L'orthographe des Français dans le Grand Débat

Baromètre sur le niveau réel de maîtrise de la langue française

Mai 2019

Une étude menée par

Orthodidacte

La maîtrise des écrits vous grandit

Le Grand Débat national a fait couler beaucoup d'encre ces derniers mois et a donné l'occasion aux Français d'exprimer leurs attentes sur de nombreux sujets. Plus de 250 000 personnes ont rédigé 170 millions de mots en quelques semaines : cela représente l'équivalent de plusieurs centaines de livres à étudier ! Impossible de faire des analyses manuelles, ni de s'arrêter sur les détails. Pour en percer les secrets, les linguistes utilisent des outils de traitement automatique des langues (TAL) pour regrouper automatiquement les idées qui se ressemblent et en faire ressortir des tendances.

Mais qu'en est-il de l'orthographe dans tous ces écrits ? Cette étude va vous faire entrer dans les profondeurs de l'IA et du big data, appliqués au plus gros corpus contributif jamais réalisé. Pour la première fois se dessine le portrait des erreurs réelles que nous faisons régulièrement.

Chez Orthodidacte, notre équipe de R&D s'est donné pour mission d'aider les Français à bien écrire. Dans le Grand Débat comme dans la vie professionnelle, une bonne orthographe permet de mieux faire entendre sa voix. Nous avons donc saisi cette occasion unique pour étudier l'orthographe des Français dans leurs écrits de tous les jours.

Bonne lecture !

Michael Hiroux
Président fondateur d'Orthodidacte

L'objectif de cette étude est de réaliser un **instantané de l'orthographe des Français**, dans leurs écrits de tous les jours, quand ils s'expriment librement. Toutes ces données sont **au plus près de la réalité** des usages contemporains.

D'autres études plus anciennes ont déjà été publiées, mais elles avaient été réalisées à partir de questionnaires incomplets, d'enquêtes au cas par cas, ou bien sur des corpus finalement très restreints.

Comment avons-nous procédé ?

Nos experts linguistes ont d'abord récupéré l'intégralité du corpus, qui a été mis en ligne par l'État. Ils ont extrait les réponses aux questions ouvertes, en laissant de côté les réponses à des QCM, et ils ont supprimé toutes les réponses en double : il reste 130 millions de mots.

Ensuite, ils ont utilisé un outil unique, confectionné et perfectionné par nos soins, capable de détecter les erreurs de langue dans des textes gigantesques et de les classer. Chaque erreur détectée vient alors alimenter une immense typologie d'erreurs, directement branchée à la plateforme de formation Orthodidacte.

Qu'est-ce qu'un corpus ?

Un corpus est une immense collection de textes, récoltés suivant une méthode. Les linguistes exploitent les corpus comme de très vastes terrains de recherche, pour étudier la langue telle qu'elle se pratique réellement.

Qu'est-ce que le traitement automatique des langues ?

Le TAL est une branche de la linguistique qui est devenue omniprésente dans notre quotidien. Derrière les chatbots, derrière la reconnaissance de la parole, derrière la traduction automatique, il y a des outils de TAL.

Première donnée : *une erreur tous les 54 mots*

Les contributeurs au Grand Débat ont fait en moyenne une erreur tous les 54 mots. Quand on jette un œil à quelques contributions au hasard, on se rend compte que personne n'a rédigé en « langage SMS », par exemple.
Les contributions sont généralement bien construites et avec assez peu d'erreurs de langue.

Mais sur un si grand volume de texte, notre outil a tout de même collecté et analysé **près de 2,5 millions d'erreurs**…

Passons-les en revue.

Les chiffres clés de l'analyse

170 000 000
mots au départ

130 000 000
mots après suppression des contributions en double

2 390 000
erreurs linguistiques détectées

1,8 %
de mots avec une erreur linguistique

Équivalent à
250 fois
Les Misérables (Victor Hugo)

12 500 000
phrases au total

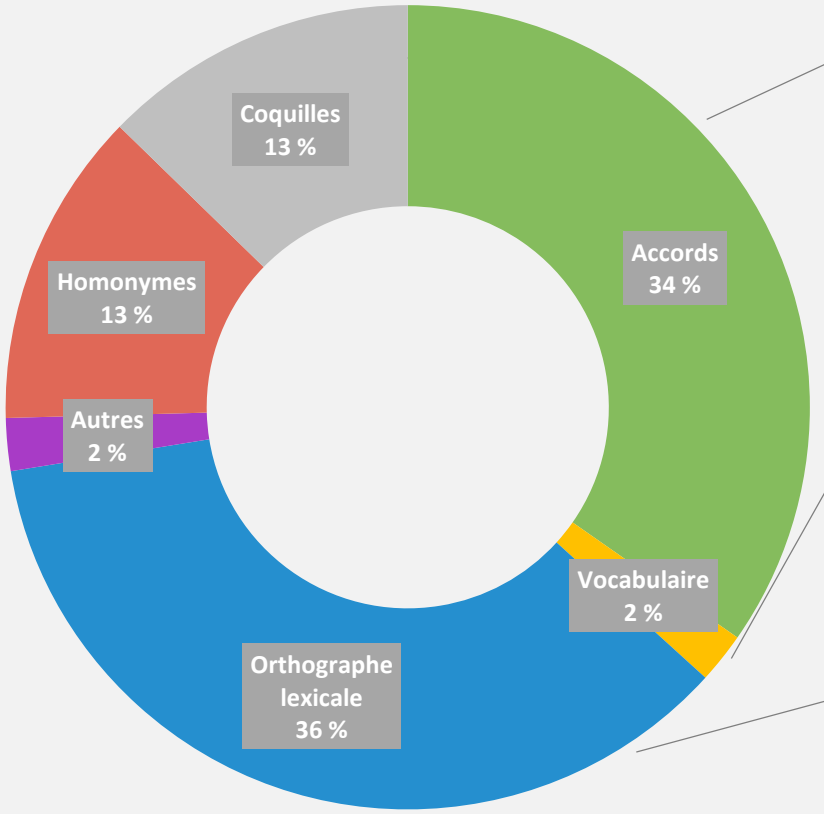
10 mots par phrase en moyenne

1 erreur tous les **54** mots

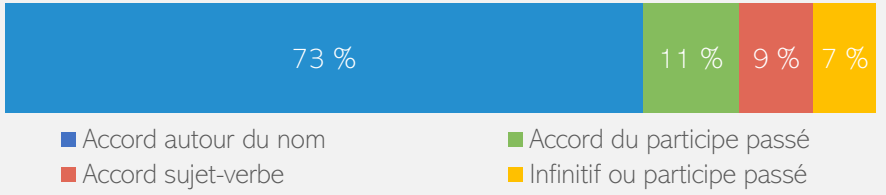
1 erreur toutes les
5 à 6 phrases
en moyenne

Le Grand Débat en chiffres

Répartition globale des erreurs

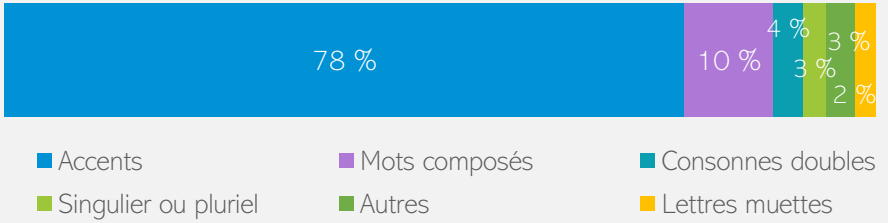


Accords : répartition des erreurs



Vocabulaire : répartition des erreurs

Orthographe lexicale : répartition des erreurs



** Pléonasme : répétition inutile. Exemple :* un monopole exclusif.
Paronyme : mot dont la prononciation est proche de celle d'un autre mot. Exemple : éditer une loi *mis pour* édicter une loi.

Impots ou *impôts* ?

Les accents

S'il fallait le résumer en une phrase : plutôt que de mettre les points sur les *i*, mettons les accents sur les *e* !

Pour pas moins de 680 000 mots, il manque un accent, à commencer par l'accent aigu sur *e* : son absence entraîne à elle seule 460 000 erreurs.

Les mots les plus concernés sont *etat* et *education*, deux mots souvent écrits avec une majuscule (quand il est question de l'État au sens d'« entité politique » et du ministère de l'Éducation nationale). Pour ces cas, ne pas mettre d'accent sur la majuscule n'est pas vraiment une erreur : les deux orthographes sont correctes, même s'il est recommandé de mettre les accents sur les majuscules.

Après *é*, les principales autres lettres accentuées sur lesquelles manque l'accent sont, par ordre de fréquence décroissante, *ô*, *ê*, *è*, *â*.

Pour 19 000 autres erreurs, la voyelle porte bien un accent, mais pas le bon : *trés* (1 771 erreurs), *régles*, *accés*, *arréter*, *étre*, *méme*, *société*, *déjà*, etc.

Top 10

etat	98 006 erreurs
education	24 049 erreurs
eviter	8 476 erreurs
ecole	7 687 erreurs
eduquer	6 791 erreurs
energie	5 430 erreurs
etats	4 917 erreurs
reduire	4 612 erreurs
deja	4 112 erreurs
depense	4 091 erreurs

Pourquoi tant d'erreurs d'accents ?

Cela provient probablement de deux raisons. D'une part, du désintérêt des Français pour ces signes qui apportent peu. D'autre part, de l'inadéquation des outils d'écriture tels que les claviers d'ordinateur, de smartphone et de tablette. Faute d'outil réellement adapté à l'écriture des accents (qu'ils soient difficiles à produire ou tout simplement inaccessibles), ceux-ci sont sacrifiés.

Ces signes accessoires, spécifiques de l'orthographe française, sont assez peu respectés, aussi bien dans l'écriture manuscrite que dans l'écriture sur écran, et sont perçus comme anodins. Le plus souvent, ces erreurs ne gênent pas la communication, même si certaines portent à confusion (surtout celles sur les homonymes).

Rond point ou *rond-point* ?

Les mots composés

Dans le domaine des mots composés, les erreurs portent sur l'utilisation du trait d'union, de l'espace, de la soudure et dans certains cas de l'apostrophe. 89 000 erreurs entrent dans cette catégorie. Elles consistent pour la plupart d'entre elles (84 %) à mettre un blanc là où un trait d'union est nécessaire.

L'absence de trait d'union concerne d'abord des mots grammaticaux : prépositions, pronoms, adverbes. Ces erreurs portent aussi, mais moins fréquemment, sur des noms comme *centre ville* : *savoir faire*, *rendez vous*, *assurance vie*, *week end*, *rond point*, etc.

Une erreur récurrente dans ce domaine concerne les noms géographiques composés, en particulier les départements français, souvent écrits par erreur sans traits d'union : *Haute Savoie*, *Seine Saint Denis*, *Alpes Maritimes*, et encore *Pays Bas*, *États Unis*, *Moyen Orient*, etc.

D'autres erreurs, comme *contre partie* (3 771 occurrences), *plate forme*, *main mise*, *bien-sûr*, *parce-que*, *parceque*, *entre-eux*, *co-propriété*, illustrent les autres types d'erreurs qui touchent les mots composés.

Top 10

au delà	11 256 erreurs
vis à vis	9 428 erreurs
au dessus	6 726 erreurs
ci dessus	3 569 erreurs
centre ville	3 243 erreurs
*peut etre**	2 567 erreurs
eux mêmes	2 120 erreurs
*eux même**	2 090 erreurs
celui ci	2 063 erreurs
*soit disant**	1 992 erreurs

* Mot comportant une deuxième erreur

Proportionelle ou *proportionnelle* ?

Les consonnes doubles

29 000 erreurs impliquant des consonnes doubles : cela semble assez peu, en comparaison des autres erreurs sur l'orthographe des mots ! **Or, contrairement aux erreurs d'accents ou de traits d'union, celles-là sont considérées comme plutôt graves.**

Quelles consonnes sont prioritairement concernées ? N, L, M et P représentent à elles seules la moitié des cas !

Les erreurs de consonnes doubles les plus fréquentes

consonne	nombre d'erreurs	une consonne de trop	une consonne de moins
N	6 638 erreurs	*carbonne*, *rationnaliser*, *innondations*	*proportionelle*, *environement*, *professionels*
L	4 992 erreurs	*familliales*, *renouveller*, *dévelloper**	*polution*, *poluants*, *reelement**
M	3 622 erreurs	*commité*, *forcemment**, *remmettre*	*évidement*, *consomation*, *suffisament*
P	2 970 erreurs	*juppe*, *handicappés*, *tapper*	*supression*, *suprimer*, *developer**

* Mot comportant une deuxième erreur

Exhorbitant ou *exorbitant* ?

Les lettres muettes

Les lettres muettes, une autre difficulté typique de l'orthographe française, sont impliquées dans 21 000 erreurs.
Pour 89 % d'entre elles, il s'agit d'erreurs sur une lettre muette à la fin du mot ; plus rarement, sur une lettre muette à l'intérieur du mot.

Exemples de lettres muettes finales ajoutées indûment : *soit disant* (1 992 erreurs), *ayants*, *parmis*, *chaques*.

À l'inverse, quelques cas de lettres muettes manquantes : *eux-même* (2 090 erreurs), *interdir*, *moin*.

Transports en communs ou *en commun* ?

Singulier ou pluriel

Dans les noms composés ayant la forme « nom + préposition + nom », il n'y a pas de règle pour savoir si le deuxième nom doit s'écrire au singulier ou au pluriel. C'est là aussi une source d'erreurs importante dans le Grand Débat, avec plus de 24 000 cas relevés.

Transports en communs (8 223 erreurs), *chiffre d'affaire*, *communauté de commune*, *moyens de transports*, figurent parmi les plus fréquentes.

Voie ou *voix* ?

Les homonymes (ou plutôt les homophones) sont des mots qui se prononcent de la même façon mais qui s'écrivent différemment. Et ils sont nombreux en français !
Sans surprise, dans cette catégorie d'erreurs, on retrouve en tête de la liste des erreurs les plus fréquentes deux paires de mots que seul un accent distingue.

D'abord *a* et *à*, pour un total de 140 000 erreurs, et ensuite *des* et *dès*, 35 000 erreurs. Ici, difficile de savoir quelles sont réellement la part d'erreur et la part de négligence car, comme on l'a vu, les accents sont souvent laissés de côté.

Quoi qu'il en soit, attention, les erreurs sur ces homonymes, dits homophones grammaticaux, sont assez mal vues. Il vaut mieux se donner la peine de chercher le moyen d'écrire *à* avec un accent grave, plutôt que de laisser croire qu'on ne fait pas la différence entre le verbe *avoir* et la préposition *à*.

D'autres paires d'homonymes ont entraîné un grand nombre d'erreurs. Il s'agit cette fois d'homonymes lexicaux, c'est-à-dire de mots tels que des noms, des adjectifs, confondus les uns avec les autres.

Les principales erreurs d'homonymes lexicaux

paire d'homonymes	exemples d'erreurs relevées
voie / voix	sur la voix publique, en voix de disparition
parti / partie	faire parti de
coût / coup	le coup de la vie, à faible coup
censé / sensé	nul n'est sensé ignorer la loi

Les hommes politique ou *les hommes politiques* ?

Quand on parle d'accords, on pense tout de suite au fameux accord du participe passé, dont les règles sont connues pour être particulièrement difficiles. Mais le français compte trois familles de règles d'accord : celle qu'on vient d'évoquer, ainsi que l'accord entre le sujet et le verbe et l'accord autour du nom.

En tête de liste : l'accord autour du nom

Contrairement à ce qu'on pourrait croire au premier abord, c'est cette catégorie qui a entraîné le plus d'erreurs dans le Grand Débat ! Et surtout l'accord de l'adjectif avec le nom auquel il se rapporte : 250 000 erreurs pour cette seule règle. Exemples : *de manière direct*, *les bénéfices important*, *sa situation social*, *les hommes politique*, etc.

Autre erreur très fréquente (175 000 cas) touchant l'accord autour du nom : l'accord entre le déterminant et le nom, même si la plupart des erreurs semblent dues à l'inattention et auraient pu être évitées par une simple relecture. Exemples : *la discriminations*, *aux suffrage*, *des campagne*, *à deux vitesse*, etc.

Attention aussi au mot *tout*, autour duquel on relève beaucoup d'erreurs : *tous le monde*, *toute les villes*, *tout embauche*, etc.

Passons au participe passé !

90 000 erreurs relevées, étonnamment réparties équitablement entre les trois groupes de règles : emploi avec l'auxiliaire *avoir*, emploi avec l'auxiliaire *être*, emploi sans auxiliaire. Eh oui ! l'accord du participe passé employé avec l'auxiliaire *avoir* a beau être réputé très difficile, il n'entraîne pas tant d'erreurs que ça dans le corpus. Pourquoi ? C'est certainement lié au format des réponses des contributeurs, donc aux questions posées. « Que faudrait-il faire pour… ? », « Quelles seraient pour vous les solutions… ? » : les questions ouvertes ont reçu des réponses de type sujet-verbe-complément dans environ 45 % des cas seulement. C'est peu ! Beaucoup de réponses étaient des phrases partielles construites autour d'un nom ou d'un verbe à l'infinitif. Autrement dit, les réponses suscitées ont une tendance à contenir peu de participes passés, donc on recueille peu d'erreurs dans ce secteur.

Mais enfin, il ne faut pas non plus négliger les 59 000 erreurs dans lesquelles un verbe en *-er* est confondu avec un verbe en *-é* ! C'est sûrement le contrecoup du grand nombre de verbes à l'infinitif.

Et enfin, l'accord sujet-verbe

Quant à la règle d'accord entre le sujet et le verbe, elle est à l'origine de 70 000 erreurs. Entre autres erreurs relevées, citons : *que les élus fasse leur travail*, *c'est cela qui les poussent*, *ce sur quoi porte les dépenses*. Pas toujours facile !

Allocation ou *allocution* ?

Les paronymes

Dans certains cas, écrire des mots qui se ressemblent est problématique, surtout quand leurs contextes d'utilisation sont proches ! Que ce soient les homonymes (mots qui se prononcent de la même façon mais s'écrivent différemment) ou les paronymes (mots dont la prononciation est proche), il peut être difficile de les distinguer. Certains de ces mots problématiques se sont frayé un chemin dans les données du Grand Débat ! Parmi les paronymes les plus fréquemment confondus : *éditer une loi* (pour *édicter une loi*), *recouvrir l'impôt* (au lieu de *recouvrer l'impôt*).

Tri sélectif ou *tri* tout court ?

Les pléonasmes

Nous avons aussi recherché la présence de pléonasmes dans les contributions : ce sont ces formules que nous utilisons parfois et qui se répètent inutilement !
Très peu de *monter en haut* dans le Grand Débat, le seul pléonasme que nous ayons repéré de façon récurrente est le fameux *tri sélectif* (14 000 occurrences) !

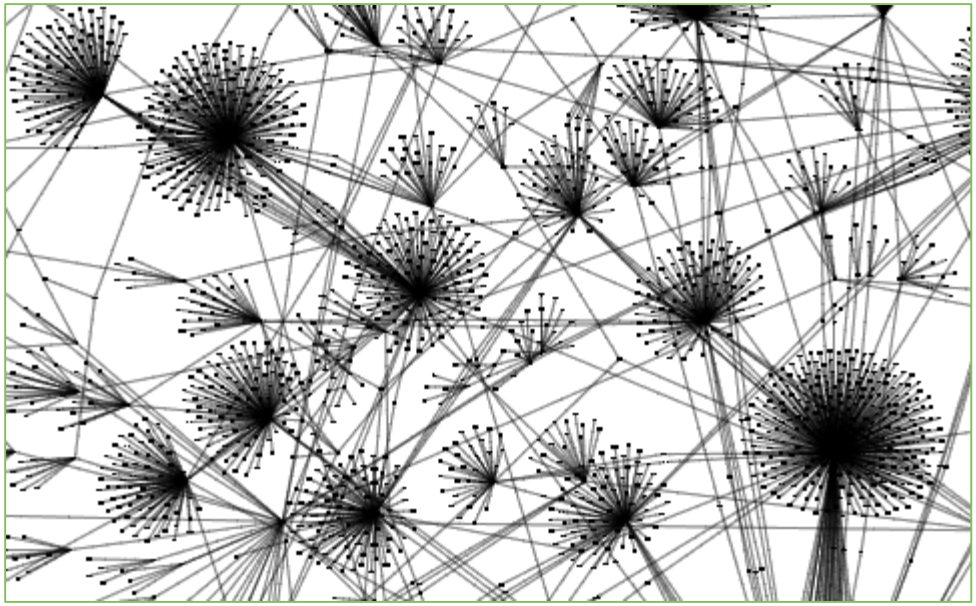
Une erreur où tout n'est pas forcément à jeter puisque, même s'il est toujours sélectif, le tri a su s'imposer dans les foyers sous cette appellation de « tri sélectif ».

Job ou *travail* ?

Les anglicismes

Les anglicismes sont bien souvent critiqués par les puristes de la langue française, mais cela ne les empêche pas de s'être immiscés dans notre vocabulaire courant !

Les mots *job*, *dumping*, *smartphone*, *business* ou encore *process* figurent parmi les anglicismes les plus fréquents dans le Grand Débat, mais il faut reconnaître qu'ils ne se laissent pas toujours remplacer par un équivalent francisé, malgré les nombreuses recommandations officielles qui les concernent.



Extrait de la visualisation des catégories d'erreurs :
un véritable champ de pissenlits !

Un (joli) champ d'erreurs

Les **2 390 000 erreurs détectées** par notre outil se connectent automatiquement dans une typologie des erreurs linguistiques.

L'IA de notre outil d'analyse nous propose cette vision artistique, qui prend parfois la forme d'un arbre à 1 000 branches, parfois celle d'un réseau connecté ou d'un champ de pissenlits…

En conclusion

C'est la première fois que le niveau de maîtrise de la langue française est étudié à partir de données réelles à une si grande échelle. Nous sommes très heureux de partager avec vous ces résultats et très fiers des outils incroyablement innovants développés par nos linguistes.

Loin de toutes considérations politiques ou sociologiques, cette étude montre plusieurs choses :

- Une erreur se glisse en moyenne tous les 54 mots.
- La majorité des erreurs sont récurrentes : un tiers des erreurs relevées portent sur les accords, un autre tiers sur l'orthographe lexicale.
- Le portrait dressé est plutôt loin des clichés sur la langue française : il n'y a pas que le participe passé qui soit source de difficultés, bien au contraire ! Les accents et les accords simples sont également des erreurs courantes.

L'étude de grands corpus et l'analyse des erreurs linguistiques nous aident à mieux comprendre les difficultés réelles des Français. Cela nous permet ainsi de construire des parcours d'apprentissage pour améliorer leurs écrits professionnels, répondant de manière optimale à leurs besoins en formation.
Nous sommes fiers qu'Orthodidacte contribue à la progression de la maîtrise du français, dans les entreprises et dans les écoles, en France et à l'étranger.

Un grand merci pour votre attention !

Orthodidacte

La maîtrise des écrits vous grandit

Chez Orthodidacte, nous développons depuis 10 ans notre **expertise en linguistique et en traitement automatique des langues** au service de la langue française.

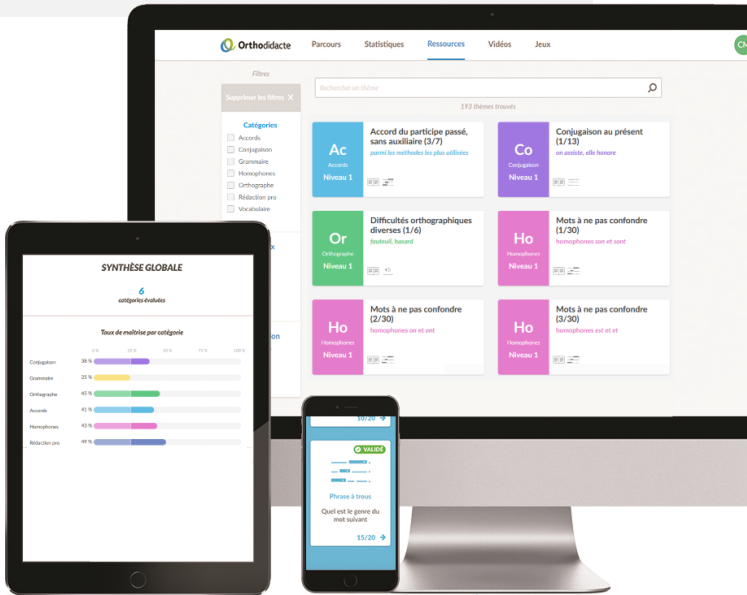
Cette expertise prend la forme de plusieurs plateformes technologiques :

- une **plateforme d'e-learning**, utilisée tous les jours par des centaines de milliers de collaborateurs, d'étudiants et de particuliers pour perfectionner leur communication écrite ;
- une **plateforme de certification**, avec la Certification Le Robert, permettant à chacun de certifier de manière fiable et rapide son niveau en langue française ;
- une **plateforme de dictées**, avec une correction instantanée des textes dictés (aux formats audio et vidéo), pour s'entraîner ou organiser des évènements géants en direct.

Nous apportons également à nos clients des services de **cartographie des compétences**, pour mesurer à un instant T le niveau de maîtrise des écrits professionnels des collaborateurs et pour identifier ceux pour lesquels une formation serait profitable.
La sensibilisation aux actions de formation prend également la forme d'évènements autour de la langue française, avec des **dictées ludiques** et des **jeux**.

Nos formations couvrent un large panel de difficultés de la langue française : des consonnes doubles au participe passé des verbes pronominaux, en passant par la concordance des temps, le registre de langue ou encore les difficultés grammaticales.

- Spécialiste des écrits professionnels
- Créé en 2009
- Basé à Grenoble
- Une vingtaine de collaborateurs



Merci aux Français !

Merci à tous les Français qui ont déposé une contribution dans le Grand Débat.

Les erreurs relevées n'enlèvent rien à la qualité des contributions.

La liberté d'expression passe avant les fautes d'orthographe !

Cette étude a été menée et rédigée par :

- Camille Martinez, docteur en sciences du langage et responsable linguistique d'Orthodidacte ;
- Baptiste Ranty, ingénieur linguiste et spécialiste TAL d'Orthodidacte.

Parlons-en ensemble !

Contact presse

Mélanie Seynat, responsable marketing et communication d'Orthodidacte

melanie.seynat@orthodidacte.com • +33 6 40 47 03 07

Contact commercial

bienvenue@orthodidacte.com • N°Cristal 0 969 395 797

APPEL NON SURTAXÉ

www.orthodidacte.com

Crédits photos : kues1 – rawpixel.com // Freepik.com