Évaluation de la Fiabilité des Données des Thèses en France : Analyse des Incohérences et Évolution des Pratiques Linguistiques

Axelle Le Poul

November 14, 2024

Résumé

Dans ce rapport, nous avons analysé les données des thèses soutenues et à soutenir en France, en utilisant une base de données accessible en ligne, recensant des thèses jusqu'à 1970 (https://theses.fr/?domaine=theses). L'objectif principal était d'évaluer la fiabilité de ces données. Nous avons observé plusieurs incohérences, notamment au niveau des dates de soutenance et des langues utilisées pour présenter les thèses. Pour étudier ces problèmes, nous avons utilisé des méthodes statistiques et des visualisations graphiques afin de mieux comprendre les tendances et les erreurs présentes dans la base. Les résultats ont montré un manque de fiabilité des données, principalement dû à leur ancienneté, ce qui a entraîné des erreurs dans les informations de temporalité et de langue. En comparant les données sur différentes périodes, nous avons pu identifier une évolution des pratiques de collecte et une amélioration progressive de la qualité des données au fil du temps. Ainsi, ce rapport porte avant tout sur la fiabilité des données proposées, soulignant l'importance d'une gestion rigoureuse et d'une mise à jour régulière des bases de données, particulièrement lorsque celles-ci couvrent plusieurs décennies.

Contents

List of Figures

Introduction

Lors de l'analyse de la base de données des thèses en France, plusieurs défauts ont été identifiés. Cette base, qui recense les thèses soutenues depuis 1970, comporte de nombreuses données manquantes ainsi que des incohérences, notamment au niveau des dates de soutenance. Par exemple, entre 1984 et 2018, nous avons constaté une surreprésentation de certaines dates spécifiques, suggérant un manque de rigueur dans la saisie des données. Ce phénomène de "dates erronées" représente un exemple typique de lacune dans la gestion des bases de données de longue durée, qui peut avoir des conséquences sur leur fiabilité et leur exploitation.

Lacunes dans les connaissances Le problème de la fiabilité des données dans les bases de thèses n'a pas été largement documenté dans la littérature. Bien que des recherches existent sur la gestion des données dans les bases de données académiques, les implications des incohérences dans les dates de soutenance n'ont pas été suffisamment explorées. Ce manque de documentation et d'analyse approfondie constitue une lacune importante dans les connaissances. Il est utile de comprendre l'impact de ces erreurs sur les recherches futures et d'identifier les meilleures pratiques pour améliorer la collecte et la gestion de telles bases de données.

Questions de recherche À partir de ces constats, plusieurs questions ont émergé :

- Quelles sont les causes sous-jacentes des incohérences et des données manquantes dans cette base ?
- Comment peut-on interpréter et exploiter ces données de manière fiable malgré leurs lacunes ?
- Quels sont les critères à prendre en compte pour structurer et organiser cette base de données de manière plus cohérente ?
- Comment la répartition des langues a-t-elle évoluée au cours des 40 dernières années ?

Méthodes Pour répondre à ces questions, nous avons utilisé le logiciel R Studio, un outil adapté à l'analyse de données massives. Nous avons filtré les données pour éliminer celles qui étaient manquantes ou incohérentes, puis nous avons effectué des analyses statistiques pour identifier les tendances et les anomalies. En particulier, nous avons comparé les dates de soutenance au fil des années, ce qui nous a permis de repérer des motifs récurrents d'incohérences, notamment la surreprésentation de certaines dates. Nous avons aussi étudié l'évolution de la complétude des données au

fil du temps, et nous avons séparé les données en intervalles temporels pour mieux comprendre leur organisation.

Hypothèses À partir des observations réalisées, nous avons formulé les hypothèses suivantes :

1. De nombreuses données manquantes sont associées à l'ancienneté des données. En effet, les thèses les plus anciennes sont souvent moins bien renseignées, en raison des pratiques de collecte de l'époque.

2. Les dates de soutenance indiquées au 1er janvier sont souvent erronées. Cela pourrait être dû à des erreurs lors de l'enregistrement des dates, particulièrement pour les thèses datant de plusieurs décennies.

3. Nous allons voir une forte augmentation de l'utilisation des langues étrangères, en particulier l'anglais, au cours des décennies.

Chapter 1

Méthodes et Données

1.1 Source des données

Les données proviennent d'un fichier CSV, "PhD.dataset.csv", que nous avons chargé dans R à l'aide de la fonction read.csv(). Ce fichier contient des informations relatives à des thèses de doctorat, notamment la date de soutenance, le sujet, les auteurs, les directeurs de thèse, et d'autres détails qui sont souvent utilisés dans ce type d'analyse.

Avant de procéder à l'analyse proprement dite, il est important de s'assurer de la fiabilité des données. Dans le cas présent, nous avons identifié des données manquantes ou incorrectes dans le jeu de données. Par exemple, certaines valeurs sont marquées comme "NA", "unknown" ou "na", et nous avons pris soin de les traiter en les remplaçant par des NA dans R à l'aide de la fonction mutate(across()). La fiabilité des données dépend également de la source d'origine et de la manière dont elles ont été collectées. Si les données proviennent d'une source fiable, comme une base de données institutionnelle ou un registre académique, elles seront probablement relativement solides, ce qui est le cas de cette base de données qui est proposée par le ministère de l'enseignement supérieur et de la recherche. Cependant, malgré sa source de nombreux problèmes subsitent.

1.2 Gestion des outliers

Les outliers, ou valeurs aberrantes, sont des points de données qui se distinguent de manière significative des autres dans un jeu de données. Dans le cas des thèses soutenues, un exemple d'outlier récurrent est celui où la date de soutenance est systématiquement fixée au 1er janvier 2005, ce qui n'a pas de fondement réel et ne correspond pas aux dates de soutenance effectives. Ce phénomène peut résulter d'une erreur de saisie ou d'un défaut dans le système informatique qui enregistre ces dates.

Jusqu'en 2005, il est fréquent de trouver cette date générique pour la majorité des thèses, ce qui crée une distorsion dans l'analyse des données. En effet, si l'on cherche à comprendre les tendances des soutenances de thèses au fil des années,

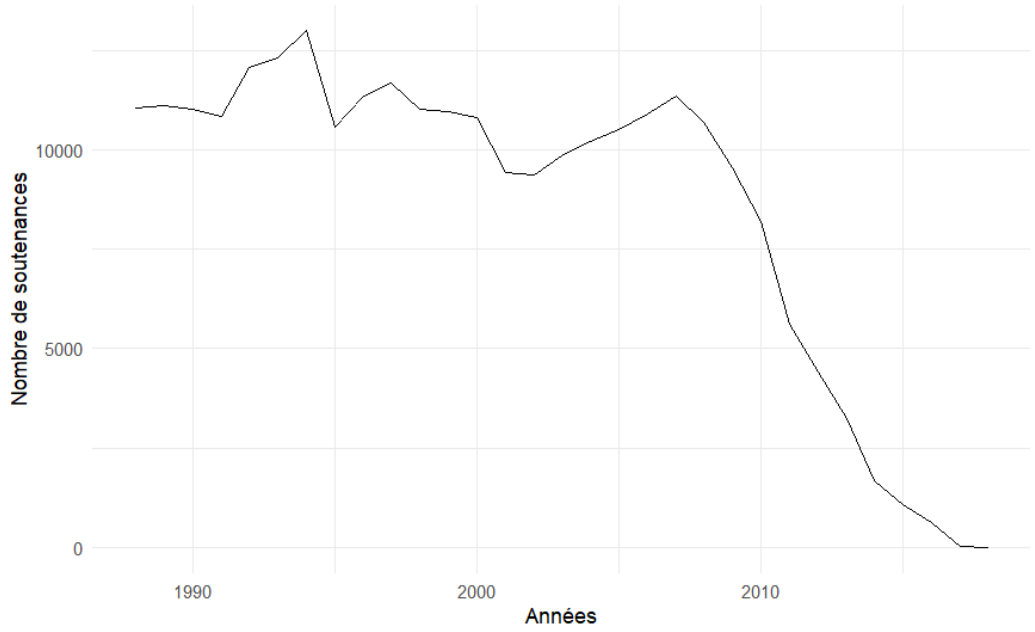

Figure 1.1: Répartition des thèses le 1er janvier par année

ces dates erronées faussent totalement l'analyse. Au lieu de refléter une répartition réelle des soutenances, elles donnent l'impression qu'une majorité de thèses ont été soutenues en ce jour unique, ce qui est évidemment irréaliste.

Ainsi, l'importance de gérer ces outliers est de permettre une analyse plus précise et fiable. Cela passe par des étapes essentielles telles que l'identification des dates aberrantes, leur correction si possible, ou leur suppression dans le cas où elles ne peuvent pas être corrigées. Nous avons choisi de supprimer les soutenances ayant lieu le 1er janvier dans notre base de données.

1.3 Techniques et outils utilisés

Passons maintenant aux méthodes et techniques utilisées dans notre code. Le prétraitement des données est une étape essentielle avant de commencer toute analyse. Nous avons d'abord chargé les données et effectué un nettoyage en remplaçant les valeurs manquantes ou incorrectes par des NA. Ensuite, nous avons calculé le pourcentage de données manquantes par colonne, ce qui nous permet de savoir où se trouvent les lacunes dans le jeu de données. Cette étape est cruciale car elle aide à identifier si une variable a trop de valeurs manquantes, ce qui pourrait justifier son exclusion de l'analyse. Nous avons utilisé des heatmaps pour visualiser la distribution des données manquantes dans notre jeu de données.

Les heatmaps permettent de voir rapidement les zones où les données sont absentes, ce qui aide à comprendre la structure des données.

Nous avons transformé les dates de soutenance en objets de type Date en utilisant la fonction dmy() de la bibliothèque lubridate. Cela permet de traiter les dates de manière cohérente dans toute l'analyse. En filtrant les soutenances entre 1984 et 2018, nous avons réduit le jeu de données à une période d'intérêt spécifique.

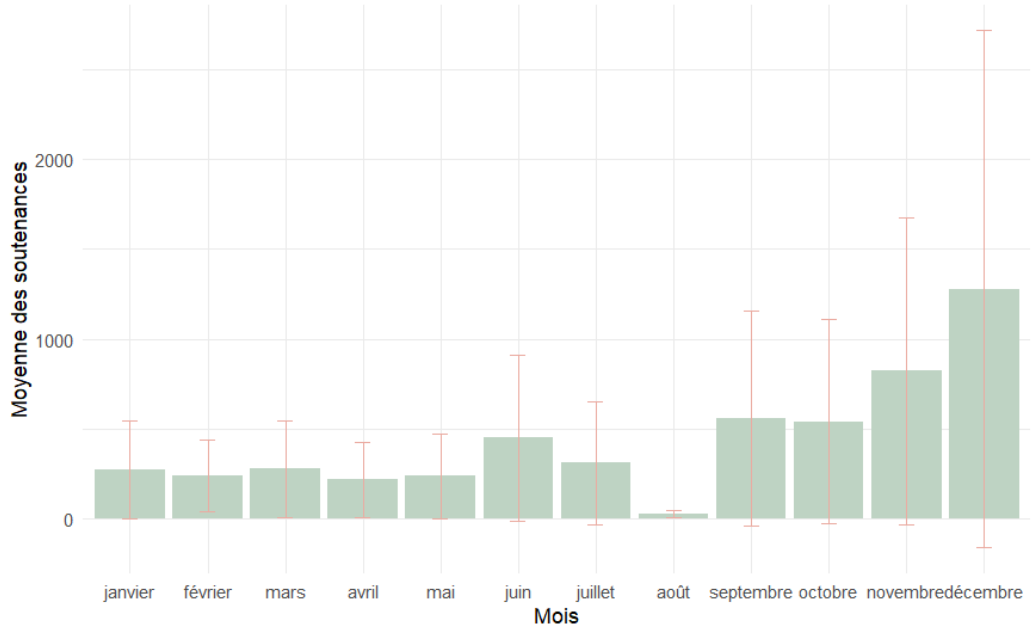

Figure 1.2: Moyenne des soutenances par mois sans le 1er janvier (1988-2018)

L'analyse temporelle est souvent cruciale pour comprendre les tendances dans le temps, par exemple, si la distribution des soutenances est stable ou s'il y a des pics ou des baisses à des moments spécifiques. Nous avons calculé le nombre de soutenances par mois et par année. Cette approche permet d'obtenir une vision claire de la répartition des soutenances au fil du temps. Nous avons utilisé la fonction 'groupby()' pour regrouper les données par mois et année, puis 'summarise()' pour compter le nombre de soutenances. Cela a permis de visualiser la fréquence des soutenances et de détecter des tendances saisonnières ou des variations au cours des années. Nous avons créé une variable qui catégorise les thèses en fonction de leur langue (français, anglais, bilingue, etc.). Cette étape permet d'analyser la répartition des soutenances selon la langue de rédaction des thèses et d'observer si certaines langues sont plus dominantes que d'autres au fil des années. Nous avons également regroupé les années en périodes de deux ans, ce qui peut aider à observer des tendances sur des périodes plus larges. Une fois que nous avons calculé le nombre total de soutenances pour chaque mois et année, nous avons calculé la proportion de soutenances pour chaque mois, ce qui permet de visualiser les tendances en pourcentage. Nous avons également utilisé des graphiques pour afficher ces proportions, ce qui aide à mieux comprendre l'évolution des soutenances au fil du temps.

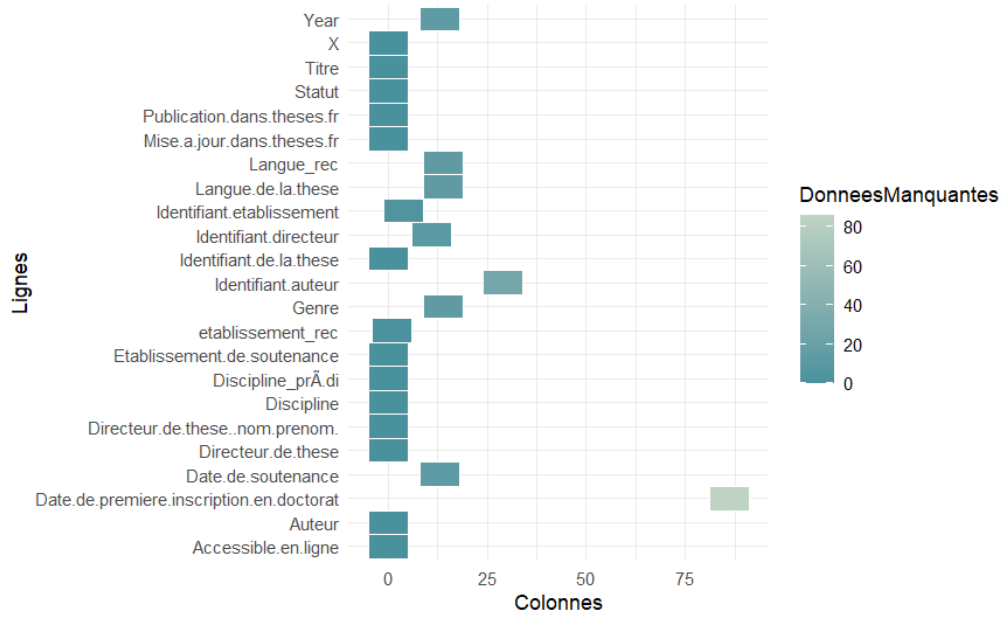

Figure 1.3: Données manquantes (1)

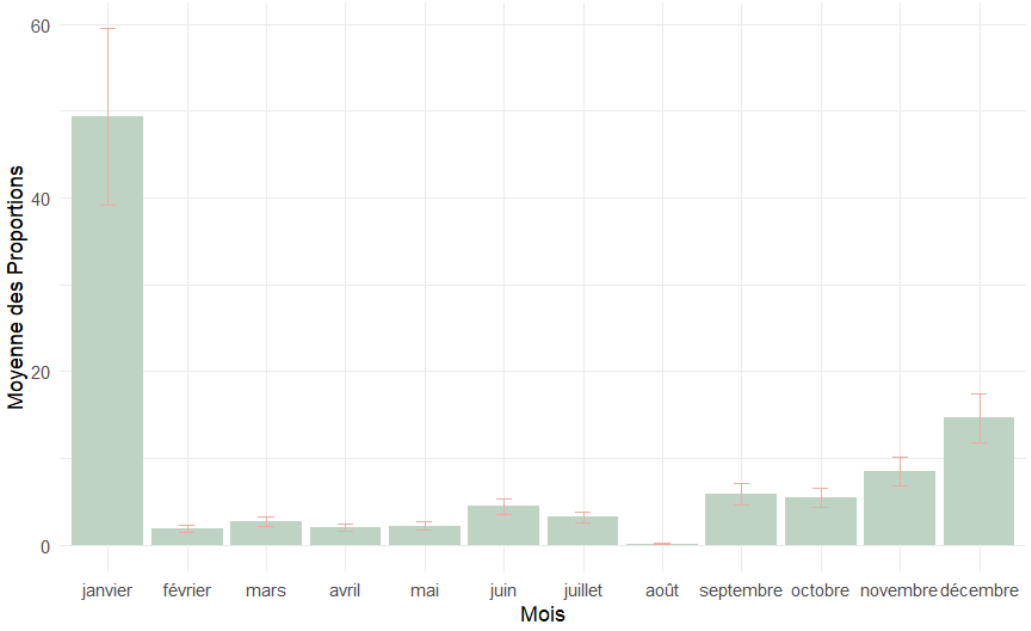

Figure 1.4: Moyenne des proportions de soutenances par mois (1988-2018)

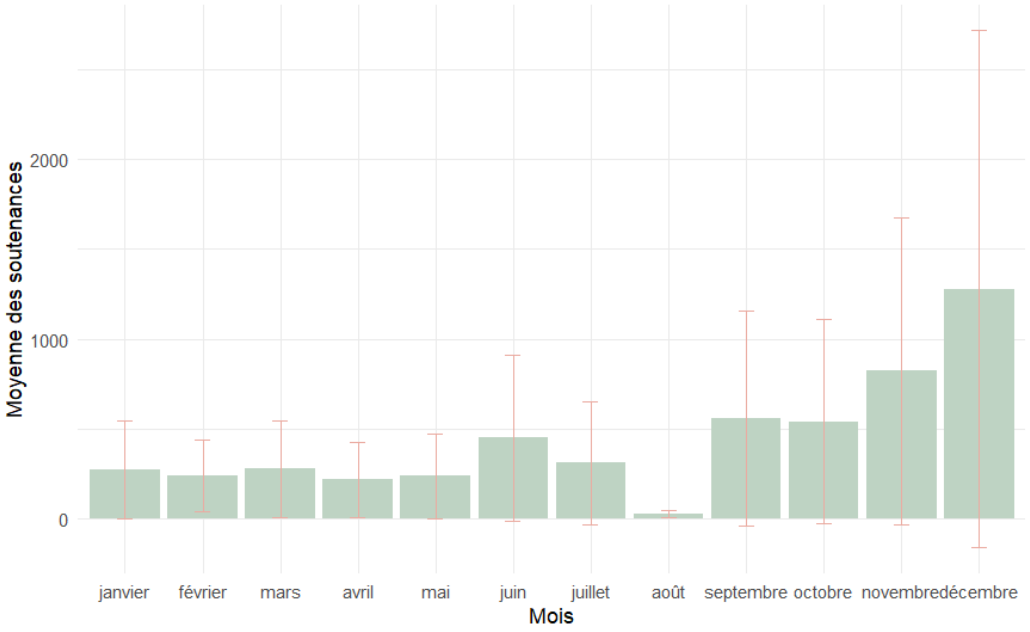

Figure 1.5: Moyenne des proportions de soutenances par mois (1988-2018) sans le premier janvier

1.4 Conclusion

Les méthodes que nous avons utilisées permettent une analyse approfondie des tendances dans les soutenances de thèses en fonction de diverses variables (mois, année, langue, etc.). Le prétraitement des données, y compris la gestion des valeurs manquantes, est essentiel pour garantir que les résultats de l'analyse soient fiables et pertinents. Les techniques de visualisation, comme les heatmaps et les graphiques, fournissent des aperçus clairs qui facilitent l'interprétation des résultats.

Chapter 2

Résultats

Introduction aux résultats

Cette analyse présente l'évolution de l'utilisation des différentes langues pour la rédaction des thèses en France entre 1984 et 2021. À travers plusieurs visualisations, nous mettons en lumière l'essor des langues étrangères, notamment l'anglais, et leur impact croissant sur la rédaction des thèses doctorales au fil des décennies.

2.1 Évolution des langues utilisées dans les thèses

Nous avons ainsi obtenu une visualisation de l'utilisation des différentes langues pour la rédaction des thèses au cours des années. L'exemple le plus parlant est le graphique suivant :

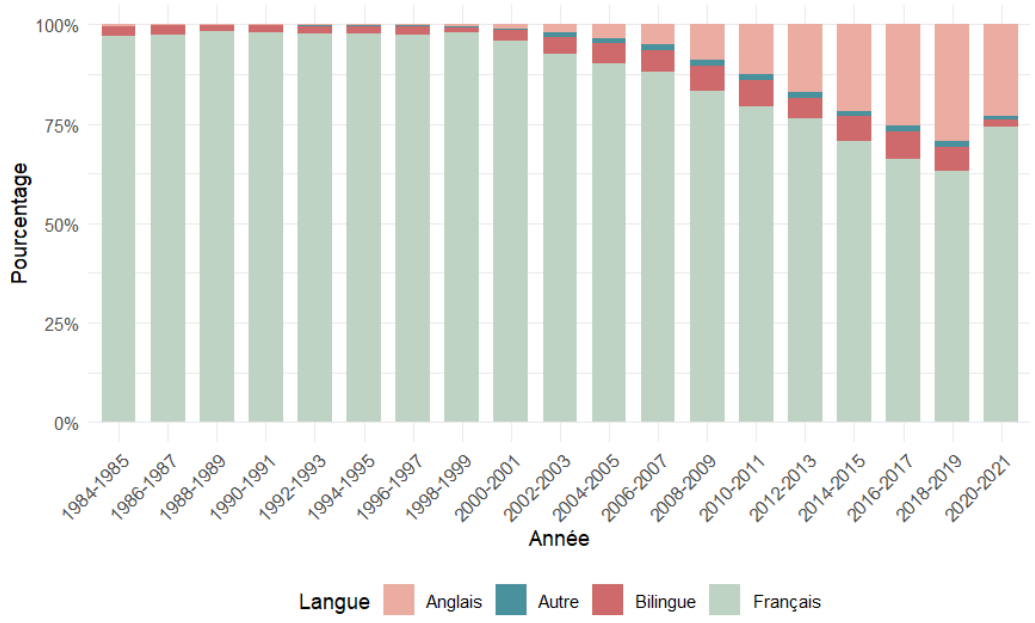

Figure 2.1: Répartition des thèses par langue chaque année.

Ce graphique montre l'évolution de la répartition des thèses par langue chaque année, de 1984 à 2021. Pour plus de lisibilité, les années ont été regroupées par paires. Nous pouvons observer sur ce graphique que jusqu'en 1999, les langues étrangères sont presque absentes dans les thèses en France, avec les thèses bilingues représentant environ 2% des thèses totales. À partir de 2000, de nouvelles langues (notamment l'anglais et d'autres langues) commencent à apparaître, mais elles ne représentent qu'une part infime des thèses (environ 5% avec les bilingues). Cependant, les présentations "non en français" évoluent fortement à partir de cette date, atteignant un pic sur l'intervalle 2018-2019 avec 37% environ des thèses dans une autre langue que le français. On remarque donc une forte évolution de l'utilisation des langues étrangères, particulièrement l'anglais (environ 30% des thèses entre 2018-2019) au cours des 20 dernières années.

Une autre manière de visualiser ces données peut se faire avec une heatmap :

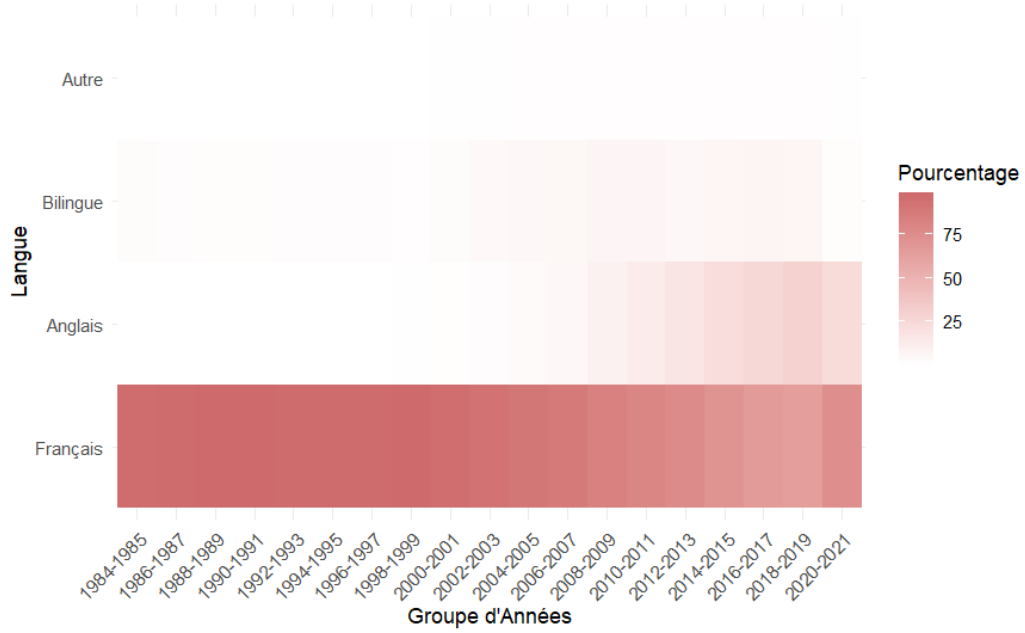

Figure 2.2: Répartition des thèses par langue et année

Cette dernière montre la même chose que le graphique précédent mais d'une autre manière. On peut ainsi voir l'augmentation des langues étrangères et particulièrement de l'anglais à partir des années 200.

Nous avons ensuite réalisé un graphique illustrant l'évolution des la présence de chaque langue dans les thèses française entre 1984 et 2021 :

Sur ce graphique, nous pouvons voir que l'augmentation de la présence des langues étrangères dans les thèses commence à partir du début des années 2000. A partir de 2007 l'anglais commence à réellement devenir la langue étrangère majeure utilisée pour les thèses (avant aussi mais bilingue). On a alors une très forte augmentation de la proportion de l'anglais entre 2007 et 2019, passant de 5 % des thèses à plus de 25%. On peut proportionnellement voir le français baisser drastiquement entre 1999 et 2019 passant de 98% des thèses à environ 63%. On voit également une baisse de l'anglais, et des langues étrangères, à partir de 2020 et donc une augmentation du français que l'on peut associer à la crise du Covid-19 (moins

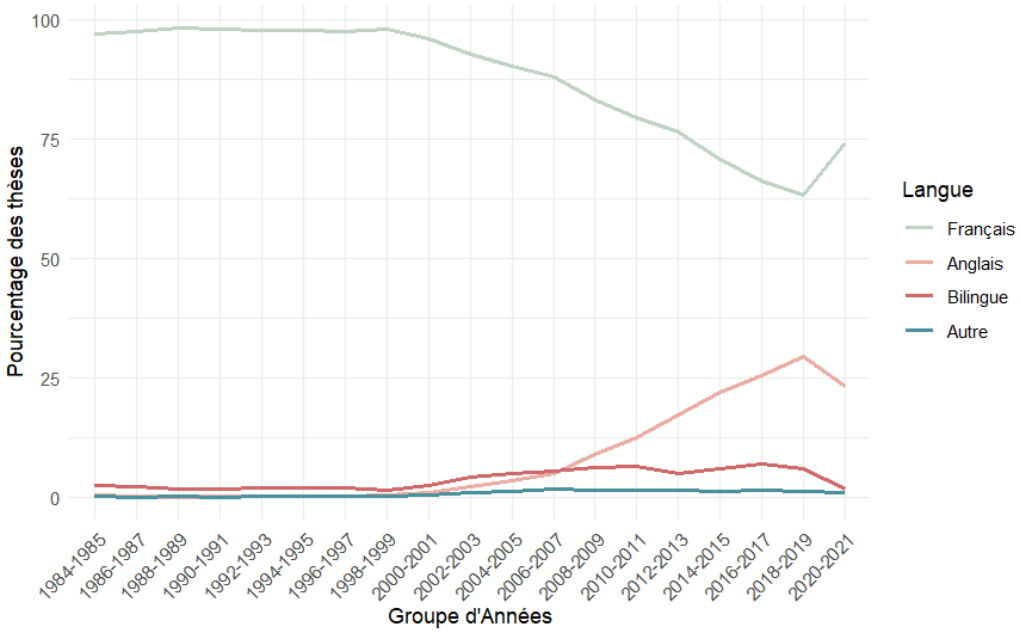

Figure 2.3: Évolution de la proportion des thèses par langue

d'étudiants étrangers).

Comme nous avons pu le voir dans les précédents graphiques, l'anglais est la langue étrangère la plus présente dans les thèses françaises. Nous avons donc réalisé un graphique portant particulièrement sur l'anglais :

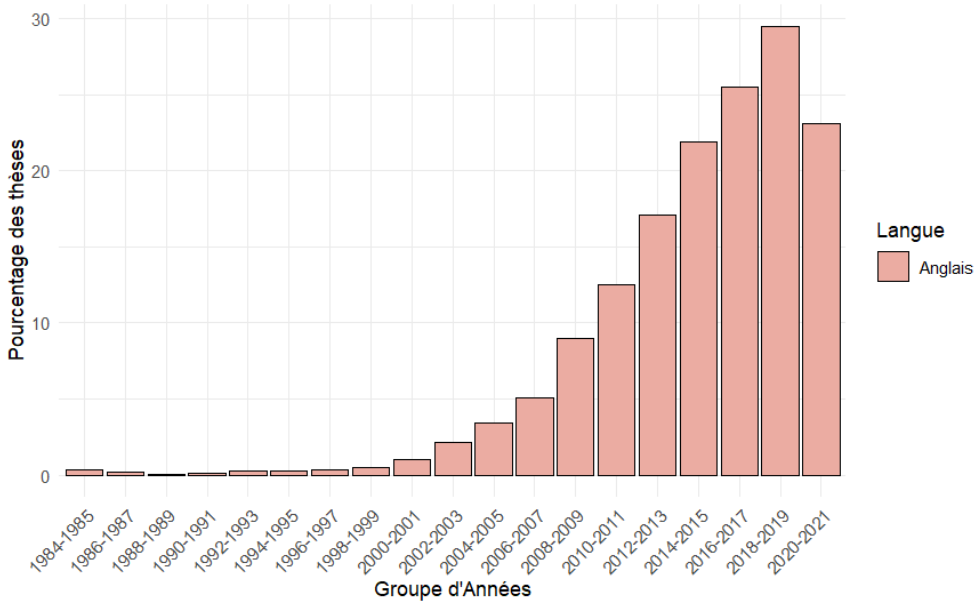

Figure 2.4: Évolution de la proportion de thèses en Anglais

Ce graphique montre la proportion de l'anglais dans les thèses en 1984 et 2021. On peut y voir plus en détail la forte évolution de la présence de la langue entre 2000 et 2019. C'est une façon de voir les résultats précédents de manière plus détaillée et précise.

Pour finir, nous avons réalisé un graphique circulaire représentant l'entièreté des thèses produites entre 1984 et 2021 et la langue utilisée :

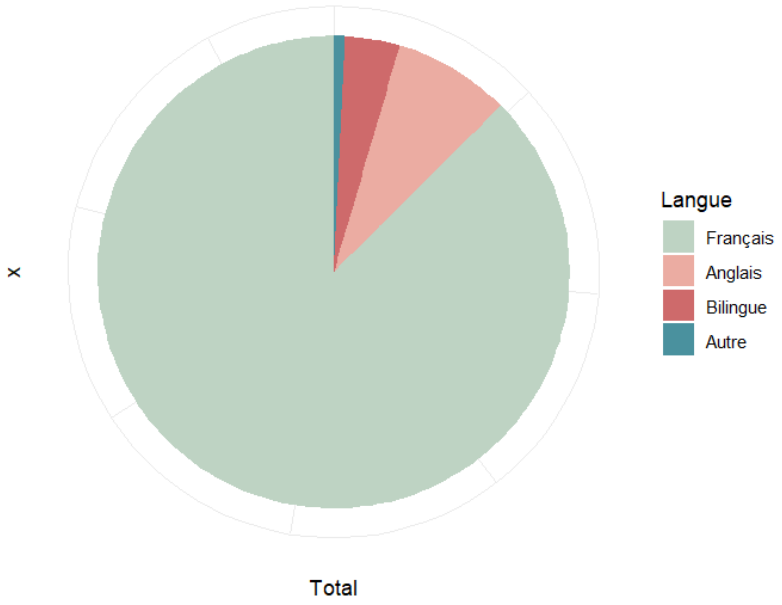

Figure 2.5: Répartition des thèses par langue

Nous pouvons voir sur ce graphique la proportion des langues utilisées pour l'ensemble des thèses de 1984 à 2021. On peut ainsi voir que la majorité des thèses sont an français, environ 88% de l'ensemble des thèses. L'anglais est très présent, environ 7,5% des thèses. On a ensuite les thèses bilingues avec 3,5% des thèses. Et enfin les autres langues avec environ 1% des thèses.

Chapter 3

Discussion

Structure de la discussion

Pour analyser les résultats obtenus, nous suivrons un plan chronologique. Nous aborderons d'abord la période antérieure à l'an 2000, puis nous examinerons la période de 2000 à 2007. Ensuite, nous analyserons l'importance croissante de l'anglais entre 2007 et 2019, avant de traiter l'impact de la crise du Covid-19 à partir de 2020.

3.1 Domination de la langue française (1984-2000)

Les résultats obtenus nous montrent une domination du français jusque dans les années 2000. En effet, cette prédominance quasi exclusive du français dans les thèses françaises avant 2000, reflet d'un milieu académique peu internationalisé. Cette forte présence de l'anglas s'explique aussi par des raisons historiques et culturelles variées comme : le rôle historique de la France dans les sciences et les arts, le rôle des institutions françaises et francophones (CNRS, la Sorbonne...) ou encore la production et diffusion des revues scientifiques en français.

3.2 Apparition de l'anglais (2000-2007)

Cependant, après 2000 on voit une apparition progressive de l'anglais dans la rédaction des thèses en France. Cela montre une ouverture accrue à l'international. Cette évolution peut s'expliquer par plusieurs facteurs. Tout d'abord, la mondialisation de la recherche scientifique a favorisé l'utilisation de l'anglais, qui est devenu la langue de référence pour les publications académiques et les collaborations internationales. En outre, de plus en plus d'étudiants étrangers choisissent la France pour leurs études doctorales, ce qui a contribué à l'augmentation des thèses rédigées dans des langues autres que le français.

3.3 Forte croissance de l'utilisation de l'anglais (2007-2019)

Entre 2007 et 2019 la forte croissance de l'anglais montre que l'anglais s'est imposé comme la langue dominante pour la recherche internationale. L'anglais devient progressivement la langue prédominante dans les domaines scientifiques et académiques, en particulier dans les sciences naturelles et les sciences sociales, où il est perçu comme une langue internationale facilitant l'échange de connaissances.

3.4 Impact de la COVID-19 sur les tendances linguistiques en 2020

Durant cette période, nous avons pu noter une réduction notable des thèses en langues étrangères, associée à une reprise du français. En effet, la crise du Covid-19 a conduit à une forte diminution de la mobilité internationale et a donc entraîné moins de thèses dans des langues étrangères.

Conclusion

Dans ce travail, nous avons commencé par explorer la fiabilité des données relatives aux thèses soutenues en France depuis 1970, en utilisant une base de données en ligne centralisant ces informations. Nous avons cherché à évaluer la cohérence de ces données, en particulier dans les aspects liés aux dates de soutenance et aux langues de rédaction des thèses. En analysant les données via des méthodes statistiques et des visualisations graphiques, plusieurs incohérences significatives ont été détectées, telles que des erreurs dans les dates de soutenance et une sous-représentation des langues étrangères.

Nos résultats montrent que l'ancienneté des données est un facteur majeur de ces irrégularités, ce qui a engendré des erreurs de temporalité et de langue dans la base. En suivant un découpage chronologique, nous avons observé une évolution des pratiques de collecte des données et une amélioration progressive de leur qualité avec le temps. L'analyse met en lumière l'importance de la gestion rigoureuse et de la mise à jour régulière des bases de données académiques pour garantir leur fiabilité, particulièrement pour celles couvrant de longues périodes. Nos résultats ouvrent ainsi la voie à de futures études visant à optimiser les systèmes de gestion des données académiques, afin d'améliorer leur utilisation dans la recherche et l'analyse à grande échelle.

Pour finir, nous nous sommes intéressés aux résultats, après filtration de la base, de la part des langues étrangères dans la rédaction des thèses en France. Ces données connaissent donc des limites pour les raisons citées plus tôt. Cependant, nous avons fait en sorte de réduire au maximum ces incohérences. Notre analyse montre une transition marquée des thèses françaises vers l'anglais, symbole d'une internationalisation croissante de la recherche en France. Alors que le français dominait jusqu'en 2000, l'anglais a pris de l'importance, atteignant près de 30 % des thèses en 2019, principalement en réponse à la mondialisation scientifique. La crise du Covid-19 a toutefois temporairement renforcé le recours au français, soulignant l'impact des échanges internationaux sur les choix linguistiques académiques.

Une analyse future intéressante proposée par la base de données que nous avons utilisée est la prépondérance de certains directeurs de thèse dans la recherche française. En effet, après s'être intéressé à certaines données nous avons pu calculer que le pourcentage de thèses encadrées par 1% des directeurs est de 10.11447 %. Nous pouvons donc imaginer étudier l'importance de certains directeurs en fonction de différents facteurs, notamment leur genre.

Références

Ammon, U., McConnell, G. D. (2002). English as an Academic Language in Europe: A Survey of its Use in Teaching. Peter Lang: Cette étude aborde l'anglais en tant que langue académique en Europe, expliquant des phénomènes similaires d'internationalisation dans d'autres pays européens.

Hamel, R. E. (2007). The dominance of English in the international scientific periodical literature and the future of language use in science. AILA Review, 20(1), 53-71. : Ce travail explore l'anglais comme langue scientifique globale et les implications de cette dominance sur les autres langues dans les travaux de recherche.