

Étude de l'usage de l'anglais dans les thèses en France

Naïma Beck

14 novembre 2024

Résumé

Ce travail s'est concentré sur l'étude de l'usage de l'anglais dans les thèses de doctorat en France, suite aux injonctions d'utiliser l'anglais comme langue d'écriture de la recherche. À travers une analyse des données issues de la base PHD, nous avons observé une tendance croissante de l'utilisation de l'anglais et avons tenté de l'expliquer au regard de la notion d'internationalisation.

Un autre angle d'analyse a été l'étude d'une saisonnalité particulière pour les soutenances de thèses, qui se révèle être les deux premières semaines de décembre, en raison d'une échéance pour l'inscription aux concours en mars. Notre base de données sur les thèses en France a été téléchargée sur le site du gouvernement. Nous l'avons exploité avec le langage R. De plus, une analyse descriptive a été réalisée pour résumer les caractéristiques des données.

Mots-clés: Thèse, Anglais académique, Internationalisation.

Table des matières

1	Introduction	3
1.1	Eléments conduisant à la question	3
1.2	Knowledge gap	3
1.3	La question de recherche	3
1.4	Méthodes et données	3
2	Méthodes et données	3
2.1	Jeu de données mobilisé	3
2.2	Outils	4
2.3	Données manquantes	4
2.4	Gestion des erreurs dans l'entrée des dates de soutenance	5
2.5	Gestion des homonymes	6
2.6	Gestion des outliers	6
3	Résultats	7
3.1	Langues d'écriture dans les thèses	7
3.2	Dates de soutenance	8
4	Discussion	9
4.1	Annonce de la structure de la discussion	9
4.2	Interprétations	9
4.2.1	Langue d'écriture	9
4.2.2	Date de soutenance	10
4.3	Conclusion	11
5	Références	11

Liste des figures

1	Corrélation entre les différentes variables concernant l'absence de données	4
2	Distribution des occurrences par mois et par année de 2005 à 2018	5
3	Proportion des thèses soutenues au premier janvier de 1985 à 2018	5
4	Proportion de thèses écrites en anglais selon les disciplines	8
5	Proportion des thèses soutenues au fil des mois de 2005 à 2018 sans le 1 ^e janvier	8
6	Somme des soutenances par jour en décembre de 2005 à 2018	9

Liste des tables

1	Classement des auteurs ayant soutenu le plus de thèses	6
2	Classement des directeurs de thèse ayant supervisé le plus grand nombre de thèses	7

1 Introduction

1.1 Eléments conduisant à la question

Le français est la langue de l'enseignement depuis la loi Toubon de 1994. Pourtant, comme le mentionnait Geneviève Fioraso, alors ministre de l'Enseignement Supérieur, l'introduction de cours en langues étrangères avait pour objectif de mettre fin à "une inégalité de fait" entre grandes écoles et universités et de répondre aux exigences de la mondialisation académique.

En parallèle, la recherche française était aussi très critiquée pour son manque d'ouverture à l'international. Qualifiée de renfermée sur elle-même car elle produisait des écrits en français limitant ainsi sa visibilité et son impact sur la scène internationale. Dès lors, il y a eu une injonction à l'utilisation de l'anglais comme langue de publication. O'Connell, A.-M., Chaplier, C. (2020) ont montré qu'étant donné les défis liés à la formation en langue à l'échelle mondiale, il arriva que, sous la pression sociale, institutionnelle et économique, l'anglais s'impose progressivement dans ce domaine, comme le stipule la loi dite Fioraso de 2013 qui encourage, entre autres, l'utilisation de l'anglais comme langue d'enseignement dans des disciplines autres que les langues vivantes.

1.2 Knowledge gap

Au cours des dernières années, plusieurs décisions politiques ont été prises pour promouvoir l'utilisation de l'anglais dans les universités et la recherche en France. Cependant, il est nécessaire d'en savoir plus sur la manière dont ces politiques affectent réellement les pratiques d'écriture des doctorants.

Cette étude est crucial pour mieux comprendre comment les injonctions à l'internationalisation dans le secteur de la recherche et de l'université se traduisent dans la réalité. Il serait donc profitable d'étudier si cette injonction s'est traduite dans les pratiques des doctorants.

1.3 La question de recherche

Dès lors, nous nous poserons la question suivante : Dans quelle mesure l'anglais a-t-il progressé comme langue d'écriture dans les thèses de doctorants en France ?

1.4 Méthodes et données

Nous répondrons à cette question en analysant les données de la base. Pour ce faire, il sera nécessaire de nettoyer les données en identifiant les valeurs manquantes, en recherchant les incohérences dans les dates, les homonymes, ainsi que les valeurs aberrantes(outliers).

2 Méthodes et données

2.1 Jeu de données mobilisé

Le jeu de donnée porte sur les soutenances de thèse en France. Il a été extrait du site gouvernemental des thèses et nous a été fourni par l'encadrant de ce papier Mr Matthieu Cisel. La recherche basée sur les métadonnées dans les archives ou les dépôts en ligne est fortement contrainte par la nature et la qualité des informations collectées

sur les manuscrits. Pour comprendre les principales variables du jeu de données, nous avons interrogé l'encadrant, examiné les valeurs associées et les noms attribués à ces variables. Il y a globalement des informations sur les doctorants (nom, genre), la thèse (nom, discipline, langue d'écriture), l'établissement, le directeur de thèse et les dates (de soutenance et d'inscription).

2.2 Outils

Nous avons utilisé le langage R avec les packages principaux **dplyr** (manipulation des données), **ggplot2** (visualisation graphique), **lubridate** (gestion des dates), et **stringr** (traitement des chaînes de caractères).

2.3 Données manquantes

Le travail préliminaire à la visualisation des données manquantes a consisté à identifier les erreurs de saisie dans les différentes colonnes. Pour ce faire, nous avons recherché dans la base de données des erreurs de saisie courantes, telles que des cases vides, des espaces, des valeurs comme 'na' ou 'unknown'. Ensuite, nous avons extrait les valeurs uniques de certaines colonnes, comme `Genre` et `Langue.de.la.thesis`, et avons remplacé les données insatisfaisantes par 'NA'. Enfin, pour les erreurs moins évidentes, c'est en naviguant et en visualisant les données que nous les avons identifiées.

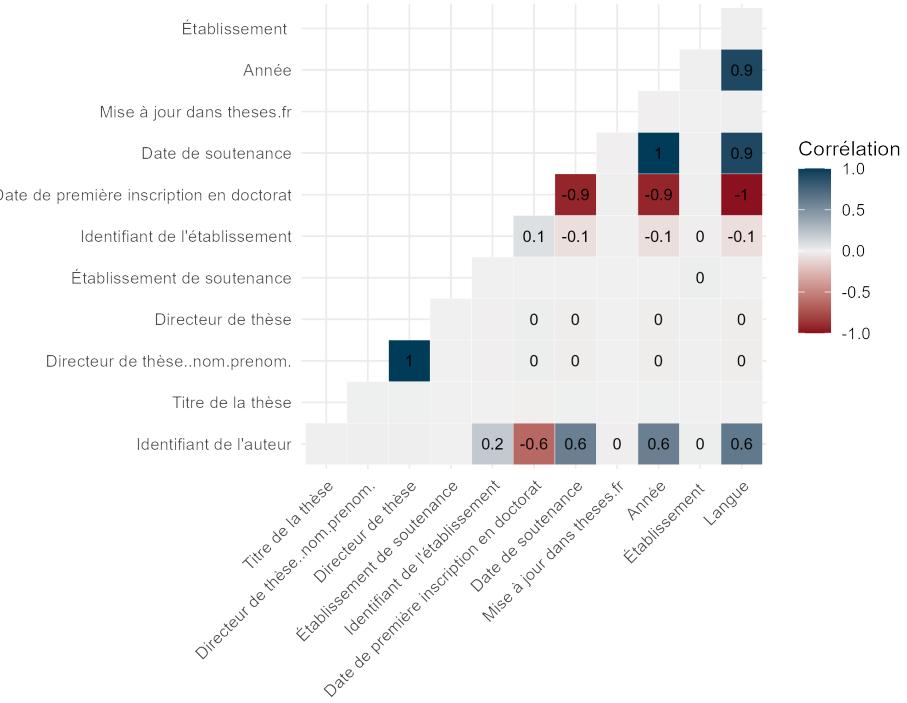


Figure 1: Corrélation entre les différentes variables concernant l'absence de données

On peut voir ici qu'il y a une forte corrélation concernant le manque de donnée pour l'année, la date de soutenance et la langue. Lorsqu'il manque l'année, il manque la date de soutenance et la langue. Par la suite, en plus de l'étude des langues, nous étudierons donc aussi la date de soutenance.

2.4 Gestion des erreurs dans l'entrée des dates de soutenance

Nous avons souhaité visualiser la fréquence des soutenances pour chaque mois, de janvier à décembre, sur l'ensemble des années de 2005 à 2018. On aimerait étudier s'il y a une saisonnalité particulière pour les soutenances, si certains mois sont plus populaires pour les soutenances que d'autres.

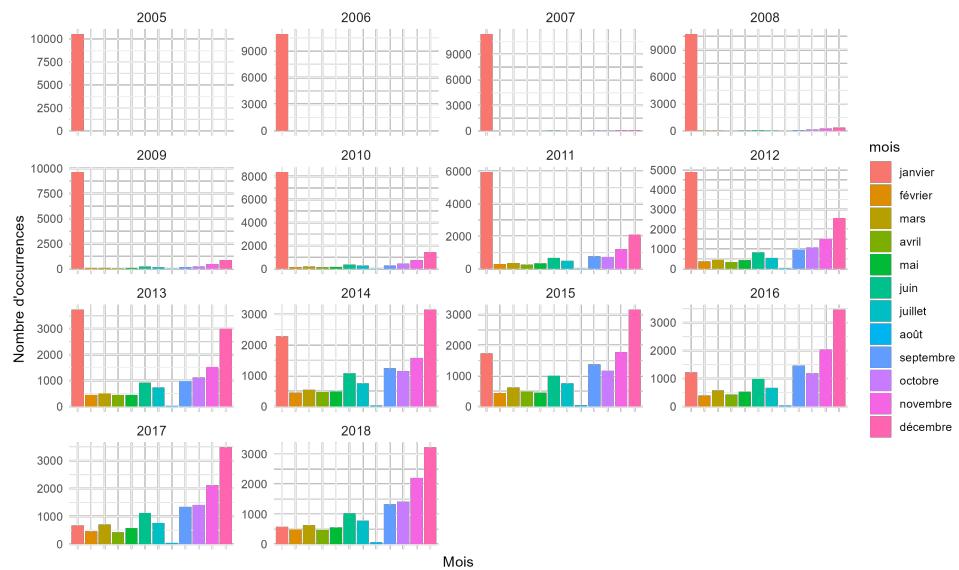


Figure 2: Distribution des occurrences par mois et par année de 2005 à 2018

On remarque que de 2005 à 2013, le mois de janvier est largement dominant en terme du nombre de thèses soutenue. Et qu'à partir de 2014, le mois de décembre comporte plus d'occurrences que les autres mois.

On pourrait en conclure que le mois de janvier était plus populaire au près des étudiants, cependant en visualisant la proportion de thèses soutenues le 1er janvier nous obtenons les résultats suivants :

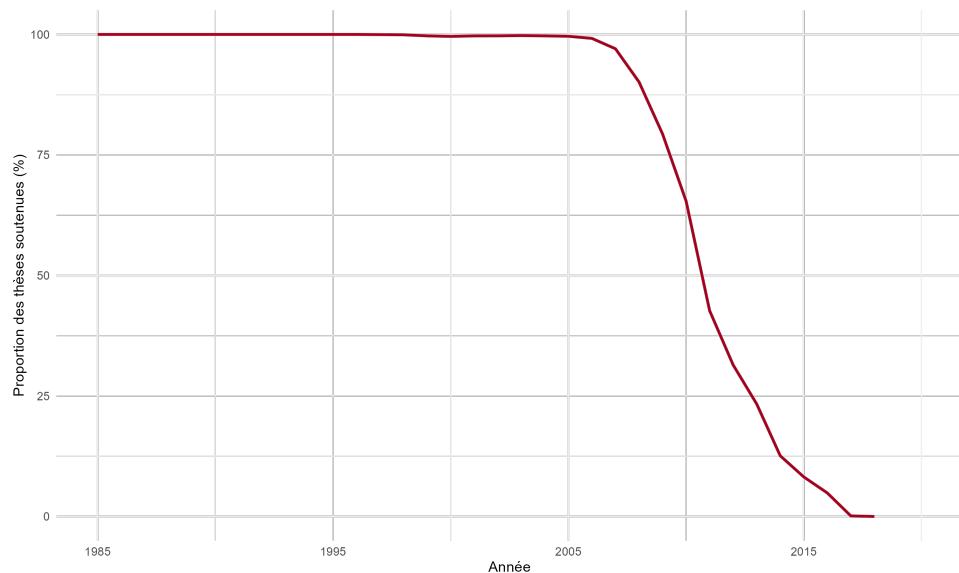


Figure 3: Proportion des thèses soutenues au premier janvier de 1985 à 2018

De 1985 à 2005, près de 100 % des thèses semblent avoir été soutenues le 1er janvier. Cependant, à partir de 2006, cette proportion diminue considérablement, jusqu'à approcher les 0 % en 2018. Il est évident qu'une telle inversion de tendance, sans cadre législatif imposant la soutenance des thèses le 1er janvier, relève d'une anomalie dans les données. Il est probable que, pour les thèses antérieures à 2006, seule l'année de soutenance ait été renseignée, et qu'une conversion automatique ait imposé le 1er janvier comme date par défaut lors de la mise au format numérique.

Nous nous accorderons donc à supprimer les données contenant le 1e janvier.

2.5 Gestion des homonymes

Pour traiter le cas des homonymes, il a fallu exploiter la colonne 'Auteur', dans laquelle plusieurs personnes pouvaient être renseignées. Nous avons donc amélioré cette colonne pour que chaque personne ait sa propre entrée. Ensuite, nous avons calculé le nombre de thèses par personne. En guise de vérification, nous avons pris l'exemple du doctorant Michel Mace. Voici le classement des personnes ayant réalisé le plus de thèses :

Rang	Auteur	Nombre de thèses soutenues
1	Nicolas Martin	16
2	Philippe Ascher	15
3	Philippe Michel	13
4	Franck Martin	12
5	Philippe Martin	12
6	Yang Liu	12
7	Celine Martin	11
8	Jing Wang	11
9	Laurent Martin	11
10	Olivier Martin	11

Table 1: Classement des auteurs ayant soutenu le plus de thèses

On peut penser que ce ne sont que des homonymes donc nous ne supprimerons pas ces entrées. Ainsi, nous avons travaillé sur la question d'homonomie qui soulève des difficultés possibles dans le suivi des individus au cours de leur carrière. C'est pourquoi nous décidons de nous concentrer sur l'étude des langues qui sera plus fiable.

2.6 Gestion des outliers

Nous avons effectué la même méthode pour le traitement des outliers, avec la colonne 'Directeur.de.these' et comme vérification le directeur Bely Lucien. Voici le classement des directeurs ayant encadré le plus de thèses :

Rang	Directeur de thèse	Nombre de thèses encadrées
1	Delebecque Philippe	68
2	Loiseau Gregoire	62
3	Sicard Michel	41
4	Franco Bernard	40
5	Dagen Philippe	39
6	Lalot Thierry	38
7	Hoffmann Christian	34
8	Vanier Alain	33
9	Peretti Jean-Marie	32
10	Gore Marie	31

Table 2: Classement des directeurs de thèse ayant supervisé le plus grand nombre de thèses

On peut expliquer ces chiffres par le fait que beaucoup de professeurs co-dirigent des thèses avec d'autres professeurs pour le prestige, ce résultat est intéressant donc nous avons décidé de ne pas supprimer ces entrées.

3 Résultats

Nous avons divisé la section résultats en trois sous-sections. Dans la première, nous étudierons la langue d'écriture et ensuite les dates de soutenance. Les résultats mettent en évidence l'évolution des pratiques linguistiques dans les thèses, montrant la progression de l'anglais face au français selon les disciplines. Ils permettent également d'observer les variations temporelles dans la soutenance des thèses, ainsi que les dynamiques qui influencent ces choix au fil du temps.

3.1 Langues d'écriture dans les thèses

Les résultats montrent que l'anglais est de plus en plus utilisé, bien que cela varie considérablement selon les disciplines et les institutions.

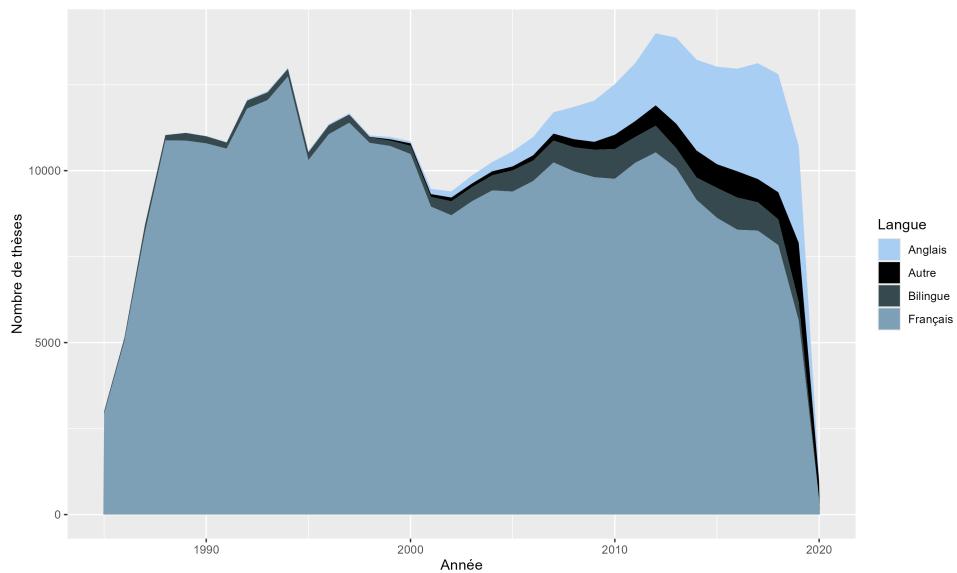


Figure 4: Proportion de thèses écrites en anglais selon les disciplines

Dans la figure 4, on voit que de 1971 à 2000, c'était le français qui était la langue la plus utilisée pour l'écriture des thèses. À partir des années 2000, on observe que l'anglais dépasse le français en tant que langue d'écriture des thèses, avec environ 12 500 thèses en anglais de 2010 à 2018, contre moins de 10 000 thèses en français durant la même période.

Nous avons vu que l'anglais est devenu la langue la plus utilisée dans l'écriture des thèses en France, mais il reste un point aveugle sur la date de soutenance des thèses, dès lors nous avons décidé d'en faire une étude supplémentaire.

3.2 Dates de soutenance

Après avoir filtré les données en retirant les dates contenant le 1e janvier, nous obtenons le résultat suivant :

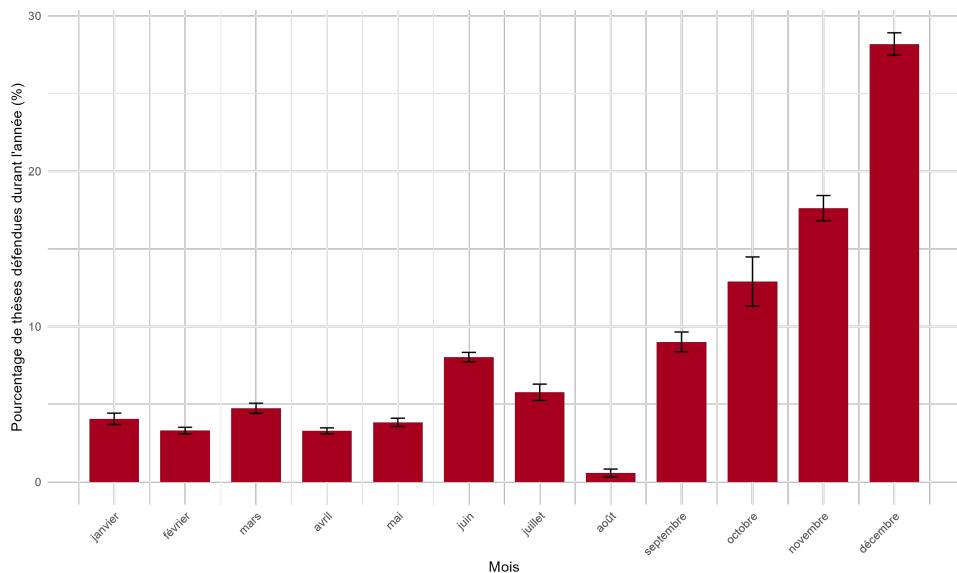


Figure 5: Proportion des thèses soutenues au fil des mois de 2005 à 2018 sans le 1e janvier

Dans la figure 5, on observe que les soutenances suivent un rythme très marqué au cours de l'année. Le mois d'août est quasiment déserté, tandis que la première moitié de l'année reste relativement calme avant un léger regain en juin. À partir de la rentrée, les soutenances deviennent de plus en plus fréquentes, culminant en fin d'année. En effet, on observe que c'est le mois de décembre qui présente la plus grande proportion de thèses soutenues de 2005 à 2018, avec presque 30 % des soutenances. Cette concentration en fin d'année reflète probablement l'organisation des calendriers universitaires et les rythmes institutionnels, avec un creux estival clairement marqué.

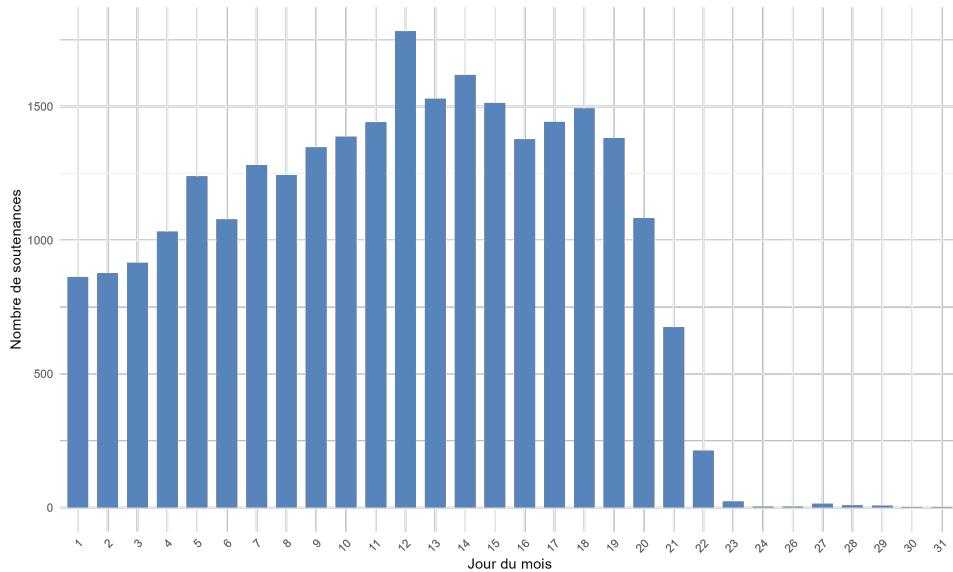


Figure 6: Somme des soutenances par jour en décembre de 2005 à 2018

Dans la figure, nous pouvons noter que la majorité des soutenances de thèses s'effectuent du 1er au 22 décembre, avec une valeur extrême le 12 décembre, avec plus de 1750 thèses.

4 Discussion

4.1 Annonce de la structure de la discussion

Cette section est divisée en plusieurs sous-sections, chacune présentant un aspect des résultats obtenus.

4.2 Interprétations

4.2.1 Langue d'écriture

On peut soutenir que si à partir des années 2000, la langue d'écriture la plus utilisée devient l'anglais, c'est parce qu'il y a de plus en plus de carrières internationales. Par ailleurs, de nombreux étudiants, notamment en école de commerce, école d'ingénieur et instituts d'études politiques, doivent pour obtenir leur diplôme, valider un semestre de mobilité internationale durant lequel la langue requise est majoritairement l'anglais. Cette expérience contribue à familiariser les doctorants avec l'usage de l'anglais dans un contexte académique et professionnel.

En outre, dans de nombreuses institutions, on observe l'introduction croissante de cours dispensés en anglais, dans des ambitions de professionnalisation des étudiants. Ces pratiques favorisent l'acquisition d'un vocabulaire technique et scientifique en anglais, qui renforce dès lors l'usage de cette langue dans la rédaction des thèses.

D'autre part, ceci traduit également l'influence des politiques publiques faisant écho à la loi Fioraso de 2013, des attentes des comités de financement et des revues scientifiques. Il est clair que ces facteurs institutionnels exercent une pression qui contribue à l'adoption progressive de l'anglais comme langue dominante.

Cependant cette tendance n'est pas homogène, il a été montré que les sciences exactes ont une plus grande tendance à l'internationalisation que les sciences humaines. Prenons l'exemple, de l'ingénierie informatique, dans ce secteur, l'anglais est quasi systématiquement la langue utilisée pour la rédaction du code, sans oublier que sur les forums et blogs cette langue est universelle.

Enfin, le choix de la langue semble également relever d'un effet générationnel, les jeunes doctorants étant plus enclins à rédiger en anglais. Ce qui peut s'expliquer par l'ère numérique dans laquelle ils se situent. Nous pouvons aujourd'hui parler de démocratie digitale et ces étudiants sont quotidiennement exposés à l'anglais à travers Internet, les réseaux sociaux, les contenus de divertissement et les œuvres de la pop culture, souvent disponibles d'abord en anglais ou avec des traductions majoritaires dans cette langue. Par conséquent, l'exposition permanente contribue à familiariser les futurs doctorants avec l'anglais et renforce leur aisance dans son usage académique. Dès lors, nous pouvons soutenir que l'anglais s'impose non seulement comme vecteur de diffusion scientifique, mais aussi comme langue d'usage quotidienne et outil d'insertion dans un environnement académique internationalisé, consolidant sa prééminence dans les pratiques d'écriture actuelles des thèses en France.

4.2.2 Date de soutenance

L'analyse des soutenances met en évidence une concentration marquée en fin d'année, avec un pic en décembre, et une activité très faible durant l'été. Par ailleurs, cette forte concentration des soutenances en fin d'année peut s'expliquer par le calendrier universitaire ainsi que les contraintes administratives liées aux concours et aux commissions d'évaluation. De plus, la forte concentration autour de certaines dates, comme le 12 décembre, suggère également une synchronisation des soutenances pour répondre aux échéances et optimiser la présence des membres des jurys.

Cette répartition n'est pas uniforme et varie selon les disciplines et les institutions. Certains secteurs privilégient des soutenances en début d'année ou à d'autres périodes, tandis que d'autres suivent les rythmes traditionnels. Elle reflète aussi des pratiques historiques et culturelles propres aux universités françaises et des logiques organisationnelles visant à faciliter le déroulement des jurys et la gestion des thèses. Enfin, cette concentration a des conséquences concrètes : elle entraîne une charge de travail importante pour les doctorants et les enseignants-chercheurs en fin d'année, mais permet aussi de regrouper les soutenances dans des périodes plus calmes pour mieux coordonner les jurys et les évaluations.

En somme, on peut aussi l'expliquer par le fait que la date des concours est majoritairement en mars et que, pour candidater, il faut avoir effectué sa soutenance de thèse avant la mi-décembre.

4.3 Conclusion

Dans ce travail, nous nous sommes basés sur le contenu de la base de données. Nous avons pu constater qu'il y avait des erreurs de saisie lors du traitement des données manquantes, des homonymes et des outliers (pour de nombreuses données, comme l'auteur, le titre, le directeur de thèse et le genre). Dès lors, il est fort probable qu'il y ait aussi des erreurs de saisie pour la langue (thèses écrites en français mais inscrites en anglais et inversement) et que les données ne soient pas le reflet fidèle de la réalité de l'utilisation des langues française et/ou anglaise en France. La fiabilité des données inscrit ainsi les limites de ce travail.

Il serait intéressant de se demander dans quelle mesure l'introduction de cours en anglais dans les universités, comme l'espérait la ministre Geneviève Fioraso, a contribué à réduire les inégalités. Une approche possible consisterait à analyser l'ethnie des doctorants et/ou des encadrants pour prédire les effets de cette politique. En outre, il serait pertinent d'examiner si certaines disciplines présentent un plus grand nombre de thèses rédigées en anglais. On pourrait également se demander si le directeur de thèse influence la langue d'écriture, ainsi que d'autres facteurs comme le genre ou l'origine ethnique.

5 Références

1. AFP, Le Monde Avec. (2013). Anglais à l'université : Fioraso dénonce une formidable hypocrisie . *Le Monde.fr* [en ligne], 21 mai 2013. Disponible à l'adresse : https://www.lemonde.fr/societe/article/2013/05/21/anglais-a-l-universite-fioraso-3410228_3224.html
2. HÉRAN, François. (2013). L'anglais hors la loi ? Enquête sur les langues de recherche et d'enseignement en France. *Population & Sociétés* [en ligne], 3 juin 2013, Vol. N° 501, n° 6, pp. 1-4. DOI : 10.3917/popsoc.501.0001
3. O'CONNELL, Anne-Marie et CHAPLIER, Claire. (2021). Les langues de spécialité dans l'enseignement supérieur en France : un exemple de littératie enseignante dans le domaine de l'anglais des sciences. *Éducation & Didactique* [en ligne], 30 juin 2021, Vol. 15, n° 15-2, pp. 85-102. DOI : 10.4000/educationdidactique.8729