de l'orthographe » (Manesse & Cogis, 2007 : 88). Afin de faire cette analyse quantitative, les textes écrits par les apprenants ont été soumis à différentes analyses automatiques proposées par Le Trameur ainsi qu'à certain nombre de calculs manuels. Dans cette partie, nous exposerons les analyses automatiques que l'on pout effectuer à l'aide du Trameur.

3.3.2.1. Extraire les éléments qui nous intéressent

Tout d'abord, *Le Trameur* nous permet d'extraire les formes qui nous intéressent d'une manière automatique. Si l'on prend l'exemple de l'accord du nombre, nous pouvons trouver tous les noms dont le contexte demande un accord du pluriel. Pour ce faire, il faut faire une recherche dans le secteur de sélection. Notons que ce secteur nous permet d'exploiter les données dans toutes quatre couches d'annotation. Afin de trouver toutes les occurrences de noms au pluriel, les consignes suivantes sont nécessaires :

- .*[^s]$ dans le crochet de « lemme », qui nous permet de trouver tous les lemmes qui ne se terminent pas par un *–s*.
- NOM dans le crochet de « catégorie », qui indique que la recherche sera faite dans la catégorie de nom.
- .*s$ dans le crochet de « a-00004 »[24], qui sert à trouver toutes les formes se terminant par un *–s*.

[24] Dans notre corpus, la quatrième couche d'annotation concerne les formes finales ou corrigées. Plus précisément, si la forme produite par l'apprenant, qui se trouve dans la première couche d'annotation, est correcte, la forme de la première couche coïncide avec celle de la quatrième. En revanche, si la forme de la première couche est erronée, la quatrième présente donc la forme corrigée.

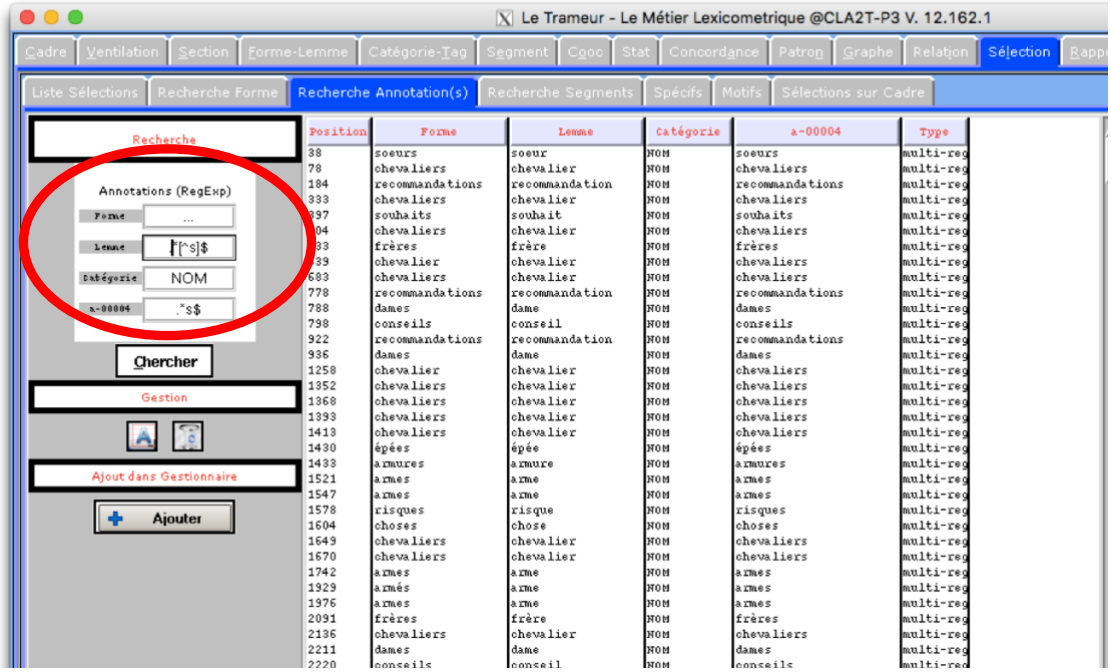


Fig. 3 : Lancement de l'enquête sur les noms au pluriel

Avec une telle opération, *le Trameur* nous présente une liste de mots, soit tous les noms qui demandent un accord du pluriel[25]. Ensuite, en ajoutant ces occurrences dans la « liste sélections », nous calculons leur nombre total.

[25] Dans ce cas nécessite une vérification, puisqu'il est tout à fait possible que dans notre corpus existent des mots se terminant par un –s et se trouvant dans un contexte du pluriel.

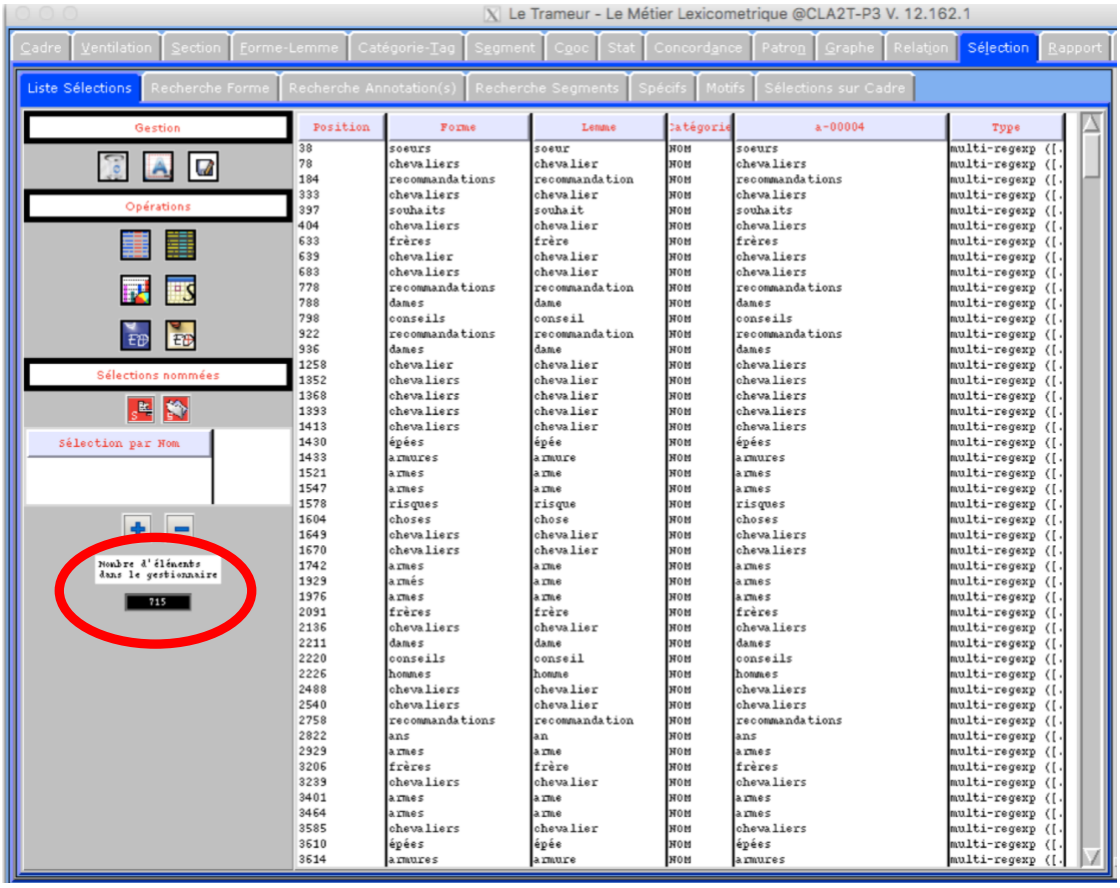
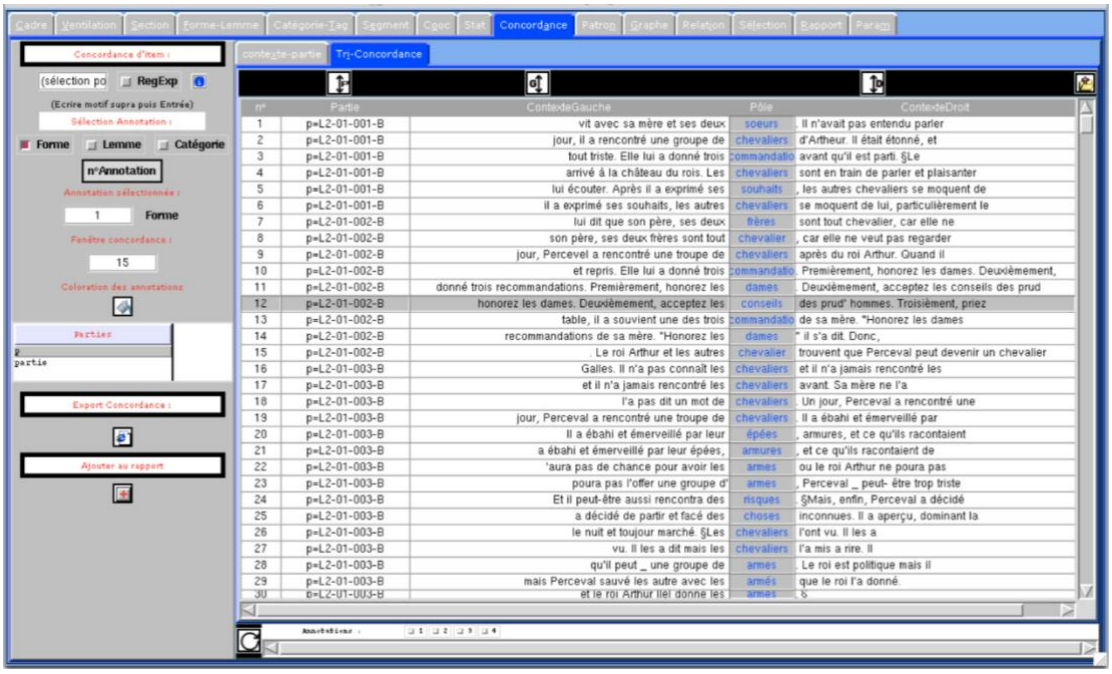


Fig. 4 : Occurrences de noms au pluriel

Comme le visualisent les deux figures suivantes (Fig. 5 et Fig. 6), nous pouvons également examiner le contexte de ces formes nominales ainsi que leur distribution en fonction de chaque apprenant ou chaque groupe d'apprenants[26].

[26] La graphie affichée dans la deuxième image permet de visualiser le nombre de formes nominales au pluriel produit par chaque apprenant.



n°	Partie	ContexteGauche	Pôle	ContexteDroit
1	p=L2-01-001-B	vit avec sa mère et ses deux	soeurs	. Il n'avait pas entendu parler
2	p=L2-01-001-B	jour, il a rencontré une groupe de	chevaliers	d'Artheur. Il était étonné, et
3	p=L2-01-001-B	tout triste. Elle lui a donné trois	commandato	avant qu'il est parti. §Le
4	p=L2-01-001-B	arrivé à la château du rois. Les	chevaliers	sont en train de parler et plaisanter
5	p=L2-01-001-B	lui écouter. Après il a exprimé ses	souhaits	, les autres chevaliers se moquent de
6	p=L2-01-001-B	il a exprimé ses souhaits, les autres	chevaliers	se moquent de lui, particulièrement le
7	p=L2-01-002-B	lui dit que son père, ses deux	frères	sont tout chevalier, car elle ne
8	p=L2-01-002-B	son père, ses deux frères sont tout	chevalier	, car elle ne veut pas regarder
9	p=L2-01-002-B	jour, Percevel a rencontré une troupe de	chevaliers	après du roi Arthur. Quand il
10	p=L2-01-002-B	et repris. Elle lui a donné trois	commandato	. Premièrement, honorez les dames. Deuxièmement,
11	p=L2-01-002-B	donné trois recommandations. Premièrement, honorez les	dames	. Deuxièmement, acceptez les conseils des prud
12	p=L2-01-002-B	honorez les dames. Deuxièmement, acceptez les	conseils	des prud' hommes. Troisièmement, priez
13	p=L2-01-002-B	table, il a souvient une des trois	commandato	de sa mère. "Honorez les dames
14	p=L2-01-002-B	recommandations de sa mère. "Honorez les	dames	" il s'a dit. Donc,
15	p=L2-01-002-B	. Le roi Arthur et les autres	chevalier	trouvent que Perceval peut devenir un chevalier
16	p=L2-01-003-B	Galles. Il n'a pas connaît les	chevaliers	et il n'a jamais rencontré les
17	p=L2-01-003-B	et il n'a jamais rencontré les	chevaliers	avant. Sa mère ne l'a
18	p=L2-01-003-B	l'a pas dit un mot de	chevaliers	. Un jour, Perceval a rencontré une
19	p=L2-01-003-B	jour, Perceval a rencontré une troupe de	chevaliers	. Il a ébahi et émerveillé par
20	p=L2-01-003-B	Il a ébahi et émerveillé par leur	épées	, armures, et ce qu'ils racontaient
21	p=L2-01-003-B	a ébahi et émerveillé par leur épées,	armures	, et ce qu'ils racontaient de
22	p=L2-01-003-B	'aura pas de chance pour avoir les	armes	ou le roi Arthur ne poura pas
23	p=L2-01-003-B	poura pas l'offer une groupe d'	armes	, Perceval _ peut- être trop triste
24	p=L2-01-003-B	Et il peut-être aussi rencontra des	risques	. §Mais, enfin, Perceval a décidé
25	p=L2-01-003-B	a décidé de partir et facé des	choses	inconnues. Il a aperçu, dominant la
26	p=L2-01-003-B	le nuit et toujour marché. §Les	chevaliers	l'ont vu. Il les a
27	p=L2-01-003-B	vu. Il les a dit mais les	chevaliers	l'a mis a rire. Il
28	p=L2-01-003-B	qu'il peut _ une groupe de	armes	. Le roi est politique mais il
29	p=L2-01-003-B	mais Perceval sauvé les autre avec les	armés	que le roi l'a donné.
30	p=L2-01-003-B	et le roi Arthur lleï donne les	armes	s

Fig. 5 : Contextes gauche et droit de noms au pluriel

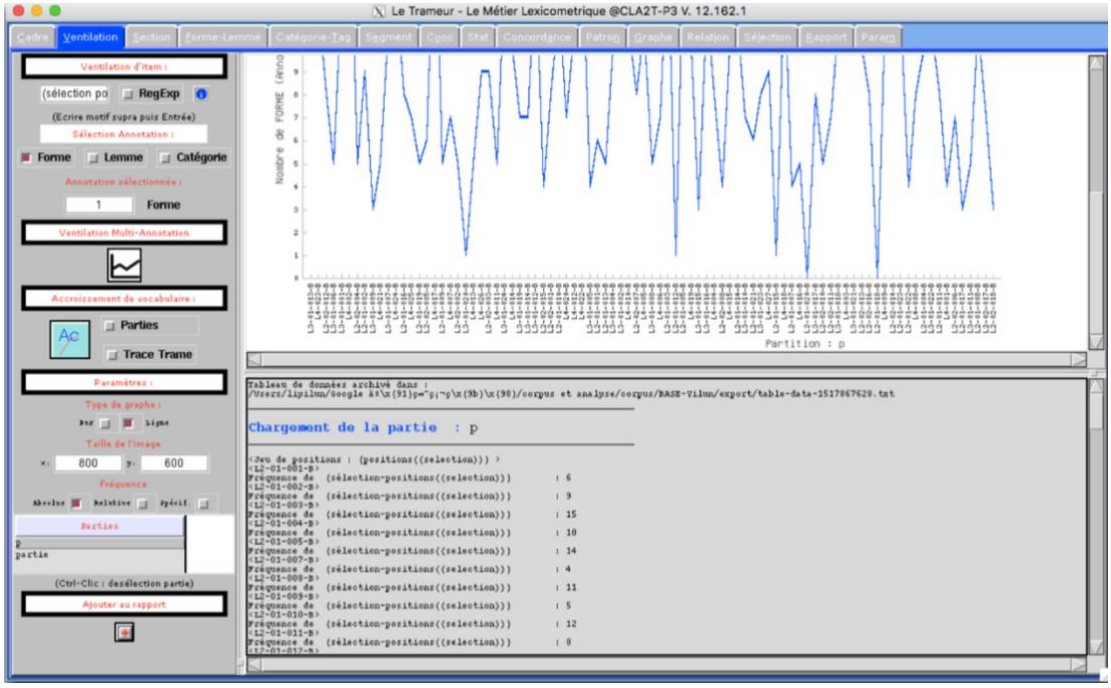


Fig. 6 : Distribution de noms au pluriel en fonction de niveau linguistique

3.3.2.2. Deux manières de repérer les formes non-normatives

Le Trameur nous permet également de repérer les formes non-normatives de manière

automatique. Si nous reprenons l'exemple de l'accord du nombre, nous pouvons trouver toutes les omissions de la marque du pluriel –s, en faisant appel à une recherche dans le secteur « sélection ». Les consignes seront suivantes :

- .*[^s]$ dans le crochet de « forme », qui nous permet de trouver tous les formes nominales produites par les apprenants qui ne se terminent pas par un *–s*.
- NOM dans le crochet de « catégorie », qui indique que la recherche sera faite dans la catégorie de nom.
- .*s$ dans le crochet de « a-00004 », qui sert à trouver toutes les formes se terminant par un *–s*.

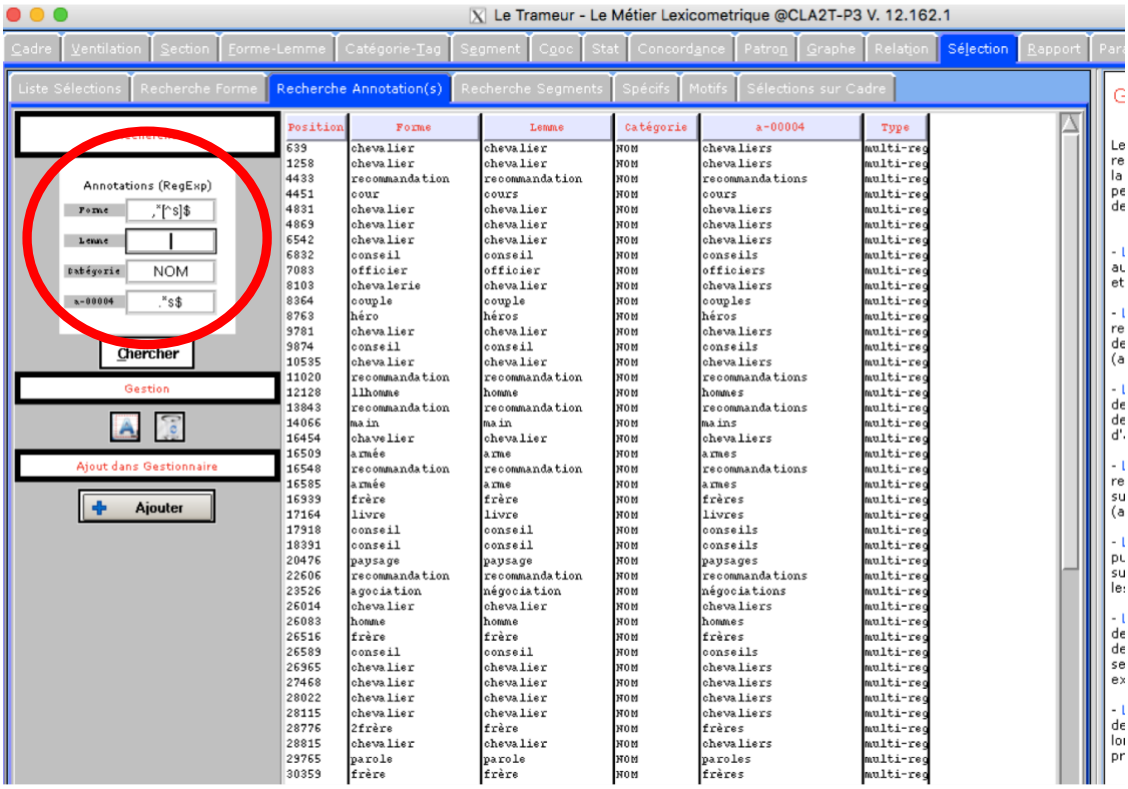


Fig. 7 : Lancement de la recherche de formes non-normatives

Une telle opération nous donnera une liste des occurrences dont la forme de la première couche d'annotation ne coïncide pas avec celle de la quatrième. De plus, la différence porte sur l'absence du morphème *–s*. Autrement dit, il s'agit des erreurs relevant de l'omission de la marque du pluriel *–s*. Remarquons qu'un travail de

nettoyage manuel sera nécessaire, puisqu'il existe des noms au singulier se terminant par un *–s* dans la liste de formes sélectionnées :

(9) a. * *ses deux frères sont {chevalier}_{chevaliers}. (L2-01-002)*

b. * *pour la première {foi}_{fois}* (L4-021)

En croisant le nombre de noms au pluriel et celui d'omission, nous pouvons déduire le taux de réussite de la morphologie nominale du nombre avec un simple calcul. Dans le secteur de « ventilation », nous pouvons également examiner la distribution des erreurs de l'omission en fonction de chaque apprenant ou chaque groupe d'apprenant. Parallèlement, une telle opération pourrait être effectuée pour d'autres catégories grammaticales, telles que déterminant, adjectif.

Le Trameur nous offre une autre possibilité de trouver les formes non-normatives sans préciser les caractéristiques de formes erronées. Nous nous concentrons cette fois sur la catégorie verbale, et souhaitons extraire toutes les formes verbales erronées produites par les apprenants dans ce corpus. Tout d'abord, comme l'illustre la figure 8, on crée une nouvelle couche d'annotation concaténant la forme produite par l'apprenant et la forme normative sous la forme X_Y. Autrement dit, la forme de la première couche d'annotation et celle de la quatrième seront concaténées dans une nouvelle couche d'annotation, la cinquième.

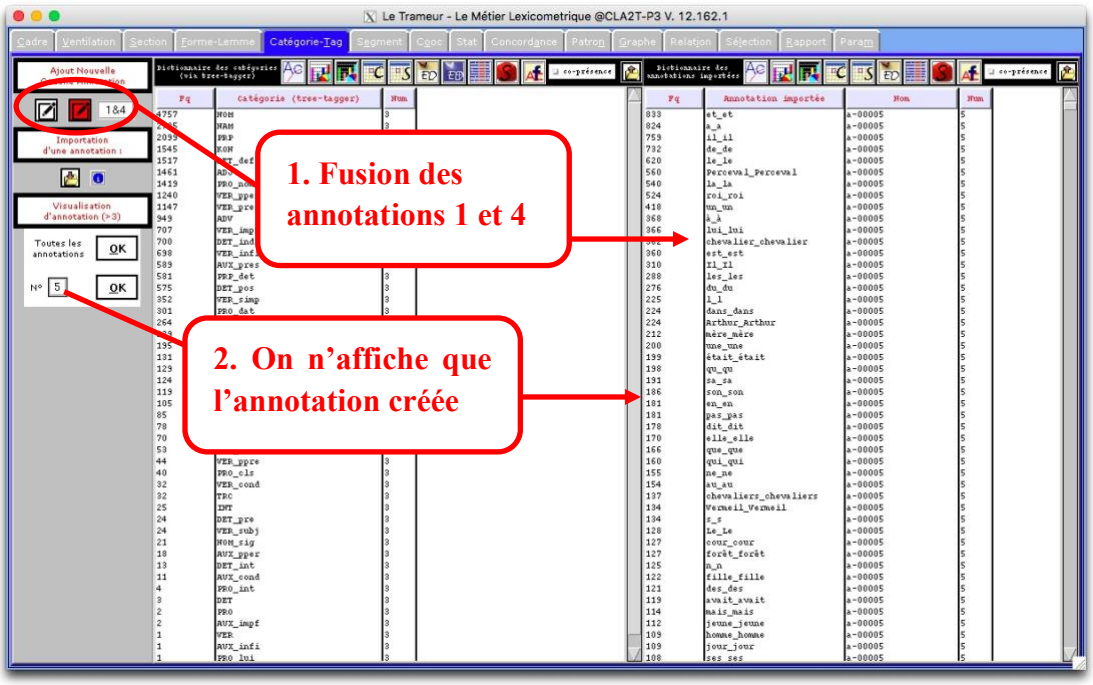


Fig. 8 : Création de la cinquième couche d'annotation

Deuxièmement, on cherche les formes verbales correctes : celles pour lesquelles X=Y dans la cinquième couche d'annotation. Et pour faire cette recherche, la consigne sera : (\w+)_\1.

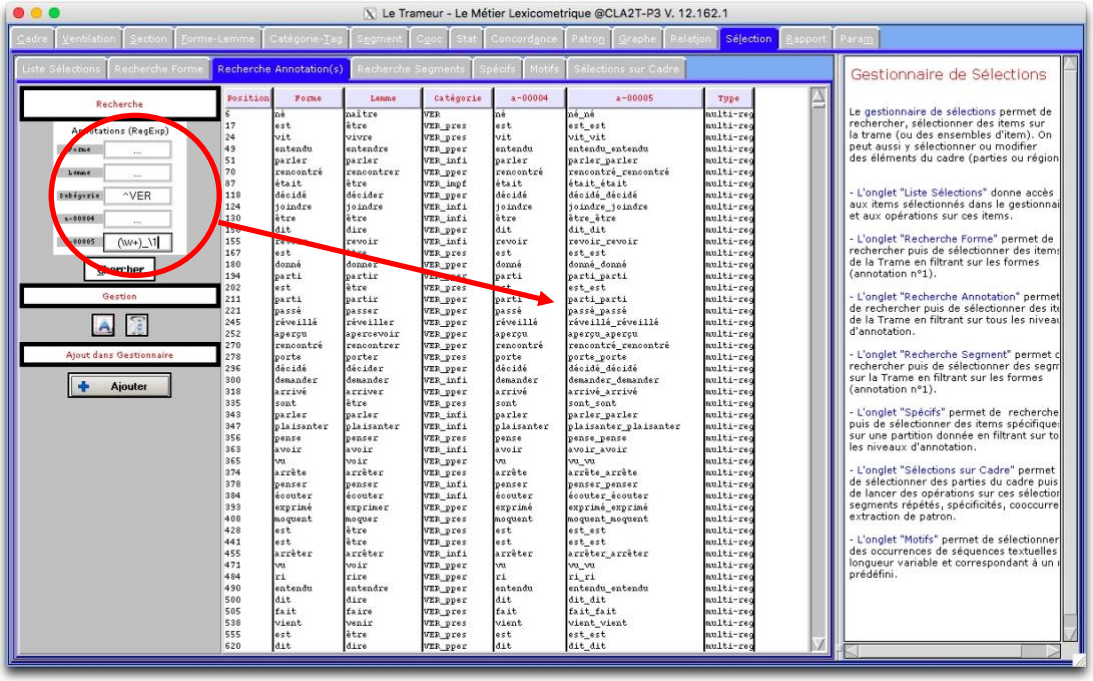


Fig. 9 : Recherche des formes normatives

A partir l'étape précédente, on a obtenu une liste de formes verbales correctes. On les sélectionne et les ajoute dans le secteur « liste sélections », où on crée une nouvelle annotation avec la valeur 0 pour ces formes verbes correctes.

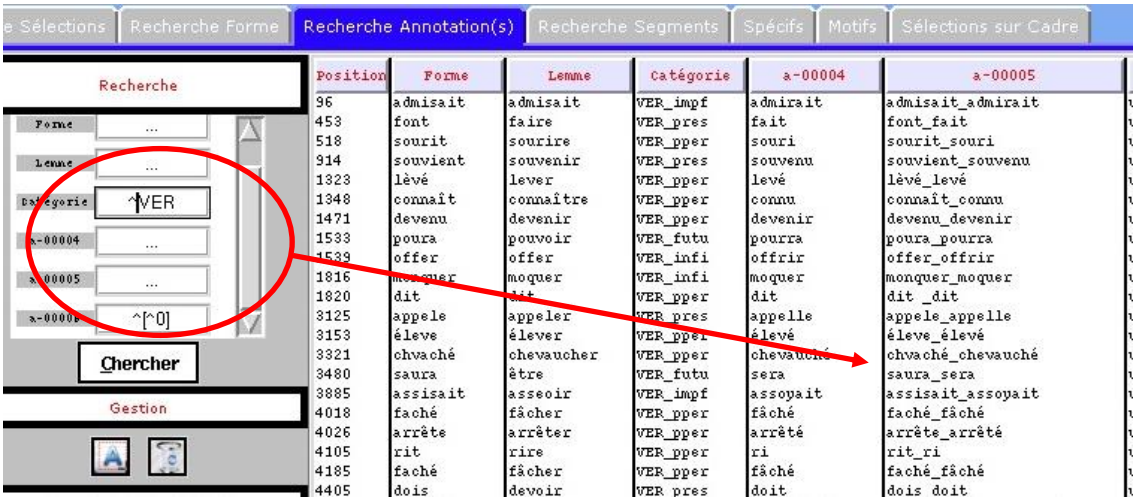


Fig. 10 : Création de la nouvelle annotation avec la valeur 0

Ainsi est créée la sixième couche d'annotation a-00006, dans laquelle les formes verbales normatives sont marquées avec la valeur 0. Dans l'étape suivante, afin d'extraire toutes les formes verbales non-normatives, on cherche les occurrences telles que l'annotation dans la sixième couche n'est pas 0. Et le consigne sera : ^[^0] dans le crochet « a-00006 ».

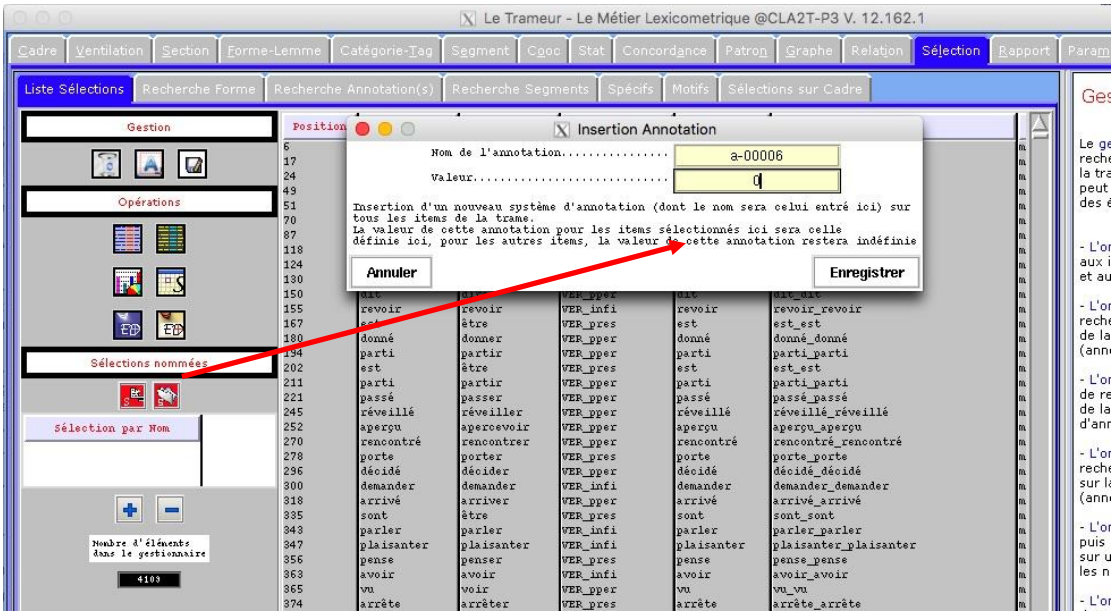
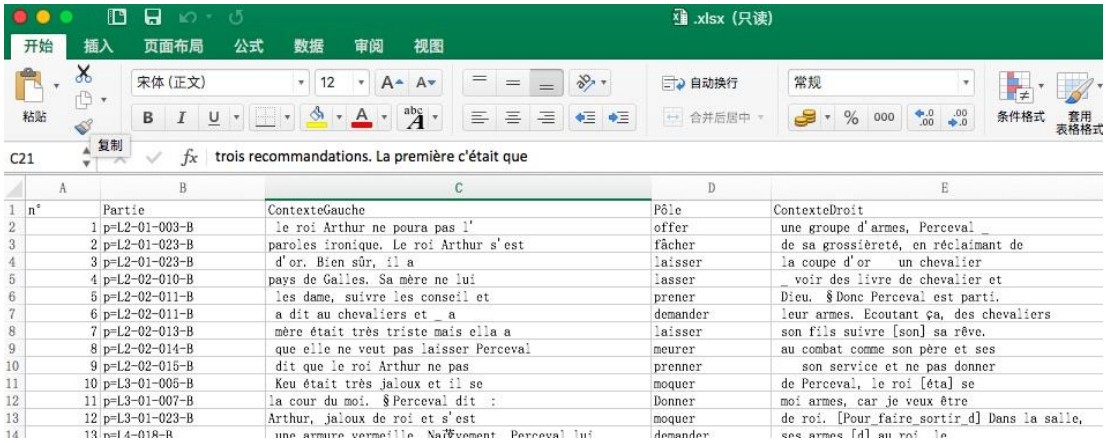


Fig. 11 : Création de la sixième couche d'annotation

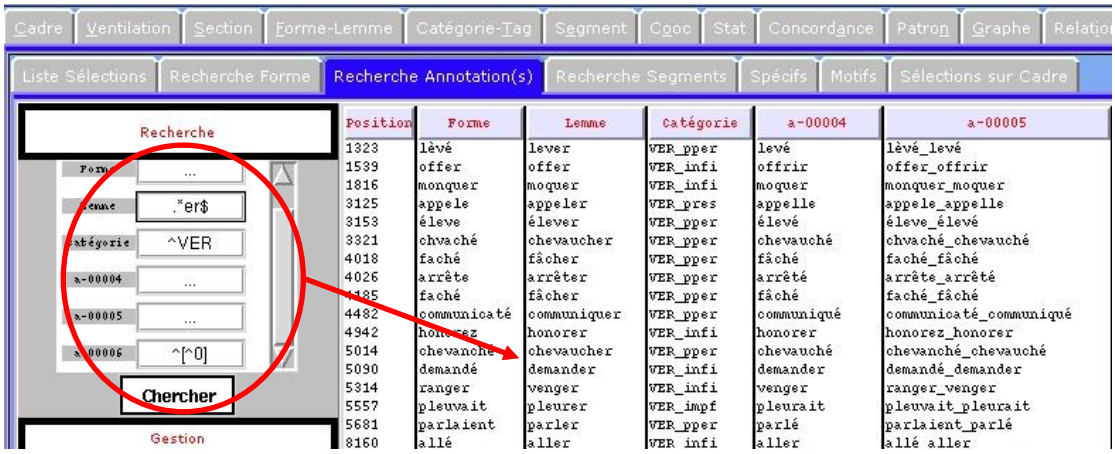
Finalement on a obtenu la liste de formes verbales erronées. Comme le montre l'image suivante, on pourrait également exporter ces formes verbales et leurs concordances (contexte gauche et contexte droite) dans un fichier au format Excel, afin de faire d'autres opérations de calcul.



	n°	Partie	ContexteGauche	Pôle	ContexteDroit
1	1	p=L2-01-003-B	le roi Arthur ne pourra pas l'	offer	une groupe d'armes, Perceval _
2	2	p=L2-01-023-B	paroles ironique. Le roi Arthur s'est	fâcher	de sa grossièreté, en réclaimant de
3	3	p=L2-01-023-B	d'or. Bien sûr, il a	laisser	la coupe d'or un chevalier
4	4	p=L2-02-010-B	pays de Galles. Sa mère ne lui	lasser	_ voir des livre de chevalier et
5	5	p=L2-02-011-B	les dame, suivre les conseil et	prener	Dieu. § Donc Perceval est parti.
6	6	p=L2-02-011-B	a dit au chevaliers et _ a	demander	leur armes. Ecoutant ça, des chevaliers
7	7	p=L2-02-013-B	mère était très triste mais ella a	laisser	son fils suivre [son] sa rêve.
8	8	p=L2-02-014-B	que elle ne veut pas laisser Perceval	meurer	au combat comme son père et ses
9	9	p=L2-02-016-B	dit que le roi Arthur ne pas	prenner	son service et ne pas donner
10	10	p=L3-01-005-B	Keu était très jaloux et il se	moquer	de Perceval, le roi [éta] se
11	11	p=L3-01-007-B	la cour du moi. § Perceval dit :	Donner	moi armes, car je veux être
12	12	p=L3-01-023-B	Arthur, jaloux de roi et s'est	moquer	de roi. [Pour_faire_sortir_d] Dans la salle,
13	13	p=L4-018-B	une armure vermeille. Naïvement, Perceval lui	demander	ses armes [d] au roi le

Fig. 12 : Liste des formes non-normatives

Pour extraire les formes verbales non-normatives, les recherches plus affinées seront également possibles avec *Le Trameur*. L'image suivante montre la recherche des formes verbales erronées dont l'infinitif se termine par –er. Avec une telle opération, on se concentre sur le premier groupe de verbes.



Position	Forme	Lemme	Catégorie	a-00004	a-00005
1323	lèvé	lever	VER_pper	levé	lèvé_levé
1539	offer	offer	VER_infi	offrir	offer_offrir
1816	monquer	moquer	VER_infi	moquer	monquer_moquer
3125	appele	appeler	VER_pres	appelle	appele_appelle
3153	éleve	élever	VER_pper	élevé	éleve_élevé
3321	chvaché	chevaucher	VER_pper	chevauché	chvaché_chevauché
4018	faché	fâcher	VER_pper	fâché	faché_fâché
4026	arrête	arrêter	VER_pper	arrêté	arrête_arrêté
4185	faché	fâcher	VER_pper	fâché	faché_fâché
4482	communicaté	communiquer	VER_pper	communiqué	communicaté_communiqué
4942	honorez	honorer	VER_infi	honorer	honorez_honorer
5014	chevanché	chevaucher	VER_pper	chevauché	chevanché_chevauché
5090	demandé	demander	VER_infi	demander	demandé_demander
5314	ranger	venger	VER_infi	venger	ranger_venger
5557	pleuvait	pleurer	VER_impf	pleurait	pleuvait_pleurait
5681	parlaient	parler	VER_pper	parlé	parlaient_parlé
8160	allé	aller	VER_infi	aller	allé aller

Fig. 13: Recherche des formes verbales erronées dont l'infinitif se termine par *–er*

La recherche pourrait également être effectuée pour une sous-catégorie de verbe, telle que les formes verbales en imparfait :

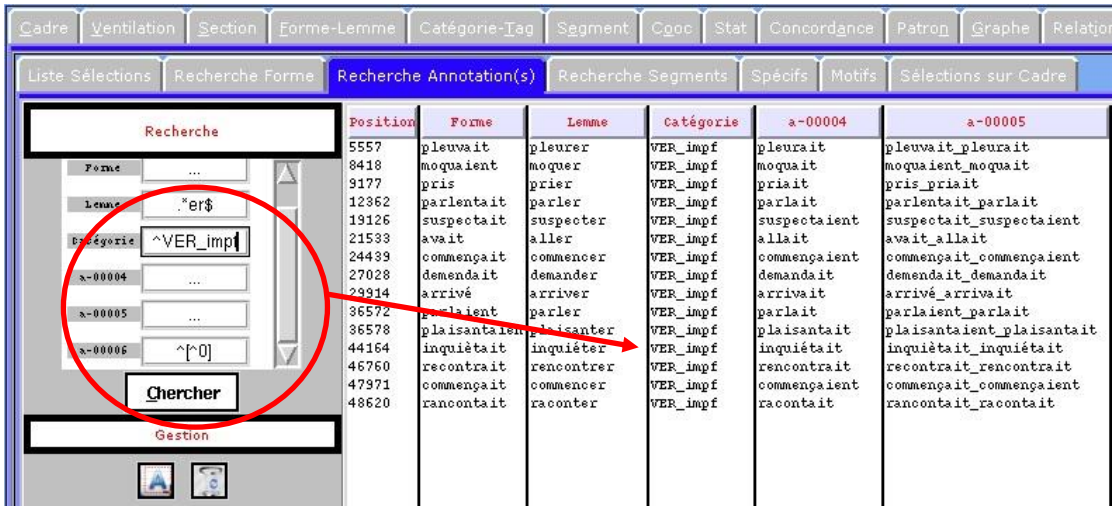


Fig. 14 : Recherche des formes verbales erronées de l'imparfait

3.3.3. Démarche d'analyse

Notons que les opérations avec *Le Trameur* susmentionnées nous permet de faire une analyse quantitative plus efficace. Et avant d'examiner les performances orthographiques des apprenants chinois, nous nous intéressons dans un premier temps les résultats bruts. Ainsi, nous pouvons répondre à une première question : quelles sont les zones de fragilité des apprenants chinois du point de vue global de leur maîtrise de l'orthographe française ? Le tableau suivant montre la répartition des erreurs en différents types et le poids relatif de chaque type d'erreur.

Tableau 7. La répartition des erreurs orthographiques

Types d'erreur		Nombre d'erreurs	Répartition d'erreurs (pourcentage)
1	Mauvais découpage du mot	9	1%
2	Substitution de mot	92	10%
3a	Erreur orthographique à dominante	133	15%

	phonétique		
3b	Erreur phonogrammique	27	3%
3c	Erreur d'orthographe lexicale sans altérer la valeur phonique	76	9%
4a	Erreur d'orthographe grammaticale : l'accord du nombre ou du genre	260	30%
4b	Erreurs d'orthographe grammaticale : les désinences verbales	284	32%
Total		881	100%

Première remarque à partir l'observation de ce tableau : les erreurs orthographiques les plus fréquentes des étudiants chinois de notre corpus sont les erreurs d'orthographe grammaticales, ainsi que les erreurs orthographiques à dominante phonétique.

Remarquons que les données se laissent certes traduire en chiffres, par le nombre des erreurs, mais comme l'indiquent Manesse *et al.*, « le sens de ces chiffres se donne dans la mise à jour des relations entre eux » (Manesse & Cogis, 2007 : 79). Ce sont les relations qui nous font comprendre les lacunes et les points forts chez les apprenants chinois lors de leur apprentissage de l'orthographe française. Donc dans la section suivante, la partie empirique, les erreurs orthographiques les plus fréquentes chez les apprenants chinois seront décrites et examinées de près. Afin de donner du sens aux données, nous nous attacherons à une analyse quantitative et qualitative.

La partie empirique se divise en trois parties. Dans la section 4.1, nous analysons les erreurs orthographiques à dominante phonétique, en les comparant avec d'autres erreurs d'orthographe lexicale, à savoir, les erreurs phonogrammiques et les erreurs d'orthographe lexicale sans altérer la valeur phonique. Nous aborderons ensuite le problème d'accord (l'accord en nombre et en genre) dans la section 4.2. La partie empirique se terminera par la discussion des erreurs orthographiques relevant des terminaisons verbales.

Cette partie empirique se veut la fois descriptive et comparative. D'un côté, pour chaque section, les données seront examinées de façon synchronique, afin de proposer une esquisse de la performance orthographique des apprenants chinois. De l'autre côté, pour mieux connaître le développement orthographique en français L2, les résultats obtenus par les apprenants chinois à des niveaux linguistiques différents seront comparés. En outre, les résultats seront commentés et les scénarios psycholinguistiques possibles pour interpréter les erreurs orthographiques seront également présentés, en faisant référence au développement langagier des apprenants L2, ainsi qu'aux divers facteurs en jeu dans le processus de la production écrite. Avec l'association des éléments de la recherche qualitative et quantitative, nous souhaitons mettre en lumière les traitements orthographiques utilisés par les apprenants chinois et leurs tendances interlangagiers spécifiques.