HAL open science

Usages du Dictionnaire Électronique des Synonymes (DES) du CRISCO : focus sur les mots inexistants

Laurette Chardon

To cite this version:

Laurette Chardon. Usages du Dictionnaire Électronique des Synonymes (DES) du CRISCO : focus sur les mots inexistants. 2022. halshs-03606075

HAL Id: halshs-03606075

https://shs.hal.science/halshs-03606075v1

Submitted on 7 Apr 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Hypotheses

OpenEdition Search
Tout OpenEdition

Usages du Dictionnaire Électronique des Synonymes (DES) du CRISCO : focus sur les mots inexistants

PAR LAURETTE CHARDON · 08/03/2022

Introduction

Une première étude a déjà été publiée sur HAL en aout 2021 intitulée Usages du Dictionnaire Électronique des Synonymes (DES) du CRISCO. Cette étude concernait les accès au DES sur les années 2019 et 2020 enregistrés dans les fichiers de logs du DES desquels ont été extraits les mots les plus recherchés.

Pour mémoire, voici un résumé des points notables :

- Les figures 1 et 2 sur les **requêtes journalières** respectivement en 2019 et 2020 et les figures 3 et 4 sur les **mots journaliers** en 2019 et 2020 nous montrent que le DES est utilisé principalement en semaine. Le week-end, la fréquentation est moindre. Il s'agit donc d'une utilisation principalement professionnelle.
- Un nombre non négligeable (en bleu ciel) des requêtes concerne des mots inexistants dans le DES.
- Avec la figure 5 représentant les moyennes, nous voyons que **le DES enregistre entre 33.259 et 43.502 mots différents et valides recherchés par jour, représentant de 135.349 à 262.924 requêtes.**
- La figure 14 nous indique qu'un mot valide est recherché de 4 à 7 fois par jour en moyenne.
- La figure 16 donne cette moyenne par mois : de 100 à 150, ce qui est cohérent avec les valeurs par jour mais on constate de grandes disparités puisque seuls 10% des mots (soit 5450 mots) ont été demandés plus de 5947 fois par mois (environ 60 fois plus que la moyenne) et que 10% ont été demandés moins de 20 fois (environ 6 fois moins que la moyenne).

L'ensemble des graphiques est accessible sur https://git.unicaen.fr/crisco-des-public/MotsLesPlusRecherches.

Cet article se propose d'approfondir les requêtes portant sur les mots inexistants dans le DES. Tout d'abord nous allons regarder la proportion de ces requêtes sur les années 2019, 2020 puis également en 2021 en essayant de répondre aux questions suivantes : est-ce des mots inexistants du à la casse ? du au manque d'accents ? est-ce des mots bien formés mais pas encore saisis dans le DES (des mots récents ou tout simplement des mots qui n'ont pas de synonymes) ?

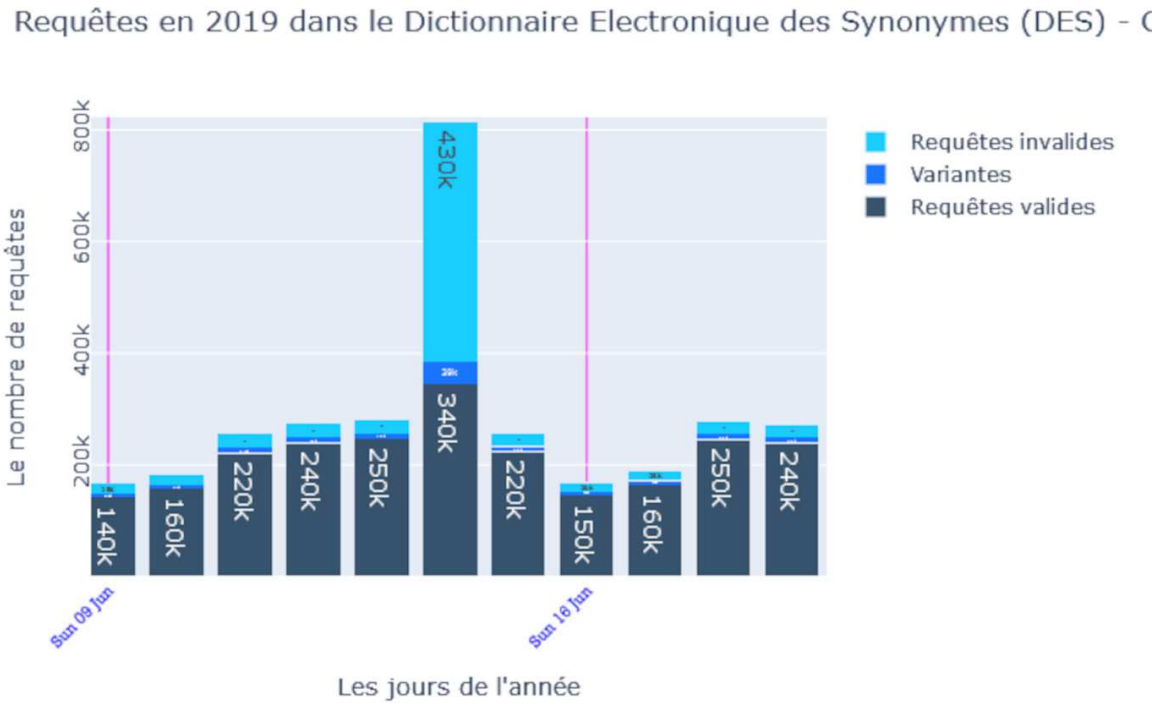
Proportion de requêtes et de mots invalides et fréquences

En **2019**, il a eu **10% de requêtes sur des mots inexistants** soit 9.998.447 sur un total de 95.365.871 requêtes avec une amplitude allant 5% à 52% selon les jours.

En **2020**, il y a eu **11% de requêtes sur des mots inexistants** soit 8.362.800 sur un total de 75.715.755 requêtes avec une amplitude de 5% à 47% selon les jours.

En **2021**, il y a eu **10% de requêtes sur des mots inexistants** soit 6.706.790 sur un total de 65.568.077 requêtes avec une amplitude importante allant de 5,8% à 47 % selon les jours.

Par exemple l'amplitude maximale de 52 % en 2019 correspond au vendredi 14 juin avec plus de 430.000 requêtes invalides comme le montre la figure ci-dessous (extraite de la figure 1) :



Pour 2020, l'amplitude maximale de 47 % correspond au lundi 21 septembre.

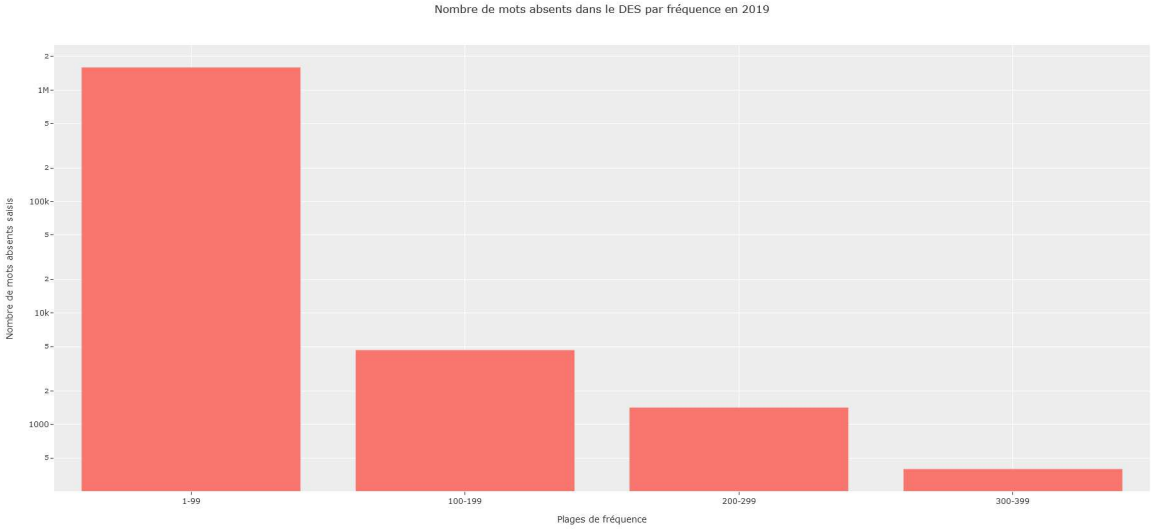
Il s'agit très probablement dans ces cas de requêtes automatiques.

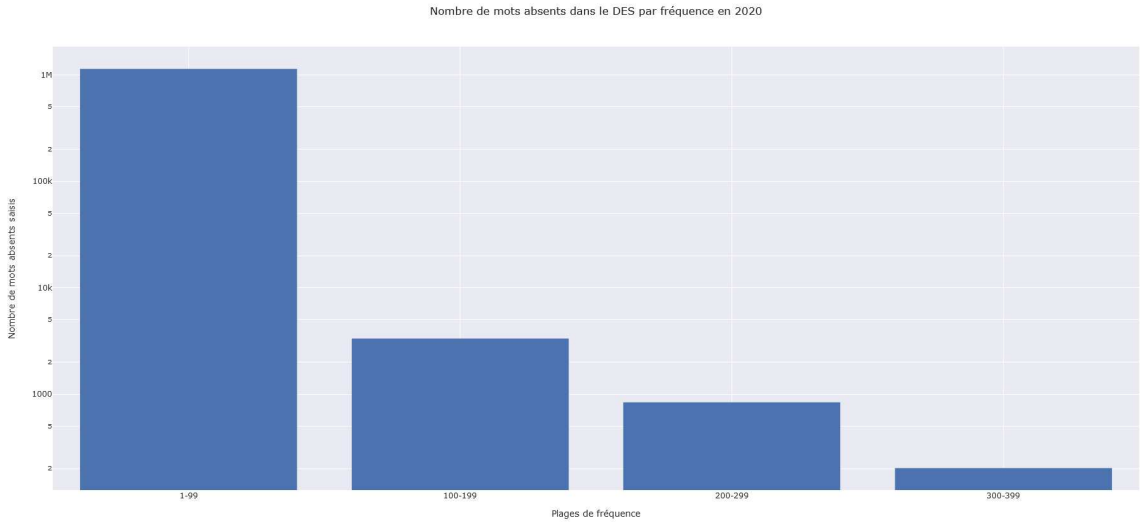
Ensuite, nous pouvons nous demander si ces mots inexistants sont demandés plusieurs fois. Si c'est le cas, quelle est leur répartition en fonction de leur fréquence ?

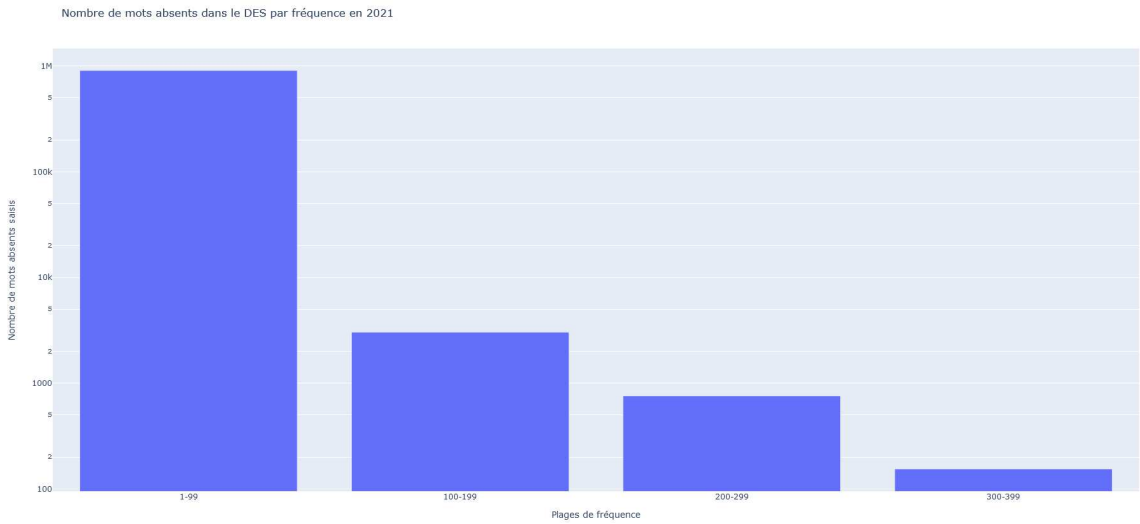
En **2019**, sur les 1.605.506 mots inexistants différents, 911.681 ont été **saisis une seule fois** soit **57%**.

En **2020 et 2021**, les proportions sont du même ordre puisque nous avons respectivement 1.145.053 mots inexistants différents dont 718.172 saisis une seule fois soit **63 %** et 915.398 mots inexistants différents dont 545.640 saisis une seule fois soit près de **60%**.

Nous pouvons aussi visualiser le nombre de mots absents saisis selon quatre plages de fréquence ci-dessous :







Une tendance globale se dessine : sur les 3 années, une très grande majorité de mots absents sont peu demandés, **seuls 400 mots en 2019, 203 en 2020 et 154 en 2021 ont été saisies plus de 300 fois, soit en moyenne une fois par jour.**

Avant de regarder d'un peu plus près ces derniers dans le paragraphe suivant, terminons en citant quelques exemples de mots demandés une seule fois dans l'année. Nous avons les différents cas de figures suivants :

- Des fautes de frappe : *enthousiasmee, enrgistrement, rigoureusemnent, rofondément, romulguer, ronfleemnt, dvouement*
- Des verbes conjugués : *rigolait, duvetassions, s'aspergeassent* …
- Des expressions et des suites de mots : *rompre la glace, durcissement par apport, réseau téléphonique public commuté, depuis ma plus tendre enfance* …
- Avec le déterminant : *la potion, la radio, la spirale* …
- Des suites de lettres sans aucun sens : *looooooooooooooooooooooooooooooookkkkkkkkkk, lucre?1610214360470, acquÃƒÂƒÃ‚©rir*…
- Caractères spéciales, langues étrangères : ***רקומקקמא*** , *ᎤᏩᏕᏥᏩᏗ, present value of new business premiums* …
- être + compléments : *être induit par, être insulté, être habilité, être exemplaire, être dépendant de* …
- etc

Si certains d'entre vous veulent approfondir cette partie (ou bien s'amuser), les graphiques ainsi que les fichiers "MotsInexistantsParAnnee.csv" sont consultables sur le serveur git de l'Université : https://git.unicaen.fr/crisco-des-public/MotsLesPlusRecherches

Quels sont les mots inexistants les plus fréquemment saisis ?

Nous allons nous focaliser sur les mots inexistants demandés plusieurs fois dans l'année : le maximum est de 360 ce qui correspond en moyenne à une requête par jour. Que peuvent signifier

ces mots qui sont saisis de façon récurrente ?

Pour le savoir, prenons quelques exemples parmi ceux les plus fréquemment saisis dans le tableau ci-dessous.

Entrée saisie	2019 : position (nombre requêtes)	2020 : position (nombre requêtes)	2021 : position (nombre requêtes)
Ainsi	6 (360)	42 (345)	15 (358)
Joie	19 (358)	94 (329)	73 (328)
Amour	22 (358)	29 (351)	55 (337)
intéret	46 (353)	38 (347)	30 (351)
focus	2 (360)	6 (361)	2 (364)
différente	8 (360)	12 (359)	14 (359)
particulière	39 (355)	22 (353)	17 (358)
compétences	11 (360)	10 (360)	9 (363)
caractéristiques	33 (356)	41 (345)	40 (357)
faire l'objet	1 (361)	14 (358)	34 (349)
il s'agit	9 (360)	1 (365)	32 (350)
il s'agit de	372 (303)	213 (297)	202 (290)
s'appuyer	10 (360)	5 (362)	1 (365)

Nous pouvons regrouper ces exemples selon plusieurs critères :

- Les mots qui commencent par une majuscule comme *Ainsi, Joie et Amour*. *Ainsi* arrive en 6ème position des mots inexistants les plus demandés avec 360 requêtes en 2019, en 42ème position avec 345 requêtes en 2020 et en 15ème position avec 358 requêtes en 2021. Ces types de mots montrent clairement qu'une amélioration serait à prévoir dans le DES à savoir **retranscrire le mot avec des lettres en minuscules avant d'effectuer la recherche dans le DES.**
- Ceux auxquels il manque un ou plusieurs accents comme *intéret*. Il pourrait être **utile d'ajouter les variantes** de type "correction orthographique" pour tous ces mots souvent saisis avec des accents manquants.
- Les **formes féminines absentes** comme *différente* ou *particulière*. Bien que certaines d'entre elles soient déjà enregistrées dans le DES (comme *spectatrice, policière*), il en reste de nombreuses autres à mentionner. Les mises à jour mensuelles du DES en ajoutent régulièrement et c'est une tâche à poursuivre.
- Des formes au pluriel : *compétences, caractéristiques*. **L'ajout de variantes plurielles**

pourrait être utile pour ces formes souvent saisies.

- Nous trouvons également des expressions comme *il s'agit* ou *il s'agit de*. *S'agir* est dans le DES avec des synonymes comme *être question, convenir* mais peut-être faudrait-il réfléchir soit à **créer de nouvelles entrées** pour ces expressions si des synonymes sont possibles, soit les **ajouter comme variantes** de *s'agir*.
- Enfin, de façon étonnante, nous voyons que *s'appuyer* n'a pas d'entrée dans le DES. Par contre *s'appuyer sur* existe. Une brève recherche dans les dictionnaires montrent que très souvent le verbe *s'appuyer* est utilisé avec la préposition *sur* mais il existe des contextes où *s'appuyer* s'utilise seul ou avec une autre préposition : *s'appuyer fortement à un mur pour exercer une poussée*, *s'appuyer d'un côté*. Des **corrections** seront à prévoir dans le DES pour en tenir compte.

Nous voyons bien, au travers de ces quelques exemples, que cette étude des mots inexistants dans le DES (en commençant par ceux qui sont souvent demandés) nous sera très profitable pour améliorer les recherches des nombreux utilisateurs en leur évitant de saisir plusieurs fois une entrée avant d'obtenir le résultat souhaité. Cela ouvrera de nouvelles pistes de corrections et d'évolution à ce produit déjà visité plusieurs dizaines de milliers de fois par jour et mis à jour mensuellement depuis de nombreuses années.

Cite this article as: Laurette Chardon, "Usages du Dictionnaire Électronique des Synonymes (DES) du CRISCO : focus sur les mots inexistants," in *Le carnet de la MRSH*, 08/03/2022, https://mrsh.hypotheses.org/5578.

Rechercher dans OpenEdition Search
Vous allez être redirigé vers OpenEdition Search
Expression ou mot-clé
Dans tout OpenEdition
Dans Le carnet de la MRSH
Rechercher