**Construire une base de données à partir de copies d'élèves**
**La base Orthocor des dictées du Balfroid 2012-2016**

Anne Dister (Université Saint-Louis - Bruxelles)
Marie-Louise Moreau (Université de Mons)

Disposant d'un ensemble de plus de 10.000 copies d'élèves, des dictées, nous souhaitions exploiter ce corpus de manière à pouvoir répondre à des questions telles que, par exemple, « Sur quels accords en genre et nombre les élèves sont-ils le plus nombreux à achopper ? », « Quelle proportion d'entre eux rencontre des problèmes avec les terminaisons verbales en /e/ : *-é, -er, -ez* ? », « Dans quelle mesure les graphies conformes aux rectifications orthographiques de 1990 sont-elles intégrées ? », « Sont-elles utilisées davantage par les élèves plutôt faibles en orthographe, ou indifféremment par les faibles et les forts ? ». La taille du corpus excluait un traitement manuel. En revanche, un tableur de type excel, qui aurait repris toutes les formes, aurait permis d'analyser facilement les données et d'apporter la réponse à de telles questions.

Nous avons donc élaboré un ensemble de procédures qui nous ont permis de disposer effectivement, pour chacune des 35 dictées dont nous disposions, de deux fichiers excel, l'un correspondant à l'ensemble des formes, correctes et incorrectes, l'autre reprenant seulement les graphies erronées. Nous avons pour ce faire suivi successivement les différentes étapes que nous allons décrire. Cet article n'a pas d'autre ambition que de mettre l'outil que nous avons conçu à la disposition éventuelle de chercheurs qui partageraient des préoccupations analogues aux nôtres.

## 1. Encodage des copies dans un fichier word

Le nombre de copies par dictée variait entre 192 et 618, le nombre de mots, entre 101 et 182. Si nous avions choisi de reporter les graphies incorrectes directement en excel, en consacrant une colonne pour les réalisations de chaque élève et une ligne pour chacun des mots, le fichier aurait comporté plus de lignes et de colonnes que celles qui apparaissent à l'écran, ce qui aurait multiplié les manipulations nécessaires. C'est pourquoi nous avons préféré encoder les graphies erronées dans un fichier word, qui sera ensuite converti en un fichier excel.

### *Le texte cible*

Le texte cible a donc été tapé en word, puis copié-collé en autant d'exemplaires que de copies, avec les aménagements suivants, qui répondent à trois grands principes :

- Toutes les informations fournies par les copies papier doivent se retrouver dans le fichier électronique.
- Chaque unité séparée des autres par une espace dans word se retrouvera dans une case distincte dans excel. Nous avons donc séparé chaque mot des autres par une espace, y compris les mots avec apostrophes ou avec trait d'union ; des mots comme *passe-montagne* ou *peut-être* ont toutefois été considérés comme une seule unité graphique et encodés sans espace.
- Lorsque pour un mot, il existe deux variantes légitimes, on en retient une (celle qu'a priori on pense la plus fréquente), et par la suite, on n'encodera que l'autre ; ainsi, dans notre cas, le texte cible comporte les graphies conformes à la norme avant les rectifications orthographiques de 1990, et on n'a entré pour les copies que les variantes rectifiées.

***Les copies***

Les personnes chargées de l'encodage des copies ont respecté les consignes suivantes.

- **Numéroter** les textes comme indiqué sur les copies. Utiliser trois chiffres dans tous les cas (001, 002… 020, 021, 033… 234, 345…), précédés de #. La cible est numérotée #000. Le numéro doit être séparé du texte soit par une espace, soit par une marque de paragraphe.
- **Reporter,** copie par copie, les mots mal orthographiés, et les mots qui diffèrent de la cible (les graphies rectifiées en 1990, p.ex.) en graissant les mots concernés. (Le marquage sert surtout l'encodeur et lui permet de voir où en est son travail, mais ce n'est pas avec le graissage qu'on isolera après les formes incorrectes.)
- **Mots manquants :** mettre autant de @ que de mots séparés par une espace, que de mots manquants. Ex. : *je ne l'ai pas vu =:: je n'ai pas vu* sera codé *je n' @ ai pas vu.*
- Lettres écrites en exposant : les faire précéder d'un point d'exclamation (*cinq*$^{ème}$ est encodé *cinq!ième*).
- **Césures** inadéquates en fin de ligne : 2 barres : *ent//housiasme.*
- **Apostrophes** en fin de ligne : 2 barres // après l'apostrophe. Ex. : *l'//*
- **Lettres ou mots illisibles** : mettre un signe § là où elles devaient apparaitre.
- **Trait d'union en plus** : le coller au mot de gauche (dans excel, si ce trait d'union est encodé au début d'un mot, il est interprété comme le signe arithmétique, ce qui engendrera des complications). Ex. : *pomme- de- terre.*
- **Mots surnuméraires** : les encadrer avec {}, sans espace. Coller avec le signe +. Si plusieurs mots sont ajoutés, les unir par le signe +, sans espace, à l'intérieur des accolades. Unir aussi par un + la ou les unités répétées au mot de gauche ou à celui de droite, selon ce qui semble le plus opportun. Cette catégorie chevauche parfois celle des mauvaises segmentations. Appliquer les mêmes décisions à tous les cas d'une même catégorie.

**Tableau 1. Exemples d'encodage en cas de mots surnuméraires**

| Cible | Réalisation de l'élève | Codage |
|---|---|---|
| *tant de cadeaux mérités* | *tant de cadeaux tant mérités* | *tant de cadeaux {**tant}+mérités*** |
| *un lieu* | *un lieu un lieu* | *un lieu+{un+lieu}* |
| *boire tout* | *boire de tout* | *boire {de}+tout* |
| *sa capacité* | *sa sa capacité* | *sa+{sa} capacité* |
| *on a* | *on n'a* | *on+{n'} a* |
| *sait- elle* | *sait-t' elle* | *sait-+{t'} elle* |

- **Mauvaises segmentations**

  C'est avec les segmentations incorrectes que l'encodage se révèle le plus difficile, au début du moins. Il faut toujours s'accrocher aux principes suivants :

  - Il doit y avoir autant d'espaces, ni plus, ni moins, dans la réalisation de l'élève, que dans la cible.
  - Pour agréger des unités, utiliser le signe + (qui indique aussi que l'élève a laissé une espace entre les deux unités).
  - Quand deux codages sont possibles, choisir celle qui crédite l'élève d'un maximum de bonnes graphies.

Les tableaux suivants reprennent différents exemples de mauvaises segmentations et indiquent comment l'encodage les a traités.

**Tableau 2. L'élève écrit plus de mots que n'en comporte la cible**

| Cible | Réalisation de l'élève | Codage |
|---|---|---|
| sait | c'est | c'+est |
| tandis | tant dis | tant+dis |
| morose | **mort rose** | mort+rose |
| ainsi | l'aissi | l'+aissi |
| lançant | l'ançant | l'+ançant |
| **ses** | **c'est** | **c'+est** |
| **pourtant** | **pour tant** | **pour+tant** |

**Tableau 3. L'élève écrit moins de mots que n'en comporte la cible**

| Cible | Réalisation de l'élève | Codage |
|---|---|---|
| **bien sûr** | **biensur** | ::biensur ::biensur |
| **c' est** | **ses** | ::ses ::ses |
| l' amande | lamande | ::lamande ::lamande |
| l' amande | la mande | ::la+mande ::la+mande |
| n' aurait- on | naurai ton | ::naurai+ton ::naurai+ton ::naurai+ton |
| n' aurait- on | n-aurai ton | ::n-+aurai+ton ::n-+aurai+ton :: n-+aurai+ton |

Répéter la séquence incorrecte autant de fois que nécessaire pour que l'encodage contienne autant d'unités séparées par une espace qu'il n'y en a dans la cible. Fait précéder les unités d'une espace et du signe ::[1].

**Tableau 4. Exemples supplémentaires d'encodage pour les mots surnuméraires, les mots manquants et les segmentations incorrectes**

| Cible | Réalisation de l'élève | Codage |
|---|---|---|
| **néanmoins** | **n'est en moins** | n'+est+en+moins |
| **si on s'en sert** | **si on senserre** | **si on ::senserre ::senserre** |
| Il y a | Il l'y a | Il+{l'} y a |
| Il y a | Il-y-a | Il- y- a |
| **Il apprit** | **Il l'apprit** | Il+{l'} apprit (*appris* n'est pas « contaminé » par la mauvaise segmentation). |
| **Il apprit** | **Il la prit** | ::Il+la+prit ::Il+la+prit (*il* est impliqué : c'est à cause du *il* que l'élève écrit *la*). |
| **Il apprit** | **Il a pris** | Il a+pris |
| On est arrivé | On n'est arrivé | On+{n'} est arrivé |
| On n' est pas arrivé | On est pas arrivé | On @ est pas arrivé |
| Rien n'est facile | Rien est facile | Rien @ est facile |
| Peu importe | Peu-in-porte | Peu- in-+porte |
| Il n' y en a pas | Il n'y en n'a pas | Il n'y en+{n'} a pas |
| Ses livres | C'est ses livres | {c'+est}+ses livres |
| Il apprit | Il la prit | ::Il+la+prit ::Il+la+prit |
| *Elle l'a convoité* | *Elle la convoiter* | Elle ::*la* ::*la convoiter* |
| **Personne n'y trouverait à redire** | **Personne ni trouverai-t-à redire** | Personne ::ni ::ni trouverai-+{t-} à redire |

## Exemples d'encodage en word

#000 Mon cher ordinateur L' époque des recherches documentaires dans les revues hebdomadaires est presque révolue Nous gardions soigneusement des articles intéressants pour en faire bénéficier les enfants Désormais ces chers petits s' attellent à une recherche poussée et sans mal à leur ordinateur Dans chaque foyer trône cet appareil indispensable à notre temps L' encyclopédie que vous auriez consultée autrefois ne s' avère plus aussi utile pensez- vous Tous les problèmes qui vous semblaient obscurs s' éclaircissent aisément Vous pourriez bien sûr aussi vous amuser avec cet objet né avant vous Cependant l' explication que vous donnera votre instituteur restera toujours le complément indispensable à votre information

#002 Mon cher ordinateur L' époque des recherches documentaires dans les revues hebdomadaires est presque révolue Nous gardions soigneusement des articles intéressants pour en faire bénéficier les enfants Désormais **ses** chers petits s' **attelent** à une recherche **pousée** et **s'+en** mal à leur ordinateur Dans chaque foyer trône cet **apareil** indispensable à notre temps L' encyclopédie que vous auriez consultée autrefois ne s' avère plus aussi utile pensez- vous Tous les problèmes qui vous **semblez obscures** s' **éclairsissent ésemant** Vous **pouriez** ::**biensûr ::biensûr** aussi vous amuser avec cet objet né avant vous Cependant l' explication que vous donnera votre instituteur restera toujours le complément indispensable à votre information

[1] Initialement, nous avions adopté en ce cas le signe >. Il est apparu cependant que ce signe se révélait problématique dans le traitement du corpus en excel (par exemple, lorsqu'on supprime les doublons d'une liste, et qu'on veut savoir à combien d'exemplaires chacune des formes est représentée dans la liste initiale. Nous l'avons donc remplacé par ::.

#003 Mon cher ordinateur L' époque des recherches documentaires dans les revues hebdomadaires est presque révolue Nous gardions soigneusement des articles intéressants pour en faire bénéficier les enfants Désormais ces chers petits s' attellent à une recherche poussée et sans mal à leur ordinateur Dans chaque foyer trône cet appareil indispensable à notre temps L' encyclopédie que vous auriez consultée autrefois ne s' avère plus aussi utile pensez- vous Tous les problèmes qui vous semblaient obscurs s' éclaircissent aisément vous pourriez bien sûr aussi vous **amusez** avec cet objet né avant vous Cependant l' explication que vous donnera votre instituteur restera toujours le complément indispensable à votre information

#004 Mon cher ordinateur L' époque des recherches **documentaire** dans les revues **ebdomadaire** est presque révolue Nous gardions soigneusement des articles **interessant** pour en faire bénéficier les enfants Désormais ces chers petits s' **appellent** à une recherche poussée et sans mal à leur ordinateur Dans chaque foyer trône cet appareil indispensable à notre temps L' encyclopédie que vous auriez **consulté** autrefois ne **::savert ::savert** plus aussi utile pensez- vous **Tout** les problèmes qui vous **semblait obscur** s' **éclersisse aisemment** Vous **pouriez** bien sûr aussi vous **amusez** avec cet objet né avant vous Cependant l' explication que vous donnera votre instituteur restera toujours le complément indispensable à votre information

## 2. Conversion du fichier word en un premier fichier excel, dit « complet »

- Faire une copie du fichier.
- Mettre des sauts de paragraphe (^p) supplémentaires entre les copies : remplacer tous les ^p^p par ^p1^p^2p^3p^4^p5^p (ceci permettra ultérieurement de repérer plus facilement les cas où l'encodage n'aurait pas respecté le principe : « Chaque unité cible doit correspondre à un ensemble de caractères pourvu d'une espace de part et d'autre »).
- Remplacer toutes les espaces par une marque de paragraphe.
- Sélectionner l'ensemble du fichier (Ctrl A), Copier =:: Coller dans excel.
- Dans le fichier excel, faire une macro :
    - Positionner le curseur en B1.
    - « Enregistrer une macro ».
    - Se positionner à la case de A où commence la 1re copie (càd après les lignes de A qui correspondent à la cible, et les 1 2 3 4 5 qui suivent).
    - Sélectionner tous les mots de la 1re copie et les 1 2 3 4 5.
    - Couper.
    - Se positionner en B1.
    - Insérer les cellules coupées.
    - Se positionner après les 1 2 3 4 5 de la cible.
    - Supprimer toutes les lignes correspondant à la copie qu'on vient de couper et coller, y compris ses 1 2 3 4 5.
    - Positionner le curseur en B1.
    - « Arrêter l'enregistrement ».
- Pour avoir les numéros des copies dans l'ordre de 001 au dernier, faire un tri du fichier de gauche à droite.
- Éventuellement, convertir les apostrophes word en apostrophes excel.

- Supprimer toutes les espaces.
- Identifier les fautes (faire disparaitre les formes correctes) : dans les colonnes (F-H, dans le tableau 5) qui suivent le corpus, utiliser la fonction « Si ». P.ex. : si B3 = A3 ; « » ; sinon B3).

  Voici un petit échantillon de ce qu'on obtient dans le fichier excel. Les quatre colonnes de droite (A-D) correspondent à la cible et à toutes les réalisations des élèves #004, #005 et #006. Dans les quatre colonnes de gauche, on a la cible, et les formes incorrectes rencontrées chez ces mêmes élèves.

**Tableau 5. Échantillon d'un fichier excel « complet »**

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | #000 | #004 | #005 | #006 | #000 | #004 | #005 | #006 |
| 2 | Les | Les | Les | Les | Les | | | |
| 3 | jeux | jeux | jeux | jeux | jeux | | | |
| 4 | électroniques | électroniques | **éléctroniques** | **éléctroniques** | électroniques | | éléctroniques | éléctroniques |
| 5 | Les | Les | Les | Les | Les | | | |
| 6 | jouets | jouet | jouets | jouets | jouets | jouet | | |
| 7 | que | que | que | que | que | | | |
| 8 | vos | vos | vos | vos | vos | | | |
| 9 | grands-parents | grands-parents | grands-parents | **grand-parents** | grands-parents | | | grand-parents |
| 10 | ont | ont | ont | ont | ont | | | on |
| 11 | connus | **connu** | connus | connus | connus | connu | | connu |

Les colonnes de gauche (E-H) sont copiées (« Collage spécial », « Valeurs ») sur un autre feuillet du fichier. C'est sur ces valeurs figées qu'on travaillera.

On peut facilement, à partir de là, calculer le nombre[2] de formes incorrectes pour chaque élève, ou pour chacun des mots cibles, mettre la référence de chaque mot cible (quelle dictée, quelle position du mot dans le texte, pourcentage de formes erronées…). On aboutit alors à un fichier comme celui proposé dans le tableau 6.

**Tableau 6 : Échantillon d'un tableau excel avec informations supplémentaires**

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | | | #000 | #003 | #004 | #005 |
| 2 | Localisation du mot | % incorrects | Cible | 2 | 1 | 4 |
| 3 | 12D-Electr-001 | 0,0 | Les | | | |

[2] Nous avons utilisé la fonction NB.VIDE pour chaque colonne, soustrait le nombre ainsi obtenu du total de mots ; de même pour chaque ligne, en soustrayant cette fois le nombre de l'effectif des élèves.

| | | | | | | |
|---|---|---|---|---|---|---|
| 4 | 12D-Electr-002 | 0,0 | jeux | | | |
| 5 | 12D-Electr-003 | 66,7 | électroniques | | éléctroniques | éléctroniques |
| 6 | 12D-Electr-004 | 0,0 | Les | | | |
| 7 | 12D-Electr-005 | 33,3 | jouets | jouet | | |
| 8 | 12D-Electr-006 | 0,0 | que | | | |
| 9 | 12D-Electr-007 | 0,0 | vos | | | |
| 10 | 12D-Electr-008 | 33,3 | grands-parents | | | grand-parents |
| 11 | 12D-Electr-009 | 33,3 | ont | | | on |
| | 12D-Electr-010 | 66,7 | connus | connu | | connu |

On peut également, dans ce fichier, ajouter les informations disponibles sur les variables « élèves » : sexe, niveau socioculturel de la famille, classe, etc.

Pour simplifier la présentation, nous emploierons désormais les termes suivants :

- les lignes, colonnes ou cases que nous dirons « bleues » sont celles où on a les données relatives à l'identification des élèves (les lignes 1 et 2 du tableau 6) ;
- les lignes, colonnes ou cases « vertes » sont celles qui contiennent les données relatives à la localisation des mots, au pourcentage des formes erronées (les colonnes 1, 2 et 3).

## 3. Conversion du fichier excel « complet » en un deuxième fichier excel, dit « liste »

Le fichier dit « Complet » peut être converti en un fichier moins lourd, ne contenant que les formes incorrectes, avec toutes les informations utiles pour leur identification.

Pour aboutir à ce résultat :

- Prendre pour point de départ le fichier « complet », dont on a une illustration dans le tableau 6.
- Juste sous la dernière ligne bleue, insérer autant de lignes qu'il y a de colonnes vertes (trois, dans notre cas).
- Y copier-coller (« Collage spécial », « Transposer ») les cases « vertes » du premier mot cible dans chacune des cases correspondant aux réalisations des élèves (D-F).
- Effacer le contenu des cases vertes qu'on vient de copier-coller.
- Positionner le curseur sur la première case verte suivante, dans la colonne A, choisir « Utiliser les références relatives » et enregistrer :

- Insérer autant de lignes qu'il y a de lignes bleues.
- Y copier-coller les lignes bleues.
- En dessous de ces lignes bleues, insérer autant de lignes que de colonnes vertes.
- Y copier-coller (« Collage spécial », « Transposer ») les cases « vertes » du mot cible suivant dans chacune des cases correspondant aux réalisations des élèves (D-F).
- Effacer le contenu des cases vertes qu'on vient de copier-coller.
- Positionner le curseur sur la première case verte remplie de la colonne A.
- Arrêter l'enregistrement de la macro.

- Répéter l'opération, avec la macro, pour les mots cibles suivants.
- Copier-coller les données sur un autre feuillet : « Collage spécial », « Tout avec le thème source », « Transposer ». On obtient alors un fichier avec la structure du tableau 7.

**Tableau 7 : Échantillon d'un fichier excel « tampon »**

| | A | B | C |
|---|---|---|---|
| 1 | #003 | #004 | #005 |
| 2 | 2 | 1 | 4 |
| 3 | 12D-Electr-001 | 12D-Electr-001 | 12D-Electr-001 |
| 4 | 0 | 0 | 0 |
| 5 | Les | Les | Les |
| 6 | | | |
| 7 | #003 | #004 | #005 |
| 8 | 2 | 1 | 4 |
| 9 | 12D-Electr-002 | 12D-Electr-002 | 12D-Electr-002 |
| 10 | 0 | 0 | 0 |
| 11 | jeux | jeux | jeux |
| 12 | | | |
| 13 | #003 | #004 | #005 |
| 14 | 2 | 1 | 4 |
| 15 | 12D-Electr-003 | 12D-Electr-003 | 12D-Electr-003 |
| 16 | 66,7 | 66,7 | 66,7 |
| 17 | électroniques | électroniques | électroniques |
| 18 | | éléctroniques | éléctroniques |
| 19 | #003 | #004 | #005 |
| 20 | 2 | 1 | 4 |
| 21 | 12D-Electr-004 | 12D-Electr-004 | 12D-Electr-004 |
| 22 | 0 | 0 | 0 |
| 23 | Les | Les | Les |
| 24 | | | |
| 25 | #003 | #004 | #005 |
| 26 | 2 | 1 | 4 |
| 27 | 12D-Electr-005 | 12D-Electr-005 | 12D-Electr-005 |
| 28 | 33,3 | 33,3 | 33,3 |
| 29 | jouets | jouets | jouets |
| 30 | jouet | | |
| 31 | #003 | #004 | #005 |
| 32 | 2 | 1 | 4 |

| | | | |
|---|---|---|---|
| 33 | 12D-Electr-006 | 12D-Electr-006 | 12D-Electr-006 |
| 34 | 0 | 0 | 0 |
| 35 | que | que | que |
| 36 | | | |
| 37 | #003 | #004 | #005 |
| 38 | 2 | 1 | 4 |
| 39 | 12D-Electr-007 | 12D-Electr-007 | 12D-Electr-007 |
| 40 | 0 | 0 | 0 |
| 41 | vos | vos | vos |
| 42 | | | |
| 43 | #003 | #004 | #005 |
| 44 | 2 | 1 | 4 |
| 45 | 12D-Electr-008 | 12D-Electr-008 | 12D-Electr-008 |
| 46 | 33,3 | 33,3 | 33,3 |
| 47 | grands-parents | grands-parents | grands-parents |
| 48 | | | grand-parents |
| 49 | #003 | #004 | #005 |
| 50 | 2 | 1 | 4 |
| 51 | 12D-Electr-009 | 12D-Electr-009 | 12D-Electr-009 |
| 52 | 33,3 | 33,3 | 33,3 |
| 53 | ont | ont | ont |
| 54 | | | on |
| 55 | #003 | #004 | #005 |
| 56 | 2 | 1 | 4 |
| 57 | 12D-Electr-010 | 12D-Electr-010 | 12D-Electr-010 |
| 58 | 66,7 | 66,7 | 66,7 |
| 59 | connus | connus | connus |
| 60 | connu | | connu |

- Copier-coller ce tableau dans un autre feuillet. « Collage spécial », « Tout, avec le thème source », « Transposé ».

**Tableau 8 : Échantillon d'un fichier excel « tampon 2 »**

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | #003 | 2 | 12D-Electr-001 | 0 | Les | | #003 | 2 | 12D-Electr-002 | 0 | jeux | | #003 | 2 | 12D-Electr-003 | 66,7 | électroniques | |
| 2 | #004 | 1 | 12D-Electr-001 | 0 | Les | | #004 | 1 | 12D-Electr-002 | 0 | jeux | | #004 | 1 | 12D-Electr-003 | 66,7 | électroniques | éléctroniques |
| 3 | #005 | 4 | 12D-Electr-001 | 0 | Les | | #005 | 4 | 12D-Electr-002 | 0 | jeux | | #005 | 4 | 12D-Electr-003 | 66,7 | électroniques | éléctroniques |

- Créer une autre macro, pour empiler les colonnes des cases bleues, celles des vertes et des roses.
  - Positionner le curseur en G1.
  - Sélectionner G1-L3.
  - « Insérer les cellules copiées » en A4.

- Supprimer les lignes G-L, maintenant vides.
- Positionner le curseur en G1.
- Arrêter l'enregistrement

- Trier sur F.
- Supprimer toutes les cases vides (elles correspondent toujours aux formes correctes).

**Tableau 9 : Échantillon d'un fichier excel « Liste »**

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **N° élève** | **N fautes** | **Localisation du mot** | **% fautes** | **Cible** | **Réalisation de l'élève** |
| 2 | #004 | 1 | 12D_Electr_003 | 66,7 | électroniques | éléctroniques |
| 3 | #005 | 4 | 12D_Electr_003 | 66,7 | électroniques | éléctroniques |
| 4 | #003 | 2 | 12D_Electr_005 | 33,3 | jouets | jouet |
| 5 | #005 | 4 | 12D_Electr_008 | 33,3 | grands-parents | grand-parents |
| 6 | #005 | 4 | 12D_Electr_009 | 33,3 | ont | on |
| 7 | #005 | 4 | 12D_Electr_010 | 66,7 | connus | connu |
| 8 | #003 | 2 | 12D_Electr_010 | 66,7 | connus | connu |

Cette base de données, un peu laborieuse à établir, s'est révélée d'une très grande commodité pour le traitement. Bien qu'elle ne perde aucune des informations des copies, elle permet en effet tous les tris, catégorisations, comptages de n'importe quel fichier excel, et offre un très grand potentiel pour les différentes analyses envisageables.