# Les méthodes de la statistique lexicale au service de l'enseignement du vocabulaire du FLE

**Dr. Itidal Abdulhalim Habib***

**(Déposé le 20 / 12 / 2021. Accepté 3 / 3 / 2020)**

## □ Résumé □

Mis au service de l'enseignement des langues, les méthodes de la statistique lexicale sont considérées comme un outil facilitant le choix du vocabulaire indispensable pour l'enseignement d'une langue étrangère. Le dispositif, servant de méthode de la statistique lexicale et proposant des fonctions et des applications diverses, constitue pour les chercheurs une source inépuisable d'exploitations diverses fondées sur des bases scientifiques précises attribuant ainsi de la crédibilité et de l'authenticité à leurs recherches.
Le travail qui suit se base sur des recherches empiriques et théoriques pour mettre en relief les atouts des méthodes de la lexicométrie. Il souligne également les problèmes que l'enseignant peut rencontrer en mettant en place une telle méthode selon la nature des textes soumis au traitement automatique.
Les critères de fréquence, de disponibilité et de rentabilité seront indispensables à déterminer le choix du vocabulaire nécessaire à l'enseignement du français langue étrangère**.**

**Mots clés:** fréquence, statistique, logiciel Hyperbase, vocabulaire, enseignement, langue étrangère.

* **Professeur adjoint à l'institut Supérieur des Langues – Section de français – Université Tichrine - Lattaquié- Syrie. itidalhabib@yahoo.com**

# دور مناهج الإحصاء المعجمية في تعليم مفردات اللغة الفرنسية بوصفها لغة أجنبية

**د. اعتدال عبد الحليم حبيب** *

**(تاريخ الإيداع 20 / 12 / 2021. قبل للنشر في 3 / 3 / 2022)**

## □ ملخّص □

تعتبر مناهج الإحصاء المعجمية أداة تسهل اختيار المفردات الأساسية والضرورية لتدريس لغة أجنبية وذلك عندما توضع في سبيل تعليم اللغات .

يشكل البرنامج الذي يعتبر منهجا في الإحصاء المعجمي بفضل تطبيقاته المتنوعة، بالنسبة للباحثين، مصدرًا لا ينضب لمختلف الاستخدامات المبنية على أسس علمية دقيقة و تضفي بالتالي صفة المصداقية والأصالة إلى أبحاثهم.

يستند العمل التالي على البحوث التجريبية والنظرية لتسليط الضوء على مزايا مناهج الإحصاء المعجمي. ويشير أيضا إلى المشاكل التي قد تواجه الباحث عند تنفيذ مثل هذه الطريقة اعتمادا على طبيعة النصوص الخاضعة للمعالجة بوساطة البرنامج الالكتروني.

ستكون معايير التكرار والتوافر والفعالية ضرورية في تحديد اختيار المفردات اللازمة لتعليم اللغة الفرنسية كلغة أجنبية.

**الكلمات المفتاحية:** التكرار ، الإحصاء ، برنامج Hyperbase ، المفردات ، التدريس ، اللغة الأجنبية .

* **استاذ مساعد – قسم اللغة الفرنسية – المعهد العالي للغات– جامعة تشرين– اللاذقية– سورية.itidalhabib@yahoo.com**

**Introduction**

Il est connu que, la statistique lexicale recherche les "différences", les "spécificités", et la "variété". Dans tout texte, un petit nombre de mots représente la plus grande partie des occurrences. L'emploi des méthodes de la statistique assure aux recherches une base scientifique. Aussi le propre de la statistique est d'être un instrument de comparaison permettant de déterminer la propriété d'une œuvre ou les constantes d'un genre.

En termes de perception, le « mot » est une réalité; ce sont les « mots » qui constituent le lexique interne des locuteurs. Et le lexique est une réalité de langue à laquelle on ne peut accéder que par la connaissance des vocabulaires particuliers qui sont une réalité de discours.

Comment peut-on déterminer le vocabulaire à enseigner? Nous allons voir que le vocabulaire qui sert de base pour l'enseignement du français, sera déterminé grâce au critère de fréquence, qui est liée aux notions d'utilité, de rentabilité et de disponibilité.

**But de l'article**

Présenter aux futurs chercheurs, étudiants en Master et/ou Doctorat, aux universités syriennes, département du français, une nouvelle méthode pour la recherche scientifique sur le vocabulaire. Cela est pour leur communiquer un outil de recherche leur permettant de mener des études sur de grands corpus de tout type en leur donnant la possibilité d'en extraire les particularités ou les spécificités en les comparant à d'autres corpus. Attribuer aussi aux chercheurs dans le domaine de l'enseignement du français la possibilité de délimiter et choisir efficacement le vocabulaire à enseigner; vocabulaire spécifique, propre à un domaine donné répondant ainsi aux besoins des apprenants correspondant à leurs niveaux d'apprentissage.

**Méthodologie de travail**

Ce travail porte sur la présentation de certaines recherches dans le domaine de la lexicométrie et la statistique lexicale pour en tirer profit pour l'enseignement du français. Premièrement, nous exposons quelques notions spécifiques de ce domaine, à savoir: "statistique lexicale", "unités lexicales", "lexique", et "vocabulaire"; nous présentons le logiciel qui facilite la réalisation des études statistiques, à savoir: Hyperbase.

Deuxièmement, nous fournissons un exemple tiré de notre étude réalisée sur le "genre" publicitaire pour montrer l'efficacité de l'utilisation de la statistique et la nouvelle technologie dans les recherches scientifiques sur le vocabulaire.

Nous essayons, troisièmement, d'aborder le rapport entre la fréquence et l'enseignement des langues en suivant la méthode adoptée pour l'élaboration du Français Fondamental.

**Statistique lexicale**

Selon des recherches statistiques, on a constaté que dans tout texte, un petit nombre de mots couvre la majeure partie des occurrences; l'emploi des méthodes de la statistique fournit à ces recherches une base scientifique. Les deux ouvrages de Charles Muller, celui de (1973) et de (1977) présentent clairement les techniques de la statistique lexicale.

On sait que l'une des caractéristiques de la statistique lexicale est de ne pas effectuer de distinction entre les différentes catégories de mots, l'ensemble étant considéré comme des formes graphiques de poids équivalents : « chaque discours est pour les programmes de calcul un sac de mots dont seul le profil de fréquences est (…) exploité » (Lebart et Salem, 1994 : 146).

Il est connu aussi que le propre de la statistique est d'être un instrument de comparaison ; la fréquence d'un mot dans un corpus donné demande à être interprétée par

rapport à sa fréquence dans un autre corpus si le calcul révèle que la différence est significative, ne pouvant être due au hasard, et non simplement aléatoire.

Muller (1979) souligne que la statistique lexicale, dans la majorité de ses applications, recherche les différences plutôt que les faits constants, la variété plutôt que l'homogénéité. Il signale que l'on ne peut mesurer un écart que par rapport à une norme. Les méthodes de la lexicométrie se fondent sur la loi binomiale et la comparaison d'effectifs « théoriques » calculés et des observations faites sur les listes de vocabulaires permet de caractériser la particularité d'une œuvre ou les constantes d'un genre. Plus on se rapproche de la moyenne représentée par le modèle théorique, plus on est dans la distribution « normale ». Plus on s'éloigne du modèle théorique, plus on est dans l'exceptionnel, associé à des probabilités faibles.

**Les unités lexicales**

Dans les études lexicales quantitatives, la complexité de la sémantique lexicale présente des difficultés. L'ordinateur, considérant comme une unité de texte celle qui est comprise entre deux caractères délimiteurs, ne peut selon Muller (1979 : 125-143), tenir compte de la polysémie d'un mot, des mots composés, des flexions. Il s'agit là des unités qui se prêtent au dénombrement automatique.

Les données expérimentales de la psycholinguistique ont prouvé que, en termes de perception, le « mot » est une réalité. Ce sont les « mots » - caractérisés phonologiquement, syntaxiquement et morphologiquement, et dotés d'une signification permettant leur usage dans des situations précises - qui constituent le lexique interne des locuteurs des langues naturelles.

Plusieurs travaux sur l'accès lexical, la sémantique psychologique et la mémoire sémantique ont confirmé que le « lexique interne » contiendrait les mots complexes et les morphèmes de la linguistique, et que « les locutions (...) constitueraient des unités stockées comme telles au même titre que les mots » (Caron, 1992 : 76-77).

Les études sur la reconnaissance des contextes thématiques, particulièrement celle de Martin (1993 : 30-35), ont montré les grandes difficultés que rencontre le chercheur à cause de la polysémie et de l'appartenance d'un mot à différentes classes.

**Lexique et vocabulaire**

En didactique du FLE, le lexique selon Robert (2002 : 100), désigne « – en langue usuelle : *le dictionnaire succinct d'une science (un lexique d'astrologie), d'un domaine spécialisé (le lexique de l'aviation)*, - en littérature : *la langue d'un écrivain (le lexique de Proust)*, - en linguistique : *l'ensemble des lexèmes (ou morphème lexicaux) d'une langue*. »

Envisageons les "*représentations*" de ces deux notions selon les adeptes de la statistique lexicale:

Pour Guiraud le lexique est «*hypothétique* », alors que le vocabulaire n'est que le reflet du lexique dans un texte donné.

Wagner (1967 : 17-18), souligne que vocabulaire désigne généralement « un domaine du lexique qui se prête à un inventaire et à une description. »

Le vocabulaire et le lexique sont pour Genouvrier et Peytard (1970 : 181) en rapport d'inclusion : « Le vocabulaire est toujours une partie, de dimensions variables selon le moment et les sollicitations, du lexique individuel, lui-même partie du lexique global. »

Dans la même direction, Picoche (1992 : 46) ajoute que le lexique est une réalité de langue à laquelle on ne peut accéder que par la connaissance des vocabulaires particuliers qui sont une réalité de discours. Ainsi le lexique « transcende les vocabulaires mais n'est accessible

que par eux. » Le vocabulaire est aussi l'ensemble des « *vocables* » (Maciel, 1997 : 13) actualisés dans le discours.

Nous imaginons le lexique comme un ensemble où les différents éléments sont liés par différents rapports. L'acquisition du lexique et du vocabulaire, a lieu au cours de l'apprentissage et l'apprenant, progressivement, s'imprègne des mots et des structures de la langue.

**Dispositifs**

*Hyperbase* est un logiciel développé par Étienne Brunet, au laboratoire *Bases, corpus et langage*, (UMR 6039, INaLF – CNRS – Nice – France), il met en œuvre tous les outils de la méthode adoptée pour le traitement statistique. Il s'agit d'un « logiciel fondé sur une logique de navigation hypertextuelle, permettant, notamment, de feuilleter les pages du texte ou d'en parcourir l'index alphabétique, et de cliquer sur une forme, pour en rechercher la fréquence ou l'environnement. » (Marchand, 1998 : 43)

Ce logiciel propose deux programmes distincts pour la recherche de l'environnement d'un mot : « *Concordance* " et « *Contexte* ». Pour un corpus particulier, ce logiciel fournit des listes et un dictionnaire des fréquences. Il offre un grand choix de sélections et de calculs et il permet de comparer un corpus particulier au *Trésor de la Langue Française*, qui opère un décompte de formes établi sur la base de la littérature des XIX$^{e}$ et XX$^{e}$ siècles.

« *Hyperbase*, écrit son auteur, s'applique à toute langue qui utilise l'alphabet latin. » (Brunet, 2011: 4). *Ce* logiciel représente une méthode statistique essentielle pour la recherche sur le vocabulaire.

Nous signalons que, dans ses publications de 2018, 2019, 2020, et 2021, Ruggia profite des corpus numériques, des méthodes de la statistique lexicale, et de l'analyse des données textuelles , pour ses recherches dans le domaine de la didactique du FLE.

Nous allons présenter certaines conclusions tirées de notre étude réalisée sur le discours publicitaire français (Habib : 2005) pour montrer l'efficacité de l'utilisation des méthodes de la lexicométrie et de la nouvelle technologie dans les recherches scientifiques sur le vocabulaire de différents domaines. Voici la liste des mots de haute fréquence.

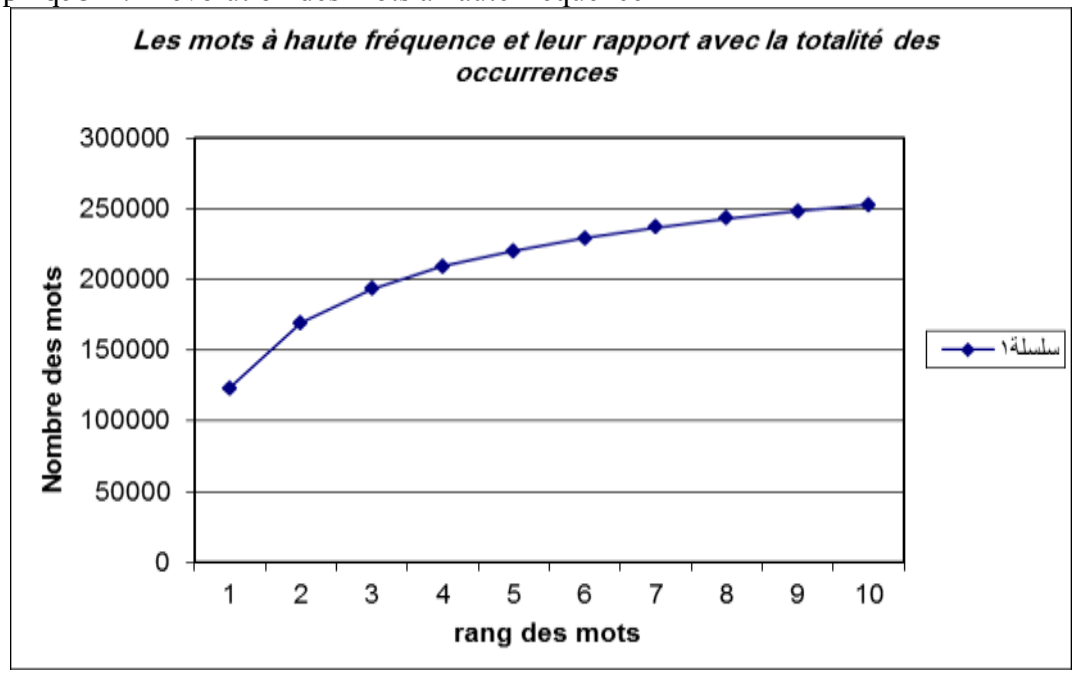
**Liste 1 - Les mots de haute fréquence dans le corpus publicitaire**

| rang | frq | mot |
|---|---|---|
| 1 | 27384 | . |
| 2 | 20683 | , |
| 3 | 16601 | de |
| 4 | 16540 | ’ |
| 5 | 8594 | et |
| 6 | 8064 | la |
| 7 | 6727 | à |
| 8 | 6210 | - |
| 9 | 6162 | le |
| 10 | 6142 | vous |
| 11 | 5907 | d |
| 12 | 5784 | l |
| 13 | 5565 | les |
| 14 | 5029 | en |
| 15 | 4507 | pour |
| 16 | 4467 | des |
| 17 | 4307 | un |
| 18 | 3608 | est |
| 19 | 3608 | : |
| 51 | 1030 | France |
| 52 | 972 | s |
| 53 | 956 | ne |
| 54 | 936 | ? |
| 55 | 932 | 3 |
| 56 | 930 | son |
| 57 | 902 | n |
| 58 | 882 | internet |
| 59 | 872 | ce |
| 60 | 862 | qu |
| 61 | 825 | ® |
| 62 | 785 | sans |
| 63 | 781 | sont |
| 64 | 775 | 4 |
| 65 | 772 | nouvelle |
| 66 | 770 | tous |
| 67 | 755 | aux |
| 68 | 735 | Paris |
| 69 | 733 | on |
| 70 | 725 | ! |

| | | | | | |
|---|---|---|---|---|---|
| 20 | 3590 | une | 71 | 676 | aussi |
| 21 | 2740 | plus | 72 | 656 | se |
| 22 | 2704 | votre | 73 | 639 | ses |
| 23 | 2636 | du | 74 | 635 | nos |
| 24 | 2552 | / | 75 | 619 | offre |
| 25 | 2432 | au | 76 | 607 | monde |
| 26 | 2368 | avec | 77 | 593 | vie |
| 27 | 2272 | sur | 78 | 587 | sa |
| 28 | 2132 | www | 79 | 576 | notre |
| 29 | 2111 | dans | 80 | 560 | nouveau |
| 30 | 2091 | que | 81 | 537 | comme |
| 31 | 1911 | ( | 82 | 531 | 5 |
| 32 | 1907 | ) | 83 | 524 | ans |
| 33 | 1690 | ou | 84 | 524 | 7 |
| 34 | 1598 | qui | 85 | 520 | bien |
| 35 | 1580 | a | 86 | 516 | si |
| 36 | 1495 | f | 87 | 508 | 6 |
| 37 | 1437 | c | 88 | 496 | « |
| 38 | 1419 | fr | 89 | 490 | être |
| 39 | 1369 | par | 90 | 478 | faire |
| 40 | 1321 | * | 91 | 475 | prix |
| 41 | 1165 | • | 92 | 470 | » |
| 42 | 1161 | tout | 93 | 466 | y |
| 43 | 1148 | il | 94 | 458 | e |
| 44 | 1146 | com | 95 | 449 | avez |
| 45 | 1132 | nous | 96 | 448 | toute |
| 46 | 1061 | … | 97 | 444 | temps |
| 47 | 1052 | 1 | 98 | 442 | h |
| 48 | 1048 | pas | 99 | 441 | 9 |
| 49 | 1039 | 2 | 100 | 429 | mais |
| 50 | 1030 | vos | | | |

Selon cette étude et à l'aide du logiciel *Hyperbase*, le rapport de ces cent mots à l'étendue totale du corpus a été calculé; les dix premiers mots constituent le quart du corpus. L'évolution des mots de haute fréquence surtout dans les trente premiers rangs est montrée dans le graphique ci-dessous.

Graphique 1 : L'évolution des mots à haute fréquence




Dans la liste ci-dessus, les cent premiers mots sont classés par ordre de fréquence décroissante. Ces résultats ne sont pas très loin des études quantitatives sur le vocabulaire français en général. Ainsi, un petit nombre de mots correspond à la majeure partie des occurrences, c'est donc la loi générale qui s'impose.

Cette liste a été comparée à d'autres tirées dans le genre littéraire telle que la liste réalisée par Étienne Brunet sur le vocabulaire de Victor Hugo, Rousseau et sur le théâtre classique (Corneille, Molière et Racine) (1981) par le même auteur. L'étude quantitative a donné une idée de la longueur des phrases dans le genre littéraire et contribué à fournir des renseignements sur la structure du vocabulaire : le discours de Rousseau privilégie la première personne alors que le genre théâtral privilégie la deuxième personne au pluriel.

La forme la plus fréquente dans la liste ci-dessus est le point avec 27384 occurrences, puis la virgule avec 20683 occurrences. La présence du point au premier rang, dans la liste de hautes fréquences, marque et caractérise le genre publicitaire. Cette présence est due à l'abondance des adresses électroniques, des abréviations et elle peut donner une idée sur la longueur de la phrase publicitaire. Puis vient la préposition *de* avec 16601 occurrences, suivie par *et* et *la* avec respectivement 8594 et 8064 occurrences. D'ailleurs, les dix premiers mots sont exclusivement des mots grammaticaux ou mots outils. Ceci n'est pas étonnant, car selon Guiraud ( 1959: 19) les mots les plus fréquents d'un texte sont les mots les plus courts, les mots les plus anciens, les mots les plus simples morphologiquement et les mots les plus étendus sémantiquement.

Nous pouvons aussi remarquer qu'il y a plusieurs formes très courtes parmi les cent premiers mots. Ce sont des "mots outils" comme par exemple, dans l'ordre de la liste, les **prépositions** (*de, à, d', en, pour, avec, dans, sur, par*), les articles ou **pronoms** (*la, le, l', les, un, une, qui*), les **adverbes** (*plus, tout, pas, ne, tous, aussi, bien*) et les **conjonctions**/**pronoms** (*et, que, ou*). Des verbes très importants *faire, être, des* **formes** de **verbes auxiliaires** y figurent aussi ; par exemple *est, a, sont et avez,* vont avec les pronoms *il, on, elle, ils, vous* et le **démonstratif** *ce* dans ses différentes formes. Les **possessifs** *son, ses et sa* s'inscrivent aussi dans le domaine de la troisième personne : dans la publicité on

parle toujours du produit ou du service. On trouve aussi des **pronoms personnels** et des **adjectifs possessifs** de la première et de la deuxième personne du pluriel, dans l'ordre de la liste *vous, votre, nous, vos, nos, notre*. L'importance de ces mots dans le discours publicitaire est incontestable

L'apparition de la suite des trois W, *www,* dans cette liste, contribue à souligner l'importance que le publicitaire attribue à l'adresse électronique du produit ou de la marque. Cette adresse est devenue un élément indispensable dans la plupart des annonces, surtout avec l'évolution rapide des moyens de communication de haute technologie. Son existence dans le texte publicitaire participe à évoquer la modernité.

Le rang 61 de cette liste nous présente le signe ®, symbole de la marque déposée; cette présence explique le fait qu'on donne beaucoup d'importance à la marque.

Plus loin dans la liste des hautes fréquences, nous distinguons des mots de signification plus "étendue". Il y a donc des **substantifs ou noms communs** comme *offre, monde, vie, prix* et *temps ;* des **adjectifs** : *nouvelle*, *nouveau*[1]*;* des **noms propres** tels que *France*, *Internet*.

Pour tenir compte des structures thématiques du corpus, nous avons adopté une méthode qui crée des regroupements autour d'un mot-pôle s'appuyant sur la récurrence des mots qui l'entourent : « Les mots entretiennent entre eux des relations (d'association, de substitution voire d'exclusion) plus ou moins fortes (…) » (Labbé, 1990 : 166. )

A l'aide du logiciel *Hyperbase,* nous avons déterminé les mots importants et précisé les thèmes caractéristiques du genre publicitaire. Les mots prédominants tels que *France*, *Paris*, *Internet*, *offre*, le *monde*, la *vie*, le *prix* et le *temps* figurant dans la liste des hautes fréquences ci-dessus ont été sélectionnés par le critère de fréquence.

Tout en étant des termes de signification plutôt large, ces mots sont aussi très révélateurs du vocabulaire publicitaire parce qu'ils sont porteurs de jugements sur la valeur, la disponibilité et l'actualité des produits et des marques. L'environnement de chaque mot a été étudié individuellement ; le mot est considéré comme un mot-pôle, les formes les plus attirées par ce mot sont aussi étudiés.

**Etude d'une forme**

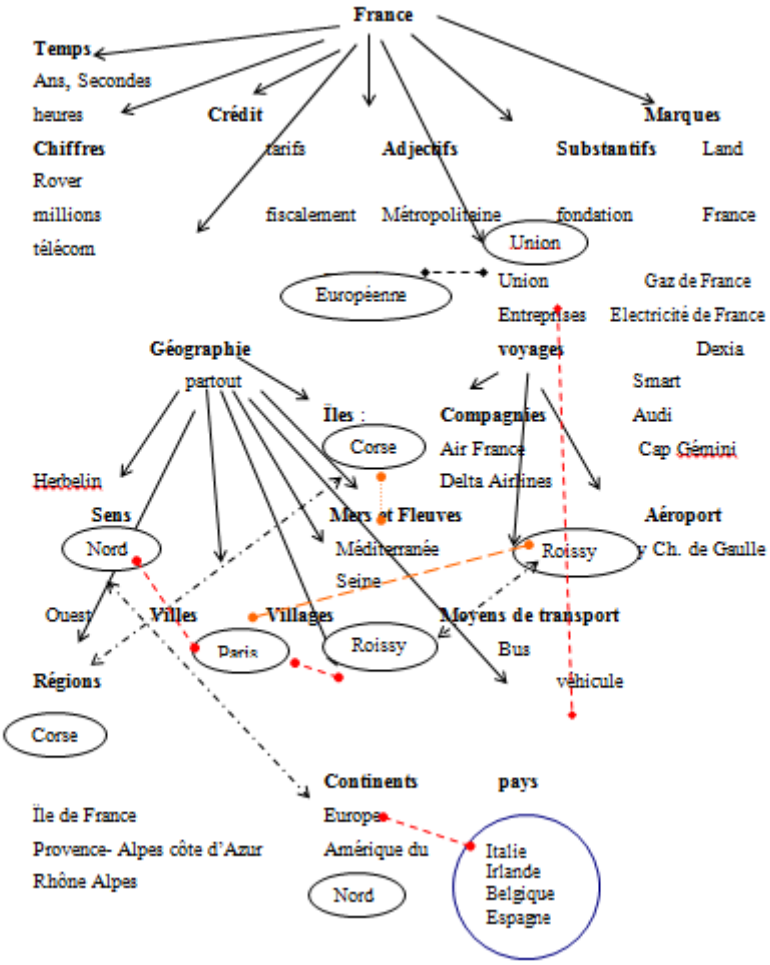
La forme *France* compte 1030 occurrences et représente un mot thème dans le corpus publicitaire. À l'aide du logiciel, chaque occurrence d'une forme est montrée dans son contexte. En comparant la forme *France* à son entourage, on obtient un fichier constitué par les mots effectuant l'environnement de cette forme. Ce sont les mots qui tournent autour du mot-pôle *France*. Nous montrons, ci-dessous, seulement une partie, faute de place, de la liste hiérarchique de l'environnement de la forme *France.*

**[1] Ces adjectifs contribuent à montrer la valeur positive du produit ; ils évoquent la modernité et l'évolution.**

Liste 2 - Environnement thématique (ordre hiérarchique)

| Ecart | Corpus | Extrait | Mot |
|---|---|---|---|
| 120.00 | 1030 | 1022 | FRANCE |
| 24.85 | 332 | 134 | ROVER |
| 23.15 | 199 | 94 | PARTOUT |
| 22.80 | 167 | 84 | TÉLÉCOM |
| 21.39 | 228 | 95 | LAND |
| 20.62 | 32 | 31 | MÉTROPOLITAINE |
| 19.23 | 26 | 26 | BELGACOM |
| 19.21 | 127 | 62 | APPELS |
| 18.83 | 120 | 59 | SMART |
| 18.81 | 29 | 27 | NEUFS |
| 18.70 | 261 | 92 | AIR |
| 18.68 | 78 | 46 | GAZ |
| 17.24 | 23 | 22 | COMMERCIALISE |
| 16.80 | 259 | 84 | HORS |
| 16.44 | 122 | 53 | DÉPARTEMENT |
| 16.14 | 524 | 126 | ANS |
| 15.59 | 111 | 48 | ENTRETIEN |
| 14.95 | 149 | 55 | CONTRAT |
| 14.90 | 136 | 52 | TARIFS |
| 14.31 | 5029 | 582 | EN |
| 14.09 | 18 | 16 | FABRIQUÉE |
| 13.65 | 37 | 23 | GÉNÉRALES |
| 12.86 | 69 | 31 | APPELER |
| 12.54 | 84 | 34 | TELE |
| 11.90 | 12 | 11 | PÉNALISER |
| 11.67 | 51 | 24 | CENTER |
| 11.56 | 48 | 23 | CORSE |
| 11.47 | 86 | 32 | TÉLÉPHONER |
| 11.46 | 45 | 22 | CONSIDÉRÉS |
| 11.31 | 9 | 9 | MÉTROPOLITAINE |
| 11.05 | 27 | 16 | COULTRE |
| 10.66 | 8 | 8 | SOUSCRIPTEUR |
| 10.66 | 8 | 8 | FISCALEMENT |
| 10.66 | 8 | 8 | ACHETÉES |
| 10.65 | 17 | 12 | RECRUTE |
| 10.63 | 32 | 17 | TÉLÉVISION |
| 10.51 | 23 | 14 | FONDATION |
| 10.32 | 300 | 64 | MN |
| 10.11 | 42 | 19 | RÉDUCTION |
| 10.05 | 59 | 23 | SECONDES |
| 10.04 | 16 | 11 | TARIFAIRES |
| 10.04 | 16 | 11 | INVESTISSEUR |
| 9.98 | 7 | 7 | RESIDENTIEL |
| 9.98 | 7 | 7 | ITALIE |
| 7.76 | 41 | 15 | DÉPARTEMENTS |
| 7.72 | 56 | 18 | PARTENAIRES |
| 7.70 | 84 | 23 | DEMANDEZ |
| 7.64 | 11 | 7 | TRAVAILLEZ |
| 7.33 | 77 | 21 | PERSONNES |
| 7.31 | 15 | 8 | DELTA |
| 7.22 | 31 | 12 | INITIALE |
| 7.19 | 619 | 85 | OFFRE |
| 7.08 | 57 | 17 | COMMERCIAL |
| 7.05 | 283 | 48 | CLIENTS |
| 7.01 | 47 | 15 | FACTURE |
| 7.01 | 16 | 8 | INTERMÉDIAIRE |
| 6.93 | 237 | 42 | PREMIER |
| 9.97 | 51 | 21 | SPÉCIAUX |
| 9.97 | 25 | 14 | BUS |
| 9.97 | 9 | 8 | CONSERVÉES |
| 9.95 | 19 | 12 | BELGIQUE |
| 9.94 | 22 | 13 | JEARGER |
| 9.84 | 52 | 21 | AVANTAGE |
| 9.67 | 23 | 13 | IRLANDE |
| 9.51 | 46 | 19 | NORD |
| 9.43 | 147 | 38 | ASSISTANCE |
| 9.37 | 10 | 8 | GRATUITÉ |
| 9.24 | 6 | 6 | PÉRIODICITÉS |
| 9.24 | 6 | 6 | INGRÉDIENTS |
| 8.85 | 11 | 8 | INSTITUTIONNEL |
| 8.85 | 11 | 8 | ATTRACTIF |
| 8.83 | 23 | 12 | COMPÉTITIFS |
| 8.77 | 16601 | 1371 | DE |
| 8.66 | 240 | 49 | GRATUIT |
| 8.45 | 7 | 6 | COMPRENNENT |
| 8.43 | 5 | 5 | BRAUN |
| 8.36 | 25 | 12 | ÎLE |
| 8.35 | 15 | 9 | NEUVES |
| 8.31 | 29 | 13 | GAULLE |
| 8.28 | 51 | 18 | ÉLECTRICITÉ |
| 8.21 | 83 | 24 | MOBILES |
| 8.16 | 62 | 20 | CENTIMES |
| 7.98 | 86 | 24 | AGENCE |
| 7.91 | 87 | 24 | CAPITAL |
| 7.81 | 8 | 6 | PRÉCONISÉES |
| 5.76 | 17 | 7 | MODALITÉS |
| 5.61 | 27 | 9 | RESPONSABILITE |
| 5.54 | 50 | 13 | INVESTISSEMENT |
| 5.54 | 10 | 5 | ADMINISTRATION |
| 5.52 | 137 | 25 | GRANDES |
| 5.51 | 91 | 19 | AGENCES |
| 5.48 | 14 | 6 | OUEST |
| 5.46 | 23 | 8 | DÉBIT |
| 5.46 | 23 | 8 | BÉNÉFICIENT |
| 5.22 | 15 | 6 | PROVENCE |
| 5.14 | 90 | 18 | AUDI |
| 5.11 | 114 | 21 | COMMUNICATION |
| 5.06 | 49 | 12 | RENSEIGNEMENT |
| 5.05 | 316 | 43 | ENTREPRISE |
| 5.05 | 31 | 9 | NIVEAUX |
| 4.49 | 8594 | 668 | ET |
| 4.46 | 36 | 9 | ÉTRANGER |
| 4.41 | 57 | 12 | OUVERTURE |
| 4.40 | 19 | 6 | MOBICARTE |
| 4.40 | 14 | 5 | NATIONALES |
| 4.37 | 105 | 18 | FRANÇAIS |
| 4.31 | 115 | 19 | ÉCONOMIE |
| 4.27 | 59 | 12 | ÉQUIPES |
| 4.26 | 38 | 9 | TÉLÉ |
| 4.24 | 91 | 16 | MICROSOFT |
| 4.23 | 20 | 6 | COUVERTURE |
| 4.18 | 15 | 5 | VÉHICULES |
| 4.09 | 54 | 11 | OPÉRATEUR |
| 4.07 | 21 | 6 | VÔTRE |
| 3.68 | 735 | 73 | PARIS |

Le thème *France* dans la publicité peut être défini à partir de son usage dans le corpus. En fait l'étude des relations existant entre les mots enrichit le sens de ces derniers. L'étude du mot thème *France* nous a révélé le fait qu'il existe une relation entre la Corse, par exemple, la France et la Méditerranée. Le mot *Nord* est associé à Amérique du Nord, Nord de la France,…
Ces rapports sont représentés dans le schéma suivant :




**L'enseignement du vocabulaire**

La recherche dans les domaines de la linguistique et de la psycholinguistique fournit quelques postulats fondamentaux. Selon ces postulats, les mots sont en rapport sémantique les uns avec les autres, ils sont organisés en réseaux d'association dans le cerveau des individus et ils sont utilisés, dans le discours, en suivant des modèles de comportement verbal généralisés. La psycholinguistique repère dans l'acquisition du vocabulaire une certaine évolution, rapide au début de l'acquisition puis il se ralentit au fur et à mesure qu'on avance (voir Graphique 1).
L'organisation sémantique du vocabulaire facilite son "intégration" et sa "mémorisation" .(Tréville et Duquette, 1996 : 27) De plus, l'ancrage cognitif joue un rôle important dans le processus d'acquisition. Le vocabulaire s'acquiert par étapes intermédiaires, par une organisation progressive : « s'il y a association, le mot nouveau

s'intègre dans des types de réseaux divers syntagmatiques, paradigmatiques, etc.» (1996 : 57).

Pour les auteurs du *français fondamental*[1], la notion de "degré de disponibilité", qui correspond à la présence plus ou moins immédiate de ces mots dans notre mémoire, offre un intérêt à la fois linguistique, psychologique et pédagogique. Pour déterminer leur degré de disponibilité, les auteurs ont eu recours à la méthode des centres d'intérêt. Ils ont signalé que le vocabulaire s'emboîte et que seule, une combinaison du vocabulaire de fréquence et du vocabulaire "disponible" donne le vocabulaire nécessaire. Des études sur le vocabulaire disponible ont été réalisées dont celle publiée par Rodriguez (2007 : 141-158).

Guiraud ( 1959 : 93) souligne qu'un très petit nombre de mots convenablement choisis « couvrent par leur répétition la presque totalité de n'importe quel texte ». Le texte publicitaire est considéré comme un outil pour l'enseignement du français, en particulier le vocabulaire. Le contenu du texte va déterminer le choix des unités lexicales à enseigner.

La fréquence, ainsi que Galisson et Coste (1976 : 242) le montrent, critère fondamental pour la sélection du vocabulaire, est étroitement liée aux notions d'utilité, de rentabilité et de disponibilité.

En s'inspirant de la méthode du *Français Fondamental*, Picoche (1993 : 57) précise que le vocabulaire qui sert de base pour l'enseignement du français, c'est celui qui figure en tête des listes de fréquences et que « tout le monde croit connaître.»

**Conclusion**

Nous avons vu que la statistique des fréquences du vocabulaire offre un critère objectif, permettant de déterminer scientifiquement les mots les plus usuels. Le vocabulaire à enseigner sera déterminé selon des critères tels que la fréquence, la disponibilité et la rentabilité. Ainsi, la progression pédagogique ne se fonde plus sur *l'intuition*, mais sur la *connaissance plus scientifique du lexique et de la grammaire du français*. (Genouvrier et Peytard, 1970 : 203).

Le logiciel proposé, *Hyperbase*, est donc un outil de recherche permettant de mener des études sur des corpus volumineux et donnant la possibilité d'en dégager les spécificités. Grâce à cette méthode de recherche, on a attribué aussi aux chercheurs dans le domaine de l'enseignement du français la possibilité de délimiter et de choisir efficacement le vocabulaire spécifique nécessaire à l'enseignement du français.

**Bibliographie**

Brunet, É. *Le vocabulaire français de 1789 à nos jours, d'après les données du Trésor de la langue française*, Slatkine - Champion, Genève – Paris, 1981.

Brunet, É. *Hyperbase, Manuel de référence*, version standard 8.0 et 9.0 janvier 2011.

Caron, J. *Précis de psycholinguistique*, Paris, P. U. F., 1992.

GALISSON, R. COSTE, D. *Dictionnaire de didactique des langues*, Paris, Hachette, 1976.

Guiraud, P. *Caractères statistiques du vocabulaire*, Paris, P.U.F., 1954.

Guiraud, P. *Problèmes et Méthodes de la Statistique Linguistique*. Holland, Dordrecht, 1959.

Habib, I. *Création et exploitation des données d'un corpus publicitaire de langue française*. Thèse de doctorat sous la direction de C. Maciel, Nice, 2005.

LABBE, D. *Le vocabulaire de François Mitterrand*, Paris : Presse de la Fondation Nationale des Sciences Politiques, 1990.

Lebart, L. et Salem, A. *Statistique textuelle*, Paris, Dunod, 1994.

[1] **Nous suivons, dans ce paragraphe, le commentaire proposé par Emile Genouvrier et Jean Peytard dans *Linguistique et enseignement du français*, Librairie Larousse 1970 : pp. 203 -204.**

Maciel, C.A.A. *Etude du vocabulaire de quatre manuels d'espagnol. Analyse statistique*. Paris, Editions Champion, 1997.
Marchand, M. *L'analyse du Discours Assistée par Ordinateur. Concepts, Méthodes, Outils*. Paris, Armand Colin, 1998.
Martin, E. *Reconnaissance de contextes thématiques dans un corpus textuel, éléments de lexico-sémantique*, Paris, CNRS-INaLF, Didier Erudition, 1993.
Muller, Ch. « Le mot, unité de texte et unité de lexique » , in *Langue française et linguistique quantitative*. *Recueil d'articles*, Éd. Slatkine, Genève, 1979, pp.125-143.
Muller, Ch. « Statistique lexicale et théorie statistique » in *Langue française et linguistique quantitative*. Recueil d'articles. Éd. Slatkine. Genève. 1979.
Muller, Ch. *Initiation aux méthodes de la statistique linguistique*, Paris, Hachette, 1973.
Muller, Ch. *Principes et méthodes de statistique lexicale*. Paris, Hachette, 1977.
Picoche, J. *Didactique du vocabulaire français*, Paris, Nathan, 1993.
Picoche, J. *Précis de lexicologie française*, Paris, Nathan, 1992.
Robert, J.-P. *Dictionnaire pratique de didactique du FLE,* Paris, Ophrys, 2002.
Rodriguez, L . "*Ruralité et acquisition lexicale au Manitoba: le vocabulaire disponible dans les écoles Saint-Eustache (milieu rural) et Provencher (milieu urbain)*", in cahiers franco-canadiens de l'ouest, vol. 19, no 2, 2007, p. 141-158.
Ruggia, S. "*La lecture contrôlée et assistée par l'analyse statistique des données textuelles.*", in le Français dans le monde, 2021, pp. 84-100.
Ruggia, S. "*Le Deep learning au service de la didactique du FLE " XIX*$^{è}$ *Congrès National de l'AMIFRAM,* 2018, Mexique.
Ruggia, S. "*Caractériser un texte en français : les passages-clés des niveaux A1 et A2 du CECRL*", 15èmes Journées internationales d'Analyse statistique des Données Textuelles, Toulouse, France, Jun 2020. 11 p.
Ruggia, S. "*Les niveaux de langue du CECRL : de la prédiction à l'analyse descriptive grâce au deep learning et à l'analyse des données textuelles.*" Trujillo, Pérou Oct., 2019.
Tréville, M.-C. Duquette et L. *Enseigner le vocabulaire en classe de langue*, Paris, Hachette F.L.E., 1996.
Wagner, R. L. *Les Vocabulaires français*, I, Paris, Didier, 1967.