Exploration diachronique des marqueurs d'intensité en anglais

Projet de linguistique outillée et traitements statistiques

M2 TAL - Université Paris Nanterre - 2021-2022

Professeur : Guillaume Desagulier

Étudiants : Santiago Herrera Yanez / Yagmur Ozturk
Numéros d'étudiant : 22000585 (P3) / 22001914 (P3)

Introduction

Des expressions anglaises comme *awfully glad* ou *terribly good* ont toujours suscité notre intérêt linguistique, car ce sont des constructions formées par un marqueur d'intensité et un adjectif qui présentent des propriétés sémantiques intéressantes. D'un point de vue lexical, la polarité de mots isolés est parfois ambiguë ou contradictoire : comme dans les exemples précédents, la polarité de l'adverbe est souvent opposée à celle de l'adjectif. On trouve aussi des cas contraire, ex. *amazingly bad*, où la valeur de la polarité est inversée. En revanche, dans les deux cas, le marqueur d'intensité ne fait que pointer dans la même direction que l'adjectif. Évidemment, ces exemples coexistent avec d'autres comme *terribly ill* ou *extremely precarious*, dans lesquels la polarité des mots est similaire. En définitive, c'est l'adjectif qui co-occurre qui marque la valeur de la polarité.

Pour approfondir cette problématique, on a décidé de mener une étude diachronique de 8 marqueurs d'intensité de l'anglais américain à partir du *Corpus of Historical American English* (COHA)[1], avec le but d'explorer les différences d'usages de ce type de construction au cours des différentes décennies comprises dans le corpus et observer, ainsi, leur évolution. Cette exploration a été faite à partir de l'hypothèse linguistique de départ suivante : l'usage du pair marquer d'intensité-adjectif des polarités contraires est d'autant plus répandu qu'on se rapproche du présent. Cette hypothèse pourrait être affirmée ou non grâce à une étude diachronique. Bien que l'on considère le corpus COHA comme assez représentatif de l'anglais américain, on ne cherche à mettre à l'épreuve cette hypothèse que dans les limites de ce corpus.

Pour cela, on a développé deux approches complémentaires. Dans un premier temps, on a travaillé sur la polarité des adjectifs co-occurrents des marqueurs d'intensité. Dans un deuxième temps, on a réalisé des observations sur les fréquences de ces adjectifs. Dans les deux cas, on a cherché des relations statistiquement significatives entre les décennies, ainsi que des informations importantes sur le pair marqueur d'intensité-adjectif. Cette méthode d'analyse a des limites évidentes, comme l'impossibilité d'analyser son usage pragmatique, notamment, l'ironie. De plus, il devient nécessaire de répondre d'autres questions sur le processus, par exemple, comment récupérer la polarité des adjectifs.

Ces questions seront abordées tout au long de ce travail, dont chaque section reflète une partie de l'exploration et de l'évaluation des données. Premièrement, on décrit brièvement le corpus pour, ensuite, expliquer comment on a créé et constitué les données sur lesquelles on a travaillé. À partir des plusieurs méthodes exploratoires multivariées, on explore les données récoltées afin de pouvoir évaluer, dans une première instance, notre hypothèse, et pouvoir l'ajuster face aux nouvelles informations obtenues. Enfin, on la teste statistiquement. Pour conclure ce travail, on a fini par une réflexion sur la démarche, les limitations et sur d'autres pistes à explorer. Tout le travail a été réalisé par des outils et scripts écrits en R que l'on joint à ce document[2].

[1] Davies, Mark. (2010) *The Corpus of Historical American English (COHA)*. Available online at https://www.english-corpora.org/coha/.

[2] Les scripts sont divisés par chaque section de ce travail. Ils sont de nature exploratoire.

Description du corpus et constitution de données

Le corpus COHA est un corpus de l'anglais américain contenant, dans sa version actuelle, 475 milliards de mots, récupérés de plus de 100 000 textes, qui couvrent les décennies comprises entre 1820 et 2010. La version du corpus qu'on a utilisé (la précédente), en revanche, va de 1810 à 2000. Bien qu'il s'agisse d'une version moins équilibrée que l'actuelle, comme dans le cas de la décennie de 1810, c'est un corpus assez homogène à l'intérieur de chaque décennie. En effet, la totalité du corpus est équitablement répartie en genre (fiction, non fiction), composé de textes de genre varié, des journaux, des romans, des livres non fictionnel, etc. Pour tout cela, sans pourtant nier que la version actuelle aurait été une ressource plus intéressante à exploiter (l'ajout de sous-titres et de la décennie 2010 sont un plus), le corpus utilisé nous permet d'avoir un aperçu approfondi de la variation de l'anglais américain.

Plus spécifiquement, nous avons eu accès au corpus en format texte brute, tokenisé et normalisé, sans aucun autre traitement. Le fichier texte en question, contenant le corpus, est composé de 20 lignes, une par décennie. Comme il s'agit d'un gros corpus, son traitement a été fait nécessairement en flux. Dans un premier temps, on avait décidé d'exploiter non seulement les tokens mais aussi les phrases. Néanmoins, la segmentation en phrases s'est avérée difficile et pas fiable, notamment dans les premières décennies, dans des textes dramatiques et poétiques.

Afin d'explorer notre objet d'étude et d'obtenir les occurrences des marqueurs d'intensité. On a choisi de travailler depuis le départ avec les 8 marqueurs d'intensité suivants, dans le but de rendre gérables les données à analyser : *extremely*, *terribly*, *awfully*, *dreadfully*, *tremendously*, *amazingly*, *insanely* et *colossally.* Ce sont tous des adverbes, qui alternent leur polarité, et qui sont susceptibles d'apparaître dans les contextes et usages décrits dans l' introduction,

Pour les récupérer, à partir de chaque ligne/décennie, on a segmenté le texte en phrases à partir desquelles on a extrait les mots cibles, les adverbes, et les mots à sa droite, les adjectifs. Afin de réaliser le filtrage nécessaire, avant l'extraction, on a étiqueté les mots avec leurs parties du discours en utilisant la librairie CoreNLP, disponible en R. La segmentation en phrase, malgré la difficulté, est donc une étape nécessaire pour réaliser l'étiquetage efficacement.

En ce qui concerne la polarité, dans un premier temps, on a utilisé le module CoreNLP pour réaliser une analyse des sentiments présents dans les phrases, ensuite, on a utilisé des bigrams et enfin seulement des adjectifs. Le fait d'avoir récolté ses différentes données, qui peuvent sembler redondantes, et le choix final à propos de quel donnée exploiter seront expliqués dans la section suivante.

Tout ce processus nous a permis d'obtenir un grand dataset contenant les occurrences des marqueurs d'intensité et leur contexte droite, avec d'autres informations concernant leur polarité, qui pourrait servir dans le cadre de plusieurs études. Dans notre cas, on a immédiatement filtré les données dans le but de garder seulement les adverbes accompagnés

des adjectifs. Pour cela, on a visé leurs étiquettes conformément au schéma d'annotation du Penn Treebank[3].

Exploration de données

Résumé des comptages de fréquences

Notre corpus est composé, finalement, d'un total de 394 449 164 tokens, avec une distribution qui augmente, en termes généraux, au fil des années (voir. Figure 1 et Annexe 1, dans l'annexe)

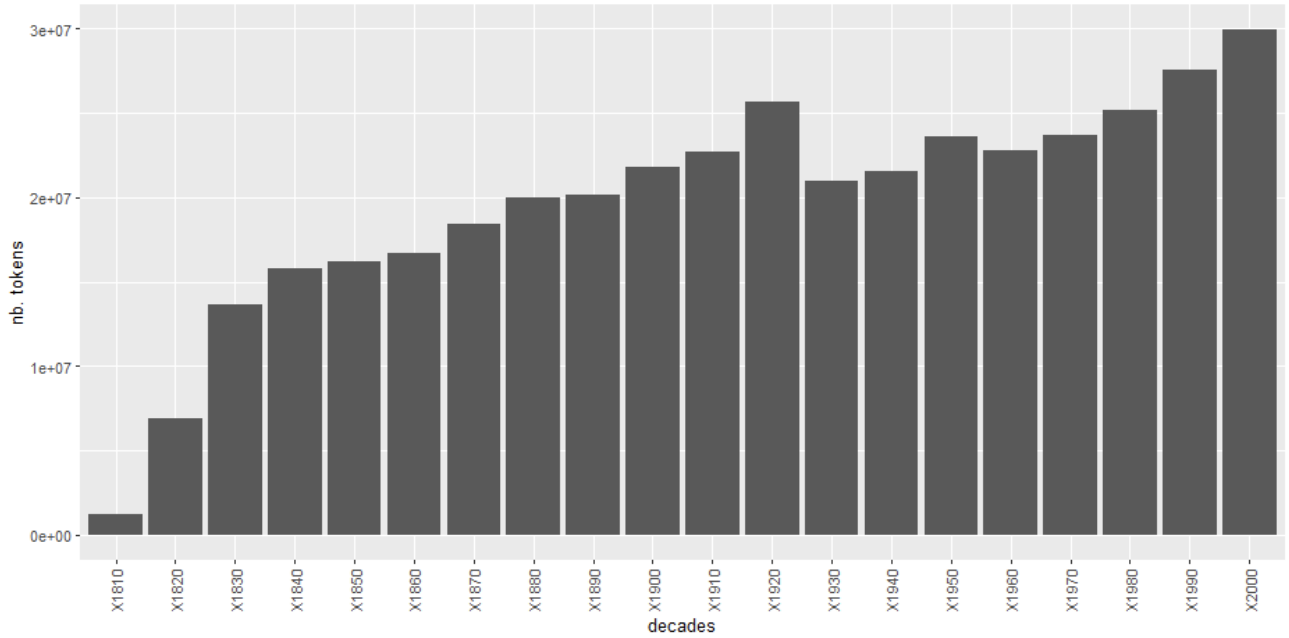
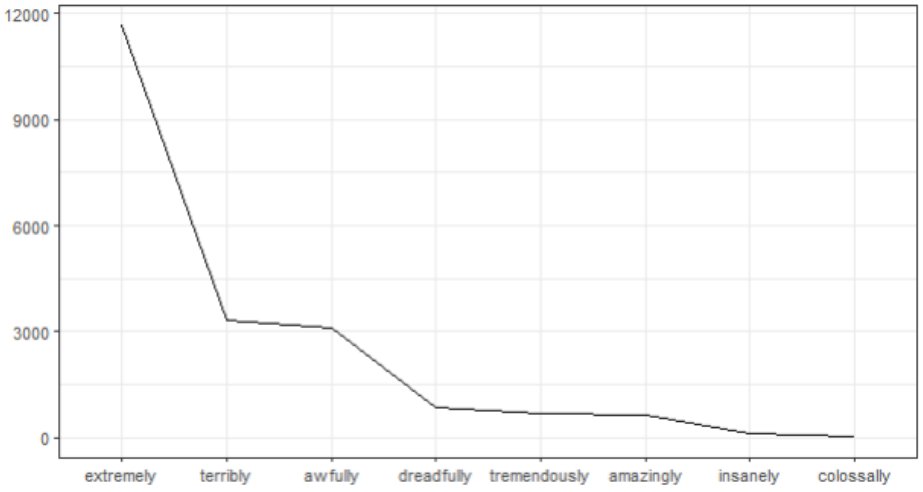


Figure 1 : Distribution des tokens par décennie

Les décennies 1810-20 sont clairement sous représentées, tandis que le 20ème siècle à une majeure présence dans le corpus. Une homogénéisation de valeurs s'avère nécessaire pour pouvoir exploiter correctement le corpus.

En ce qui concerne les marqueurs d'intensité, on observe une disparité importante entre les différents adverbes choisis. *Extremely, terribly* et *awfully* ont des valeurs de fréquence plus élevées que les restes, étant donné que le premier est celui qui dépasse considérablement la moyenne. Les mots comme *insanely, ou colossally*, bien qu'intéressants d'un point de vue sémantique, ont une présence très réduite dans la totalité du corpus, motifs pour lesquels on a décidé de les exclure tout suite de notre étude. En outre, si on constate aussi que le ratio nombre des marqueurs / tokens (Figure 3 et Figure 4) montre une dispersion assez importante entre la première moitié du 19ème siècle et le reste du corpus. La décennie 1970 représente une exception dans la mesure où nous n'avons pas trouvé un nombre significatif de marqueurs d'intensité de ce type.

[3] M. Marcus, B. Santorini and M.A. Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. In *Computational Linguistics*, volume 19, number 2, pp. 313–330.



	targetWord	n
1	extremely	11651
2	terribly	3339
3	awfully	3088
4	dreadfully	846
5	tremendously	659
6	amazingly	638
7	insanely	126
8	colossally	16

Figure 2 et Tableau 1 : Fréquence des marqueurs d'intensité

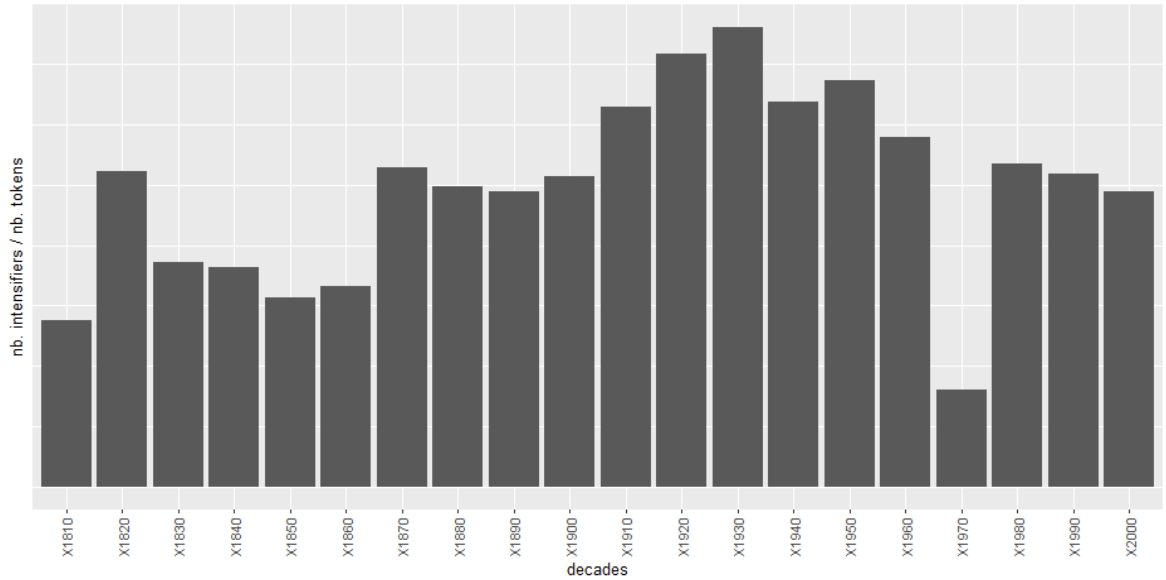


Figure 3: Ratio du nombre des marqueurs d'intensité pour chaque décennie

Étant donné cette dispersion et la nécessité de trouver des périodes qui nous permettent d'analyser le changement ou non de l'usage de ses marqueurs, on a décidé de regrouper certaines décennies entre-elles. Cela nous a permis de réduire le nombre d'observations à analyser, sans perdre la variable temps. On s'est décidé à réaliser une segmentation classique par demi-siècles : de 1810 à 1850, de 1860 à 1900, de 1910 à 1950 et de 1960 à 2000. En effet, ce nouveau regroupement permet de rendre l'analyse plus interprétable et, aussi, de réduire la dispersion de fréquences dans nos subsets (Figure 4). Néanmoins, le groupement de 1810-1850 continue à être moins comparable que les autres.

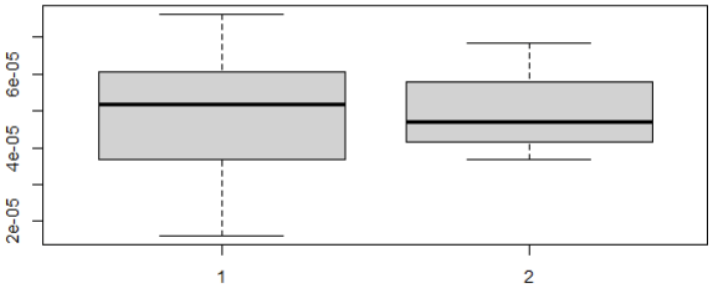


Figure 4 : Boxplots du ratio nb de marqueurs d'intensité / nb de tokens. 1 pour les décennies séparées, 2 pour les décennies regroupées.

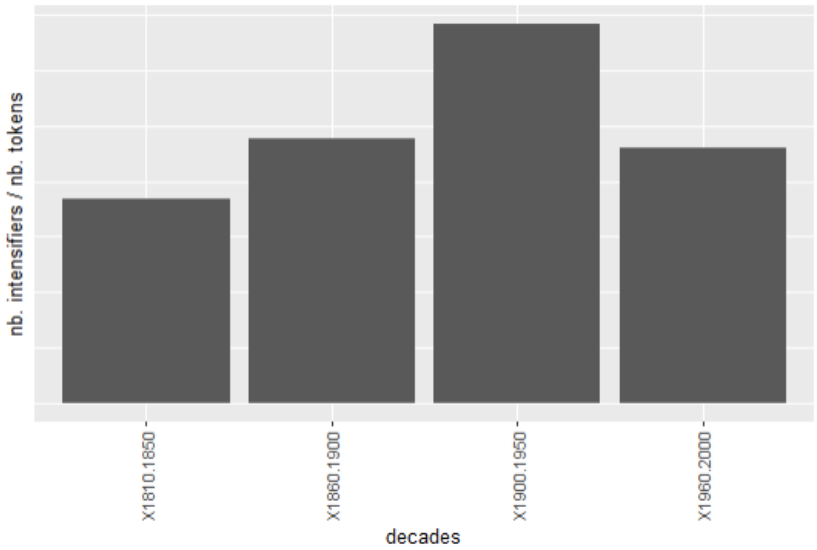


Figure 5 : Ratio du nombre des marqueurs d'intensité pour chaque groupe de décennies.

Polarité

Comme on l'a dit dans l'introduction, étudier la polarité de ses constructions est une des approches que nous avons choisies. Dans un premier temps, on a décidé d'utiliser l'analyseur de sentiment de CoreNLP[4], basés sur une architecture de réseaux de neurones et des règles, qui comporte 5 niveaux de polarité : neutre, très positif, positif, très négatif et négatif. Pour simplifier nos variables, les niveaux très positif et très négatif ont été classés respectivement dans les classes de positif et négatif.

On a extrait la polarité des phrases où les constructions apparaissent, la polarité du bigram (adverbe-adjectifs) et, enfin, la polarité de l'adjectif seulement. Il s'agissait de voir si c'était possible de considérer que la phrase, comme le contexte, déterminait la polarité de la construction et pouvoir ainsi étudier les occurrences du bigram dans un contexte d'usage plus large. Néanmoins, étant donné que la segmentation en phrases n'est pas fiable et que, pendant une observation préliminaire des résultats, on a constaté quelques résultats douteux, on a décidé de ne pas utiliser la polarité des phrases. Pour un autre problème, on a aussi exclu la polarité obtenue à partir de bigrams : l'analyseur de sentiment avait la tendance à

[4] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60. [pdf] [bib]

classer les constructions qui avaient des polarités lexicalement opposées comme neutres, de plus cette façon de procéder n'était pas systématique. Tous ces problèmes ont amené à un changement notre méthode d'analyse de la polarité à faveur de l'utilisation du lexique *Opinion Lexicon*[5], une ressource stable, systématique et linguistiquement fondamentale. Les observations des polarités des adjectifs co-occurrents à partir du CoreNLP est disponible dans l'annexe (voir Annexe 3 et 4).

En ce qui concerne le lexique de polarité en question, il s'agit d'une liste de mots d'opinion ou de mots de polarité positive et négative en anglais. Il est composé d'environ 6 800 mots. Cette liste a été compilée pendant de nombreuses années, à partir des expériences de Bing Liu. Ce lexique est assez vaste, non seulement en ce qui concerne le nombre de mots, mais aussi en ce qui concerne les abréviations de mots et les éventuelles fautes de frappe. On a donc étiqueté chaque adjectif selon l'appartenance de mots positifs ou négatifs aux differentes listes. Les mots qui ne se trouvaient dans aucune liste, ont été classé comme NA.

Comme on l'a dit, après avoir analysé les données, on a décidé d'annoter uniquement les adjectifs, le contexte droit du mot cible, ce qui a donné des résultats plus précis en termes d'analyse de la polarité. À cet égard, on pense aussi que l'analyse des adjectifs grâce au lexique est plus précise sur un seul mot. À cause de ces réflexions et afin de créer des graphes permettant d'explorer notre hypothèse, on a été obligé de modifier notre dataframe de début. Plus spécifiquement, les graphes ci-dessous ont été créés en utilisant uniquement la polarité du contexte droit (adjectif), qui a été ajouté au dataset final.

Dans les graphes ci-dessous, on peut observer la proportion des adjectifs cooccurrents positifs et négatifs sur la totalité de chaque marqueur d'intensité. On peut aussi voir que la combinaison des décennies permet d'obtenir une analyse diachronique plus équilibrée (Figure 6 vs Figure 7). Dans ces graphes, on a utilisé les trois termes les plus fréquents, *extremely, awfully et terribly,* et *amazingly*, dont l'occurrence est plus équilibrée en termes de polarité.

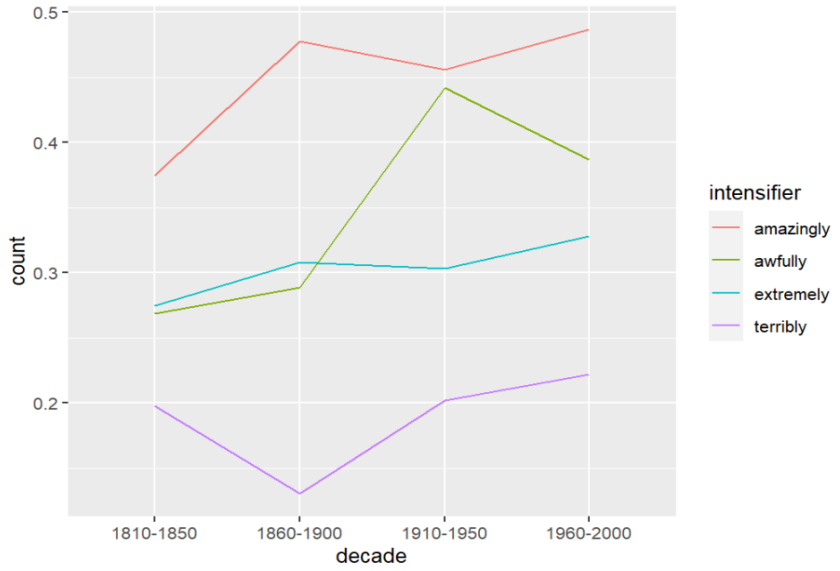


Figure 6: Ratio adjectifs positifs / toutes les adjectifs pour chaque marqueurs d'intensité, annotés avec Opinion Lexicon

[5] Minqing Hu and Bing Liu. "Mining and summarizing customer reviews." *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper)*, Seattle, Washington, USA, Aug 22-25, 2004.

Le changement le plus grand de l'usage positif dans le temps est observé dans le cas de *awfully*, alors que le changement le moins prononcé est observé dans *extremely*. L'usage de *amazingly* est resté majoritairement positif tout au long des décennies, tandis que *terribly* a l'usage positif le plus faible. Toutefois, au cours des cinq dernières décennies, l'utilisation positive de *awfully* a diminué. La plus grande différence que nous observons se situe entre le deuxième et le troisième groupe de décennies pour le marqueur *awfully*. Puisque ces deux groupes ont un taux de marqueurs d'intensité plus équilibré, il semblerait y avoir un changement dans l'usage de awfully et possiblement une relation statistique significative entre la polarité de la construction et le temps. Cette significativité statistique reste à être vérifiée. En outre, ces graphes montrent que de tels changements ne se sont pas produits pour tous les marqueurs d'intensité que nous avons choisi d'analyser. L'une des raisons peut être que dans le corpus l'utilisation de ces marqueurs d'intensité avec des d'adjectifs est peu fréquente. En tous cas, les piques vers le bas nous avertissent d'une anomalie dans le corpus, voir 1850. Comme nous l'avons dit, il semble préférable de ne pas travailler avec le premier groupe de décennies.

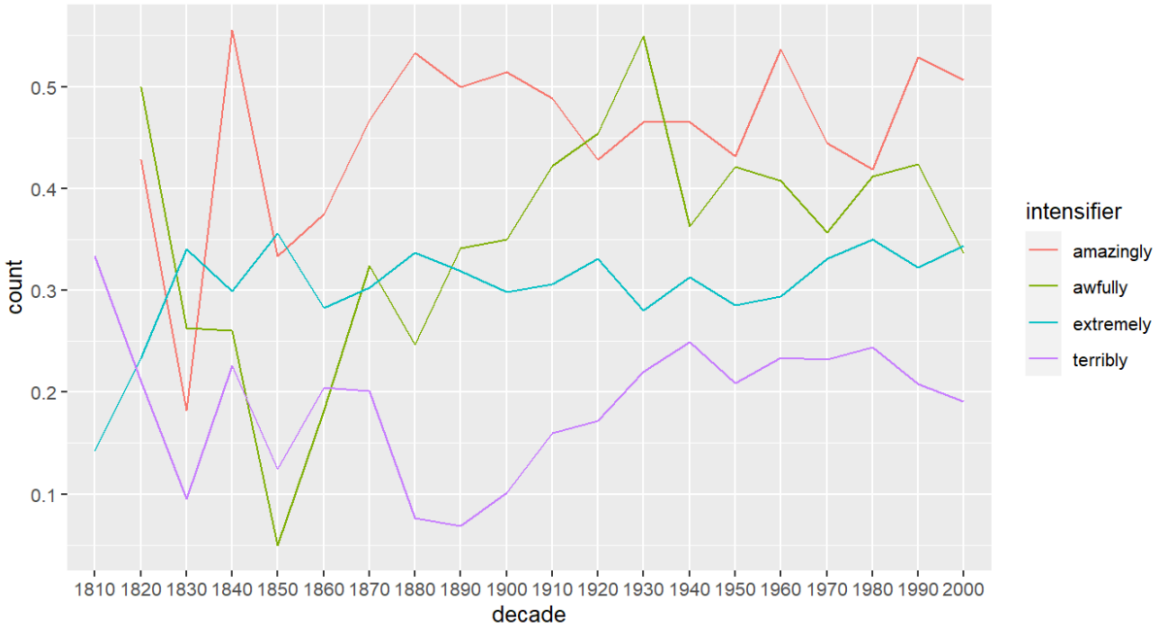


Figure 7: Ratio adjectifs positifs / tous les adjectifs pour chaque marqueurs d'intensité, annotés avec Opinion Lexicon. Les décennies ne sont pas regroupées.

À travers la méthode *Multiple correspondence analysis (MCA)*, on a analysé la relation entre nos observations (marqueur, polarité de l'adjectif et groupe de décennies) et nos variables, afin de pouvoir inspecter encore plus en profondeur l'usage de *awfully* et *terribly.* La première dimension, dans les deux cas, montre une importante séparation entre les décennies 1810-50 et le reste du corpus. Cette différence était attendue. A continuation, on verra que les adjectifs qui sont co-occurrent avec les marqueurs sont aussi très différents[6]. Dans la deuxième dimension, on observe, qu'effectivement, de 1910 à 2000 l'usage de *terribly* et *awfully,* est plus positif que pendant les décennies 1860-1900. Il faut souligner que *terribly* est de façon générale plus proche de la polarité négative que awfully.

[6] Il se doit principalement à la conformation du corpus : les genres sont mal équilibrés. Voir https://www.english-corpora.org/coha/

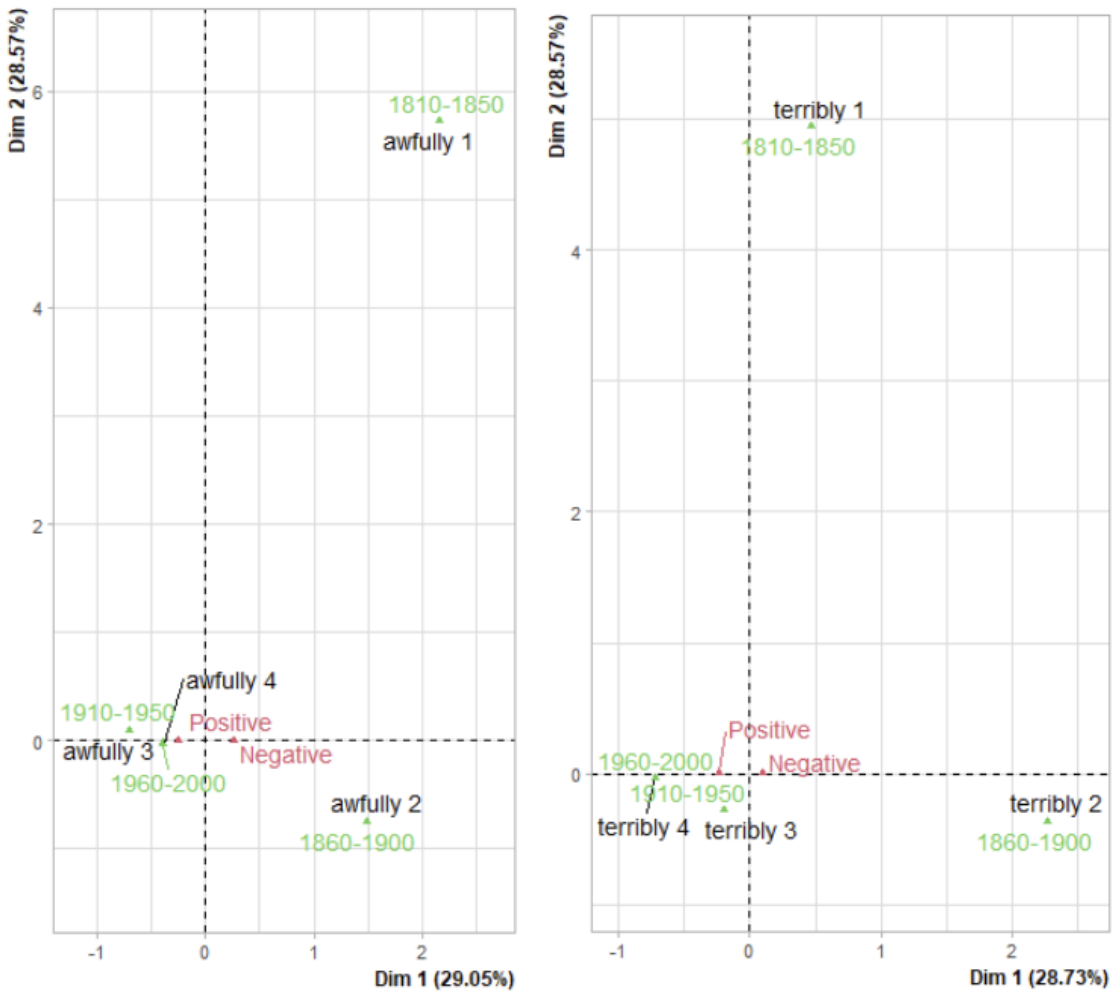


Figure 8: MCA sur les observations 1) marqueur, 2) polarité de l'adjectif et 3) groupe de décennie pour *awfully* et *terribly*.

Adjectifs co-occurrents

L'autre approche à considérer est l'étude des adjectifs co-occurrents. Connaître quelles sont les occurrences les plus fréquentes nous permet surtout de comprendre leur usage. Il s'agit de pouvoir caractériser chaque adverbe selon leur contexte adjectival immédiat. Pour profiter de nos observations dans la section précédente, on a choisi de nous focaliser plus spécialement sur *awfully* et *terribly*, en espérant trouver d'autres informations complémentaires pour solidifier notre hypothèse, maintenant, plus restreinte dans sa portée.

Une première approximation nous montre que ces adverbes partagent seulement deux adjectifs co-occurrents : *sorry* et *hard*. En effet, étant donné que nous travaillons avec de hautes fréquences, on est dans le domaine de la collocation et chaque adverbe semble avoir ses propres bases collocatifves (ici, les adjectifs). De plus, on observe que *terribly* co-occurre plus avec des mots qu'on peut considérer comme neutre ou négatif que *awfully*.

	targetWord	rightContext	n
1	terribly	sorry	245
2	terribly	afraid	110
3	terribly	wrong	105
4	terribly	important	84
5	terribly	hard	82
6	terribly	frightened	71
7	terribly	upset	45
8	terribly	lonely	41
9	terribly	tired	40
10	terribly	hot	37

	targetWord	rightContext	n
1	awfully	good	296
2	awfully	sorry	278
3	awfully	nice	175
4	awfully	glad	130
5	awfully	hard	96
6	awfully	tired	57
7	awfully	funny	44
8	awfully	long	43
9	awfully	fond	42
10	awfully	clever	41

Tableau 2 : les 10 bigrams les plus fréquents dans le corpus pour *terribly* et *awfully*

Ensuite, on a décidé d'appliquer une autre méthode multivariée pour analyser nos données : le positionnement multidimensionnel (MDS). L'idée était de profiter des fréquences de bigrams pour réaliser une matrice de distance ou de disimilarité capable de nous donner des informations sur la proximité d'un mot par rapport à un autre, selon sa distribution dans le corpus. Pour cela, au lieu d'évaluer les adjectifs en fonction des adverbes, on a préféré essayer d'évaluer la distance entre les marqueurs selon leur usage dans différents groupes de décennies. On a pris en compte les 50 adjectifs les plus fréquents du marqueur pour chaque époque. Il s'agissait de créer une matrice à partir d'une information notable et c'est pour cela qu'on a ignoré les basses fréquences.

En effet, on a trouvé des résultats similaires à ceux qu'on avait vu dans la section précédente. Ici, dans la version clusterisée (K-means), on voit que la distance entre l'usage du marqueur d'intensité *awfully* de 1860-1900 et 1910-1950 est plus courte, ayant une similarité sémantique plus proche.

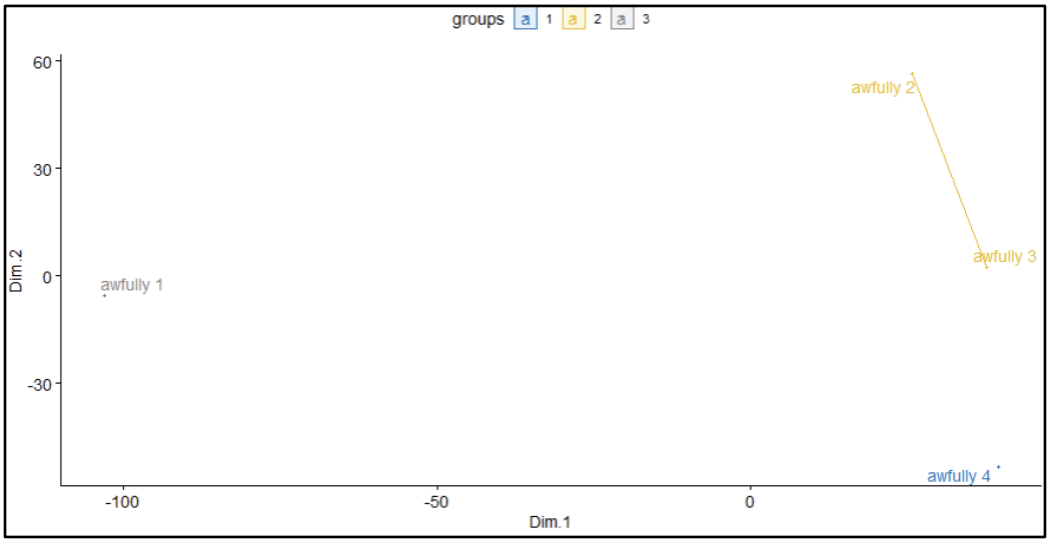


Figure 9 : MDS + K-means. *Awfully* est mappé selon la distribution qui l'a avec les 50 plus fréquents adjectifs co-occurrents dans chaque époque. 1 : 1810-1850, 2 :1860-1900, 3 : 1910-1950, 4 : 1960-2000

Si on regarde la version du MDS pour le marqueur d'intensité *terribly,* la tendance, vue dans la section précédente, se réaffirme : l'usage dans les décennies entre 1910 et 2000, où les adverbes sont sémantiquement proches, coïncide avec le moment où la polarité est la plus positive.

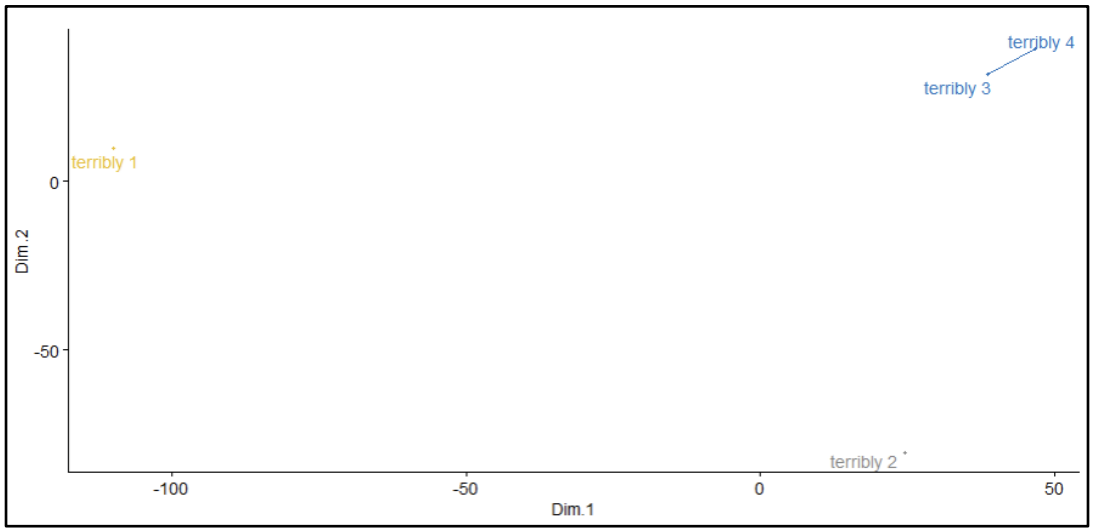


Figure 10 : MDS + K-means. *Terribly* est mappé selon la distribution qui l'a avec les 50 plus fréquents adjectifs co-occurrents dans chaque époque. 1 : 1810-1850, 2 :1860-1900, 3 : 1910-1950, 4 : 1960-2000

Dans les deux cas, la première période de décennies reste substantiellement différente aux reste, marquant plutôt une particularité du corpus que de l'évolution des marqueurs d'intensité.

On a essayé aussi de présenter de façon interprétable les lignes et les colonnes en même temps, à travers la méthode CA, mise à échelle (*scaled symmetric biplot*). On a projeté les 7 adjectifs les plus fréquents pour les deux marqueurs d'intensité, à travers les différents groupes de décennies (Figure 11 pour *awfully*), mais les résultats sont plus difficiles à interpréter. Il n'y a pas assez d'adjectifs de co-occurrents pour tirer une conclusion partielle différente à celle qu'on a déjà exprimée. En ajoutant plus d'adjectifs on rendrait le graphe illisible.

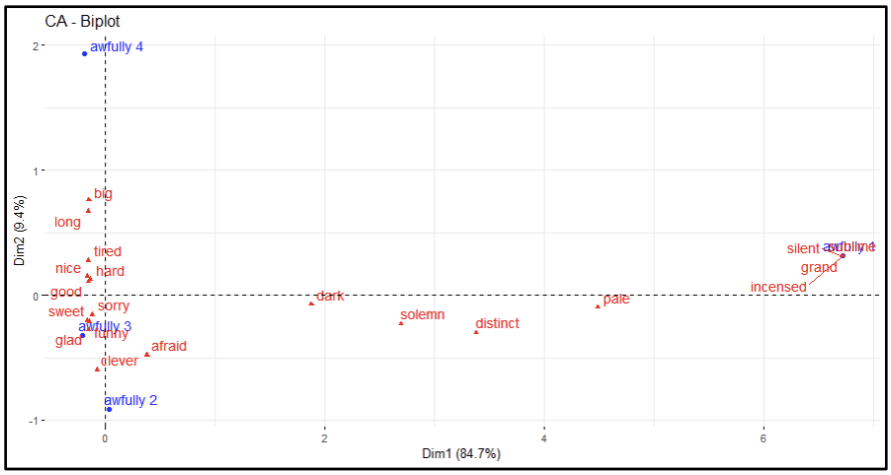


Figure 11 : CA *scaled symmetric biplot* pour représenter lignes et colonnes du marqueur d'intensité *awfully* avec ses 7 plus fréquents co-occurrents, par groupe de décennies.

Reformulation d'hypothèse et tests

Arrivés à ce point, à la lumière de l'exploration des données, notre hypothèse de travail à subit des modifications. Elle était, au départ, trop ambitieuse et vague. Assurer que les constructions marqueurs d'intensité-adjectif de polarité contraire sont plus répandues dans l'actualité est finalement difficile à prouver. En revanche, on a trouvé dans nos corpus des tendances de changement dans l'usage des deux marqueurs d'intensité qui semblent être statistiquement significatives. Notamment, dans le cas de l'adverbe *terribly,* on observe une tendance à la hausse de son usage avec des adjectifs de polarité positifs au cours du dernier siècle. Dans le cas de l'adverbe *awfully,* son usage dans un contexte positif semble augmenter depuis la moitié du siècle passé et diminuer au fur et à mesure que l'on se rapproche des années 2000.

Ainsi, une première reformulation de l'hypothèse est la suivante :

- H0: l'usage des marqueurs *terribly et awfully* (ses adjectifs co-occurrents et la polarité de ces derniers) et le temps et le moment il est utilisé sont indépendants.
- H1: l'usage des marqueurs *terribly et awfully* (ses adjectifs co-occurrents et la polarité de ces derniers) et le temps et le moment utilisés sont interdépendants.

Une deuxième pair d'hypothèses plus spécifiques est possible :

- H0: l'usage des marqueurs *terribly et awfully* avec un adjectif de polarité lexicalement opposé n'est plus répandu à mesure que l'on se rapproche des années 2000.
- H1: l'usage des marqueurs *terribly et awfully* avec un adjectif de polarité lexicalement opposé est plus répandu à mesure que l'on se rapproche des années 2000.

Afin de vérifier ses hypothèses, en sachant où cibler grâce à notre exploration, on a appliqué le Test exact de Fisher. Il s'agit de constater qu'il y a une relation statistiquement significative entre l'usage des marqueurs et les différents groupes de décennies qui justifierait une étude plus approfondie de ces marqueurs. En ce qui concerne le choix du test, étant donné la taille de nos observations et de nos variables, on a considéré qu'il était tout à fait pertinent de l'utiliser, malgré son coût.

Pour cela, on a créé une matrice 2x2 avec les occurrences de chaque marqueur selon leur contexte de polarité pour chaque période et on les a comparées. De nouveau, on a exclu la période 1810-1850 à cause de son caractère problématique, comme nous l'avons déjà expliqué.

	1860-1900 / 1910-1950	1910-1950 / 1960-2000	1860-1900 / 1960-2000
awfully	2.727e-09	0.4563	3.35e-05
terribly	0.3948	0.001753	0.0002495

Tableau 3 : Valeurs de la p-value du Test exact de Fisher, calculé à partir d'une matrice 2x2 avec les occurrences de chaque marqueur selon leur contexte de polarité pour chaque période.

Dans le cas de l'adverbe *awfully*, on voit qu'entre les périodes 1910-1950 et 1960-2000, il n'y a pas de relation significative entre les variables. Cela coïncide avec les périodes où awfully, au début, a tendance à être utilisé avec des adjectifs de polarité positive, puis diminue à la fin. Mais, en revanche, si on compare ces deux périodes avec la période de 1860-1900, on observe que son usage "positif" semble être relié au moment de sa réalisation. Du côté de *terribly*, les valeurs sont faiblement significatives pour les comparaisons entre les périodes 1910-1950 / 1960-2000 et 1860-1900 / 1960-2000. En effet, ce sont des périodes où son usage a changé. Enfin, ces résultats nous encouragent à continuer l'étude de ses marqueurs.

Conclusion

À travers de différents aspects du phénomène étudié, en utilisant des données sur les co-occurrences et la polarité, en tirant partie des ressources externes, comme le lexique, et de plusieurs techniques exploratoires multivariées à travers un nouveau langage de programmation, on a réussi avoir un aperçu de l'usage de quelques marqueurs d'intensité dans le cas de l'anglais américain. Les résultats obtenus justifient de continuer à étudier ces constructions afin de pouvoir déterminer plus nettement leur changement au fil du temps. L'utilisation de la nouvelle version du corpus COHA nous intéresse beaucoup, notamment pour l'équilibrage de dernières décennies et l'ajoute de la période 2010-2019, éléments qui seraient idéaux pour continuer notre étude. Enfin, il serait nécessaire d'élargir notre liste de marqueurs à étudier.

Bibliographie

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60. [pdf] [bib]

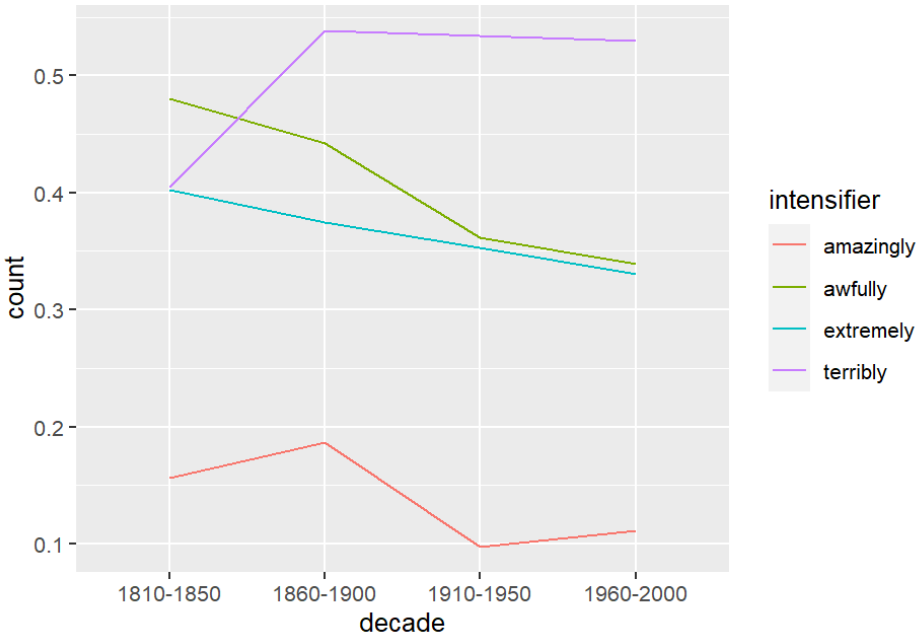
M. Marcus, B. Santorini and M.A. Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. In *Computational Linguistics*, volume 19, number 2, pp. 313–330.

Minqing Hu and Bing Liu. "Mining and summarizing customer reviews." *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper)*, Seattle, Washington, USA, Aug 22-25, 2004.

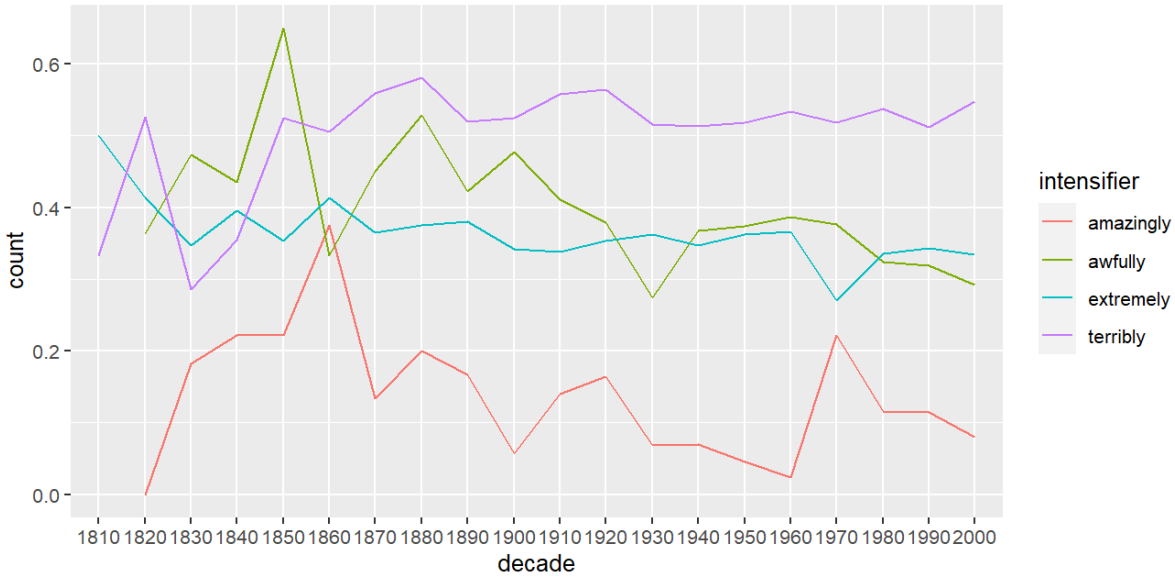
Annexe

décennie	nb. de tokens	décennie	nb. de tokens
1810	1 198 660	1910	22 659 982
1820	6 879 778	1920	25 640 921
1830	13 641 864	1930	20 960 877
1840	15 821 313	1940	21 516 313
1850	16 234 862	1950	23 617 883
1860	16 705 700	1960	22 810 220
1870	18 411 598	1970	23 708 597
1880	19 978 425	1980	25 182 408
1890	20 180 783	1990	27 583 376
1900	21 802 294	2000	29 913 310
TOTAL			394 449 164

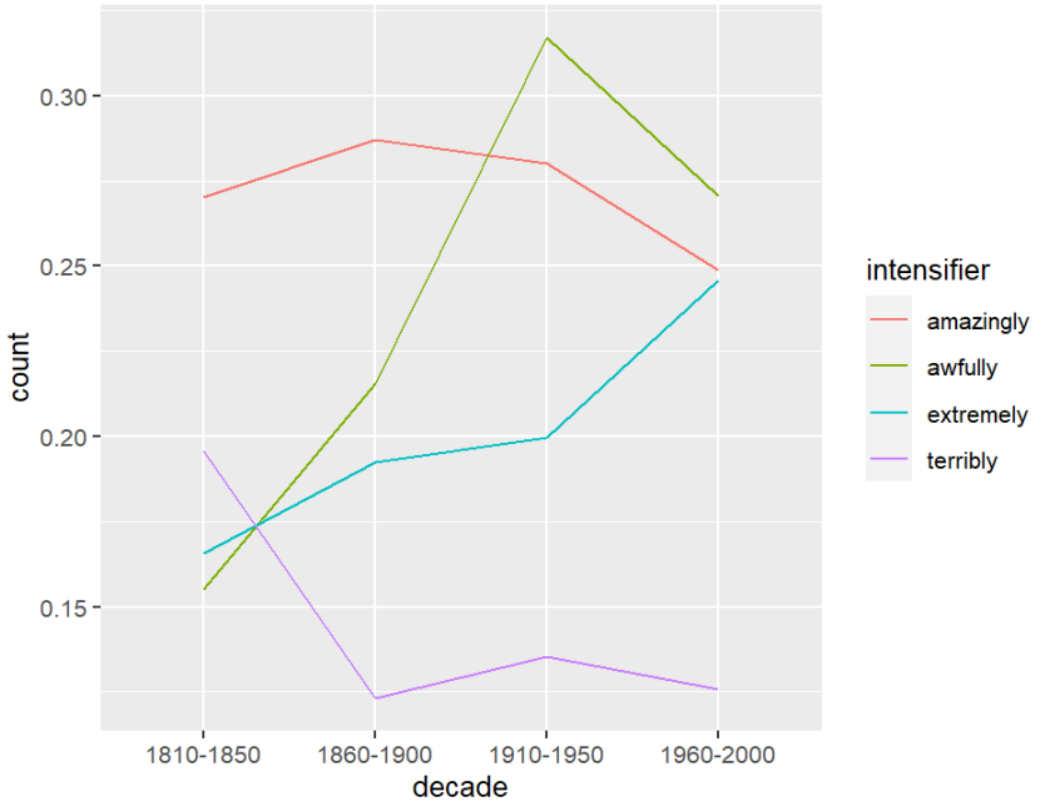
Annexe 1 : Exploration des données du corpus



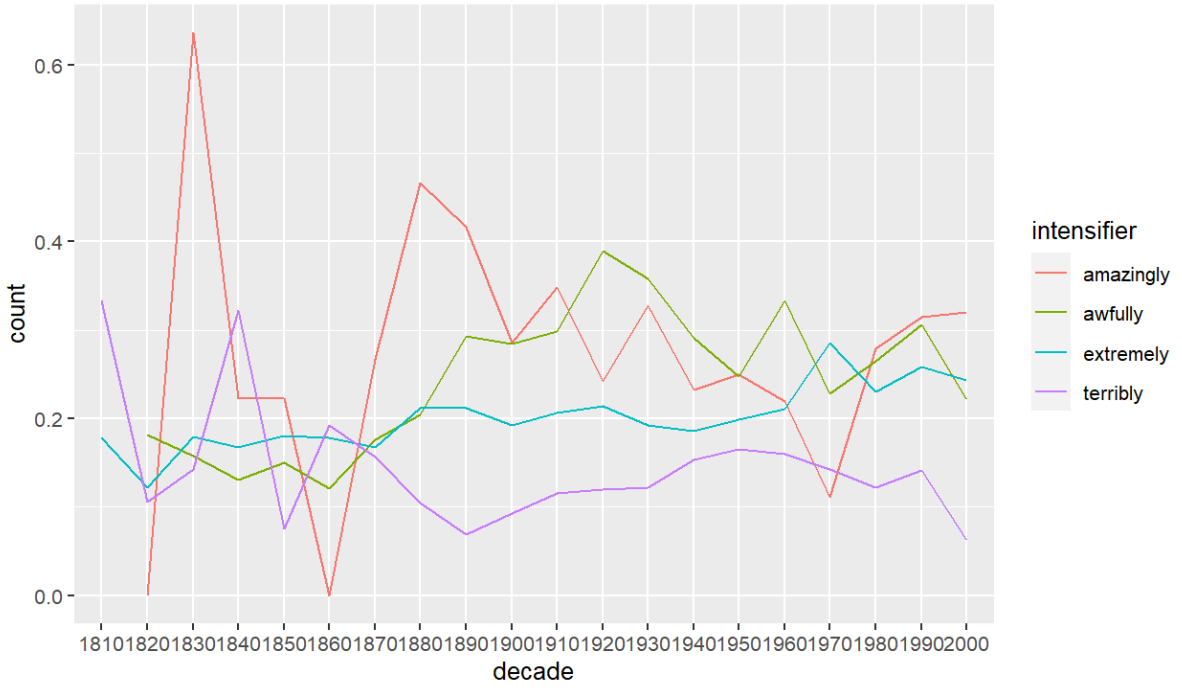
Annexe 2 : Ratio adjectifs négatifs / toutes les négatifs pour chaque marqueurs d'intensité, annotés avec Opinion Lexicon



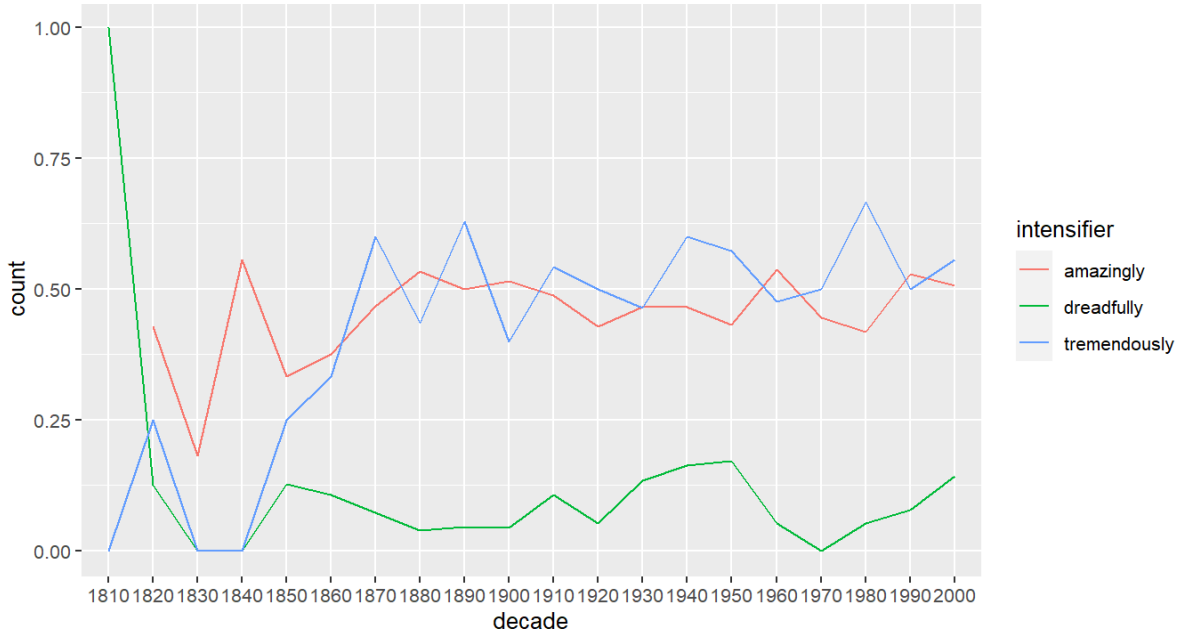
Annexe 3 : Moyens d'utilisation *négatifs* des marqueurs d'intensité annotés avec CoreNLP



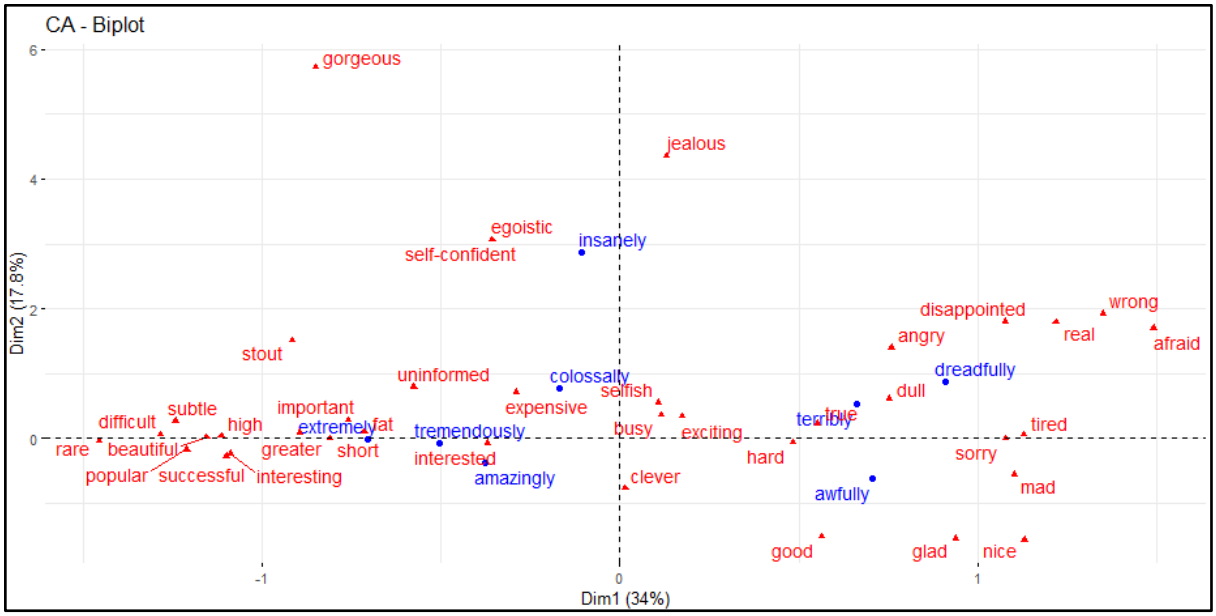
Annexe 4 : Moyens en groupe d'utilisation *positifs* des marqueurs d'intensité annotés avec CoreNLP



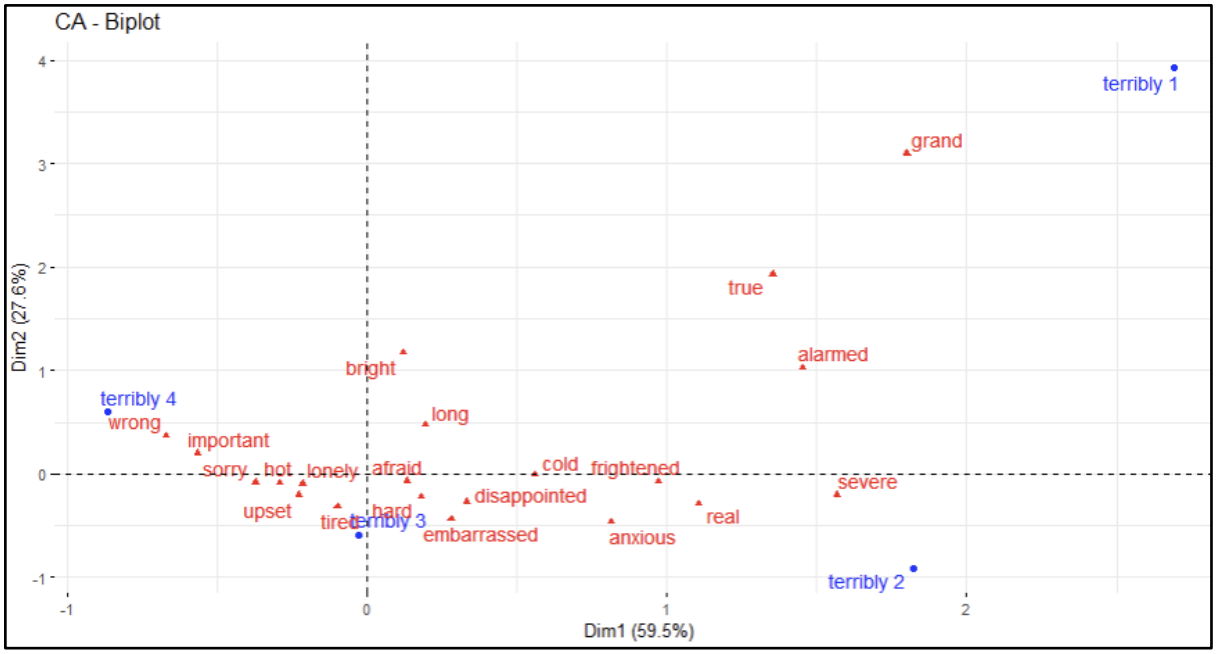
Annexe 5 : Moyens d'utilisation *positifs* des marqueurs d'intensité annotés avec CoreNLP



Annexe 6: Exploration des autres termes avec Opinion Lexicon



Annexe 7 : Biplot mis à l'échelle pour représenter lignes et colonnes de chaque marqueur d'intensité avec ses 5 plus fréquents co-occurrents.



Annexe 8 : Biplot mis à l'échelle pour représenter lignes et colonnes du marqueur d'intensité *terribly* avec ses 7 plus fréquents co-occurrents, par groupe de décennies.