# Outils Linguistiques

Fassi Valentine

# Étude des contextes d'apparitions des occurrences « sorte de » et « espèce de » avec l'outil Sketch Engine

Dans le contexte du cours d'outils linguistiques, nous avons mené une étude de cas sur les occurrences « sorte de » et « espèce de » afin de montrer que l'analyse sur corpus permet de donner des indications précises sur les fonctions sémantiques de certaines unités. Nous avons étudier un phénomène spécifique pour que des informations générales en ressortent.
L'objectif de notre étude était d'étudier les similarités et les différences de ces deux occurrences qui sont quasi-synonymes. Ces unités ont elles vraiment une relation synonymique ? Quelles sont les différences sémantiques entre elles ?
Nous avons également voulu savoir si certaines hypothèses que nous pouvons faire d'instinct ont la même profondeur que les hypothèses que l'on fait en s'appuyant sur les données, on prouve la nécessité de l'analyse sur corpus pour l'analyse d'un phénomène sémantique. Les corpus permettent également de voir la langue dans sa dynamique, dans son contexte.

Dans une optique de sémantique distributionnelle, nous nous sommes appuyés sur l' hypothèse distributionnaliste de Harris pour réaliser cette étude. Selon cette hypothèse, il y a une corrélation entre la distribution d'un élément et sa signification (le contexte).

Lors de la réalisation de notre étude, nous avons travaillé avec l'outil informatique Sketch Engine sur le corpus frenchtenten2012. L'avantage de ce corpus réside en le fait qu'il contienne plusieurs milliards de mots et qu'il possède la lemmatisation, le taging et la tokenisation.

## 1ère étape : Définition des occurrences et formulation des hypothèses d'intuition :

• Étymologie des occurrences

**Sorte** *Subst.* ***1. a)*** *ca 1220 « groupe de gens, compagnie, société » (Auberon, éd. J. Subrenat, 501), en a. et m. fr.;* ***b)*** *1458 « condition, rang d'une personne » roys de noble sorte (Arnoul Greban, Mystère de la Passion, éd. O. Jodogne, 5806); cf. estre de ma sorte (Id., ibid., 10898); 1558 estre de bonne sorte (B. des Périers, Nlles récréations et joyeux devis, éd. Kr. Kasprzyk, 41, p. 175);* ***2.*** *1327 « catégorie d'êtres animés ou de choses, espèce, genre » (Watriquet de Couvin, Li Tournoi des Dames, 692, éd. A. Scheler, p. 253); 1723 de première sorte « de qualité supérieure » (Savary); 1803 une beauté de la première sorte (Chateaubr., Génie, t. 2, p. 280);* ***3.*** *ca 1485 « manière de faire une chose, façon » (Myst. Vieux Testament, 13470, éd. J. de Rothschild, t. 2, p. 193: dites moi la* ***sorte****Comme vostre père se porte); 1549 de la bonne sorte « de la bonne manière » (Est.); 1668 faire en sorte que (La Fontaine, Epitre a Monseigneur le Dauphin ds Œuvres, éd. H. Regnier, t. 1, p. 4); 1678 faire en sorte de (Id., Le Florentin, sc. VIII, 395, ibid., t. 7, p. 427);* ***4.*** *1664 une sorte de « ce qu'on ne peut qualifier exactement et qu'on rapproche d'autre chose » (Corneille, Rodogune [Epitre dédicatoire] à Monseigneur le Prince);* ***5.*** *1689 quelque sorte de temps « pour un certain temps » (Sévigné, Corresp., éd. R. Duchêne, t. 3, p. 627);* ***6.*** *1723 impr. (Fertel,Imprimerie ds IGLF: nous comptons dix-neuf* ***sortes*** *ou corps de caractères);* ***7.*** *1765 pharm. manne en sorte(Encyclop. t. 10, s.v. manne, p. 45a).*

**Espèce** ***1. a)*** *xiies. vraie espesse « signe, révélation (de Dieu) » (Alexis, ms. S, éd. Paris et Pannier, 1298), attest. isolée;* ***b)*** *1314 « apparence » (H. de Mondeville, Chirurgie, éd. A. Bos, § 210 et 215);* ***c)*** *1545 spéc. théol. « présentation matérielle de l'eucharistie, apparence (le pain) sous laquelle est offert le corps du Christ » (Calvin,Institution chrétienne, éd. J.-D. Benoît, IV, XVII, 13); av. 1656*

*« id. » au plur. (A. Arnauld, Lettre citée par Pascal,Provinciales, XVI, éd. L. Lafuma, Œuvres, p. 447);* ***2. a)*** *1269-78 « catégorie d'êtres vivants du même type » ici « genre humain » (J. de Meun, Rose, éd. F. Lecoy, 6939);* ***b)*** *1314 plus gén. « catégorie, sorte » (H. de Mondeville, op. cit., 1826); d'où* ***3.*** *1587 espèce de « sorte de » ceste espece de magiciens (Lanoue, Discours politiques et militaires,136 ds Littré); 1725 report de l'accord de l'article sur le nom complément un espèce de Dictionnaire (Grandval, Le Vice puni ou Cartouche, p. IV).* ***4.*** *1670-81 dr. (O. Patru, Plaidoié, 9 ds Rich.). Empr. au lat. class. species « vue, regard » d'où « apparence, aspect, type, cas particulier (terme de dr.), catégorie » et spéc. en lat. chrét. « matière d'un sacrement (en parlant du sel du baptême) »*

• Définition des occurrences

SORTE → *Caractéristique commune qui définit, à l'intérieur d'un ensemble d'êtres ou de choses de même nature, un groupe, un genre, un type identique.* (LAROUSSE)
ESPÈCE → *Caractéristique commune qui définit, à l'intérieur d'un ensemble d'êtres ou de choses de même nature, un groupe, un genre, un type identique (LAROUSSE)*

• Premières hypothèses

Intuitivement, nous avons émis des hypothèses sur les différences et les ressemblance entre « sorte de » et « espèce de ». Nous avons d'abord constaté que ces deux occurrences n'apparaissaient pas toujours dans le même contexte, elles n'ont pas pas les mêmes propriétés distributionnelles. Cette remarque nous a conduit à nous interroger sur le caractère grammatical de ces unités. L'hypothèse majeure résultant de cette analyse serait que « espèce de » est une occurrence ayant un caractère plus nominal que «sorte de », elle conserverait les propriétés du nom et donc sa sémantique lexicale originelle. « Sorte de » aurait moins de propriétés lexicales. D'un point de vue sémantique, il n'indiquerait pas forcement un groupe, un ensemble (contrairement à « espèce de »). Étymologiquement, « sorte » signifie « espèce de » mais cette occurrence a perdu de son sens. On indique de manière étymologique que l'individu est une sous catégorie mais il s'est desémantisé pour indiquer uniquement l'opération de rapprochement, d'approximation. Un nom qui se grammaticalise perd de son sens et de ses propriétés nominales. Une unité lexicale indique quelque chose qui se passe dans le monde (qui a une valeur référentielle) alors qu'une unité grammaticale sert à donner plus d'informations sur d'autres mots. En diachronie, la tendance à aller du sens lexical au sens grammatical pour un mots correspond à la javélisation sémantique. Ce type de processus est partout dans la langue.
Après toutes ces remarques, nous possédons donc l'instrument théorique qui nous permet d'expliquer que « sorte de » est plus grammatical que « espèce de ». On a besoin de corpus pour vérifier ces hypothèses, et pour vérifier que notre intuition est fondée. On veut également voir si on peut considérer que le lien entre « sorte » et « de » (coalescence) est plus fort que celui entre « espèce » et « de ». (Ce n'est pas exactement le cas encore car le processus de grammaticalisation est un processus lent et qui est en cours) mais on veut vérifier ça. )

**2ème étape : Analyse des données sur Sketch Engine**

**Première observation → fréquence d'apparition**

Afin de mesurer le degré de grammaticalité des occurrences, nous avons premièrement analysé la fréquence d'apparition des occurrences car celle-ci nous donne des indications quant à la grammaticalité d'une occurrence (plus une occurrence est fréquente, plus elle est grammaticale).

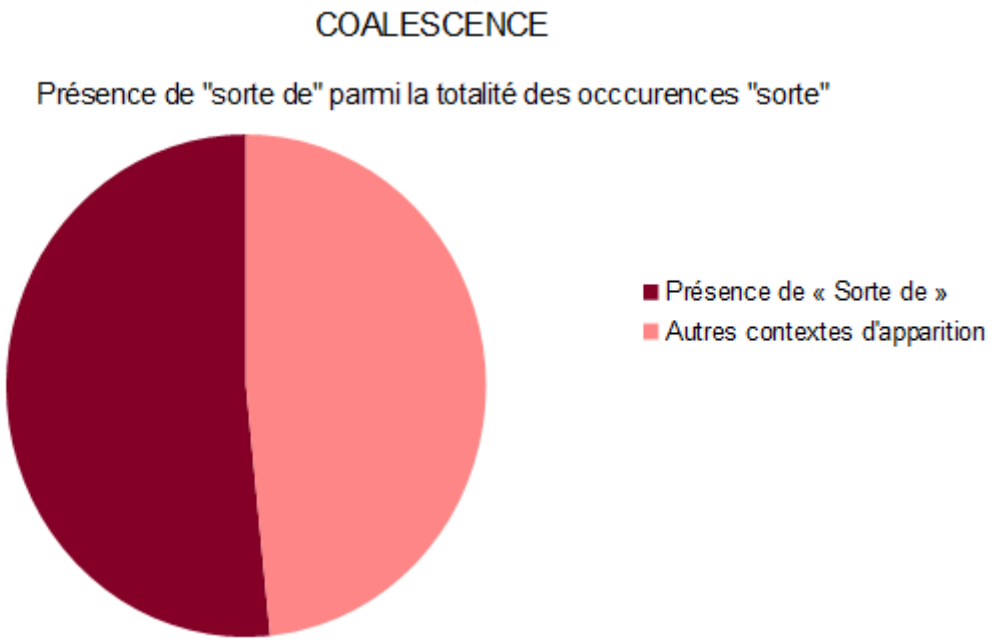
sorte → **1,978,247** dans tout le corpus (172.85 per million)
espèce → **840,529** (73.44 per million)

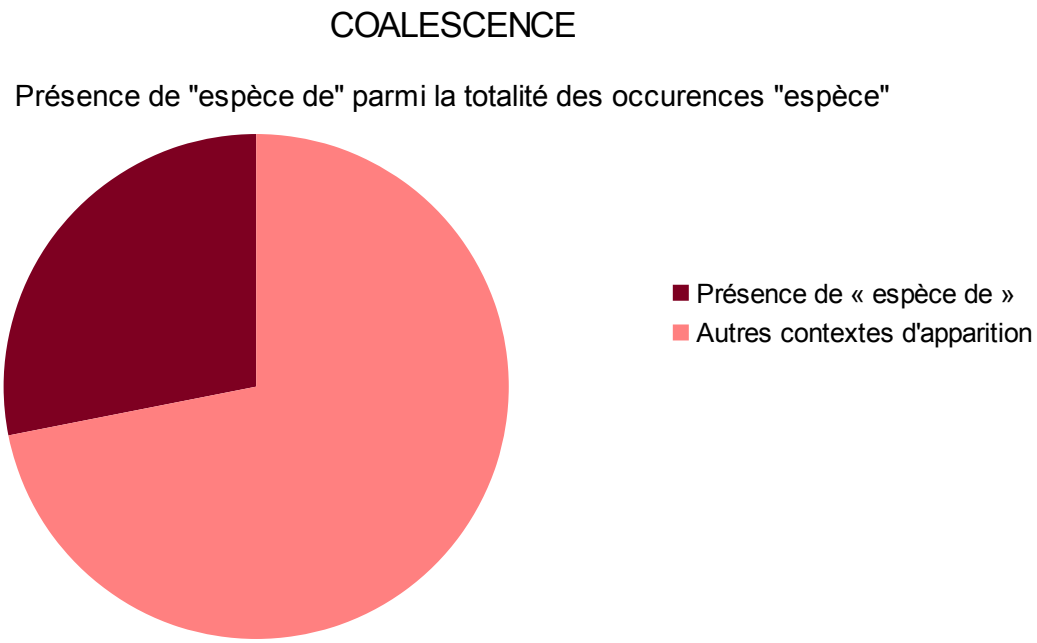
**<u>Deuxième observation → la coalescence</u>**

Pour étudier « sorte » et « de » en même temps, on a besoin des expressions régulières. Les expressions régulières sont des notations algébriques qui nous permettent de chercher dans un corpus des séquences de caractères qui ont les mêmes propriétés. On les utilise quand on veut chercher toutes les occurrences de la lettre /a/ par exemple ou encore les caractères compris entre a et v [a; v].
Sur Sketch Engine, on utilisera donc le Corpus Query Language. C'est un langage basé sur la notion d'expressions régulières. On tape donc dans CQL [lemma="sorte"][lemma="de"]. Attention aux erreurs, parfois des mots sont mal lémmatisés et étiquettés.

sorte de → **1,020,892** (89.20 per million)
espèce de → **236,165** (20.63 per million)

Les résultats confirment que la coalescence entre « sorte » et « de » est plus forte que la coalescence entre « espèce » et « de » . L'expression « sorte de » est également plus présente que l'expression « espèce de », ce qui est un indicateur quant à la grammaticalité de ces mots.






**<u>Troisième observation → étude de la nominalité des occurrences en voyant si ils sont modifier par des adjectifs, des déterminants, etc</u>**

**Quelles formes faut-il exclure de nos corpus ? Comment les exclure sur Sketch Engine ?**

3 exemples de constructions contenant l'occurrence « sorte de » :

• Toutes sortes (pluriel) - "hier j'ai mangé toutes sortes de gâteaux" → toutes les sortes qui existent (tous les éléments)
• Toute sorte (singulier) - "je refuse toute sorte de gâteau" → n'importe quel gâteau (n'importe quel élément)
• Une sorte - "hier j'ai mangé une sorte de gâteau" - un gâteau qui n'était pas prototypique, un élément qui se rapproche de l'ensemble, qui se situe à la limite de l'ensemble (un élément qui se rapproche)

Dans les trois cas, on est face à une opération qui consiste à dire quels types de relation ce référent a avec un ensemble d'éléments auxquels il appartient ou presque.

Il faut exclure "faire en sorte de" mais on peut garder les trois autres constructions (c'est un choix arbitraire). On demande à sketch engine d'exclure toutes les formes de "sorte de" précédées de "en" La formule régulière est [lemma!="en"][lemma="sorte"][lemma="de"]. On a une différence de plus de 34 000.

Comme ça ne change pas grand chose d'étudier des milliers et non des millions d'occurrences et pour limiter la puissance et le temps de travail informatique utilisé/ nécessaire, on limite ce nombre d'occurrence à 10 000. Pour cela, on va sélectionner « sample » dans la colonne de gauche sur Sketch Engine et choisir 10000.

Pour continuer à analyser la grammaticalité des occurrences il faut rendre en compte de certains

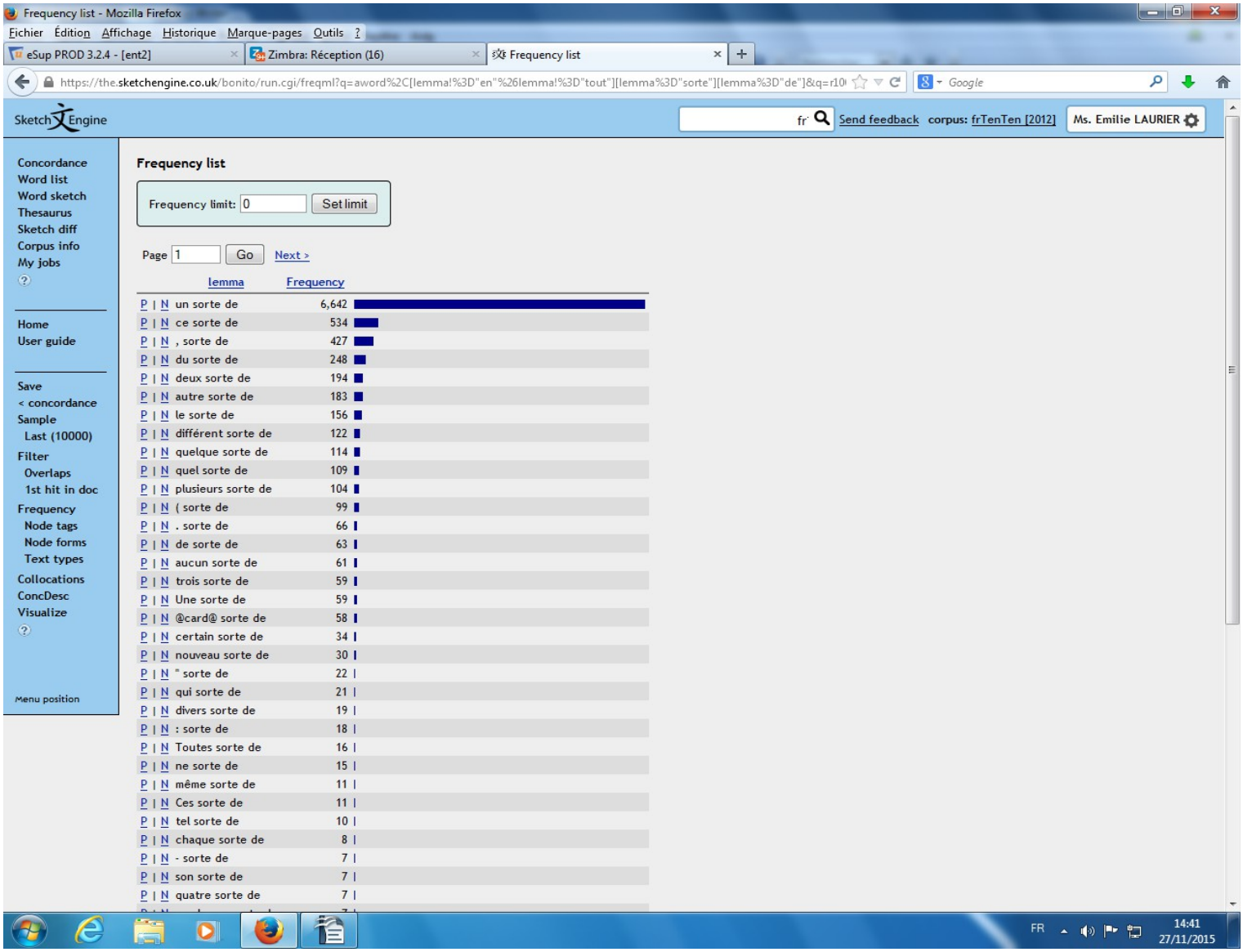
éléments :

- statut nominal de « sorte de » et « espèce de » dans les corpus. Ce statut est validé par le fait qu'il soit précédé d'un déterminant et/ ou suivi d'un adjectif.
- On doit également vérifier qu'on trouve « sorte de » dans un plus grand nombre de contexte. Pour cela on doit regarder/ analyser les noms qui suivent « sorte de » et « espèce de ». Hypothèse sous-jacente : « espèce de » est suivi d'un lexique plus spécifique.

Ensuite on veut élimer toutes les occurrences de « toute » et « toutes » dans notre analyse, on élimine donc le lemme « tout ». Cela nous permet donc de récupérer les occurrences de « sorte de » qui ne sont pas précédées par « en » et « toute » et « toutes ».

En ce qui concerne la formule régulière a utiliser, on ne peut pas écrire [lemma != « en »][lemma!= « toute »][lemma = « sorte de »] car Sketch Engine va croire que l'on met à la suite « en » et « toute » et « sorte de » (=« en toute sorte de ») donc ça ne fonctionne pas.
On doit écrire [lemma!= « en »&lemma!= « tout »][lemma = « sorte »][lemma= « de »]

Ensuite, on va étudier le lemme le plus fréquence à gauche de notre occurrence. On va sur « frenquency » et on demande une liste de fréquence.



Lors de la formulation des hypothèses d'intuition, nous pensions que « espèce de » pouvait se trouver en début de phrase et pas « sorte de ». Or, grâce à cette outil, on s'aperçoit que « sorte de » peut être précédé d'un point et donc apparaître en début de phrase. En regardant les occurrences, on voit qu'elles ont toutes la même forme :

**/SENT Sorte /NOM de /PRP** → « sorte de fable allégorique » → cela fait référence à une double prédication, c'est à dire une construction dans laquelle on prédique un élément sans utiliser une construction verbale. On antépose une prédication sans verbe = apposition nominale. Un élément nominal aposé à un autre nom et qui sert à prédiquer (dire quelque chose d'une autre chose).

Résultat : on peut donc trouver « sorte de » en début de phrase, notre hypothèse n'était donc pas justifiée.

Ensuite, on a chercher à voir si les déterminants précédents les occurrences s'accordent avec « sorte » ou « espèce » ou avec les noms qui les suivent.

Par exemple, on dit « un espèce de con ». Dans ce cas, « un » s'accorde avec « con » et non avec « espèce », le caractère nominal de « espèce de » et donc remis en cause, le déterminant s'accorde parfois avec le nom situé après plutôt qu'avec « espèce », cela est il pareil pour « sorte de » ?

« Un » et « une » sont deux tokens du même lemme. En observant leur fréquence, on va voir quel type d'élément compose notre nœud. On obtient 35 occurrences « un sorte de... », où le déterminant n'est pas accordé avec « sorte de... » Est-ce une erreur de français ? Comment faire pour le savoir ? Dans une analyse comme celle ci, le linguiste peut juste faire des observations.

Remarque : Jusqu'à maintenant, nous avons prouver que « sorte de » se grammaticalisait . Or l'analyse des tokens précédant cette occurrence prouve que « sorte de » possède encore un caractère nominal dans le sens où le déterminant qui le précède s'accorde avec « sorte de ».

Pour vérifier le caractère nominal de « sorte de », on va également analyser si il est modifié par un adjectif.
On veut savoir si « sorte » est modifié par des adjectifs. On utilise pour cela la dépendance, elle permet d'identifier toutes les occurrences d'un adjectif qui modifie « sorte de », à n'importe quelle distance.

Problème → Sketch Engine n'est pas annoté pour la dépendance.

Solution → On va chercher les adjectifs qui suivent « sorte de ». Pour savoir que la notation de adjectif est ADJ, on se sert du tagset summary qui recense les abréviations de toutes les catégories syntaxiques.

• Pour « sorte de »

L'expression régulière utilisée sera :

[lemma!="en"&lemma!="tout"][lemma="sorte"][tag="ADJ"][lemma="de"]

On trouve 3,845 (0.34 per million) occurrences. On utilise un critère positionnel mais cela ne veut pas dire que tous les adjectifs qui suivent « sorte de » ont une relation de dépendance avec cette

occurrence. On change « word » dans frequency par « lemma » parce que sinon on a toutes les formes d'un même lemme avec « word ».

Seulement il y a toujours des adjectifs qui ne dépendent pas de « sorte de », on doit donc éliminer les occurrences qui ne qualifient pas « sorte de ».
On revient à la recherche « sans adjectif » et on fait une recherche de frequency avec 1R. Cela veut dire qu'on s'intéresse à l'élément qui suit le nœud. Ici, le nœud ne correspond pas à « en », pas à « tout » mais à autre chose + « sorte de ».

Nous avons également trouvé que « sorte de » pouvait être précédé d'un chiffre, comme dans : « 4 sortes différentes de poils ». La notation algébrique correspondant aux numéros est @card@. Nous avons décidé de garder ces contextes car ils représentent toujours plus ou moins une opération sur un ensemble.

**Quatrième observation→ on étudie les adjectifs placés entre « sorte » et « de » et entre « espèce » et « de »**

Exemple : une espèce importante de.... Si il y a un adjectif entre « espèce » et « de » alors cette occurrence sera plus nominale.

• Pour « espèce de »:

[lemma="espèce"][]{0,1}[lemma="de"]

résultat → 261 135

[] = n'importe quel type d'élément, de token
{0,1} = soit un élément, soit rien

On veut observer quand il y a un élément entre « espèce » et « de » et on souhaite que cet élément soit un adjectif

Nouvelle formule = [lemma="espèce"][tag="ADJ"][lemma="de"]

Résultat = 9500 = 3,64 %

• Pour « sorte de » :

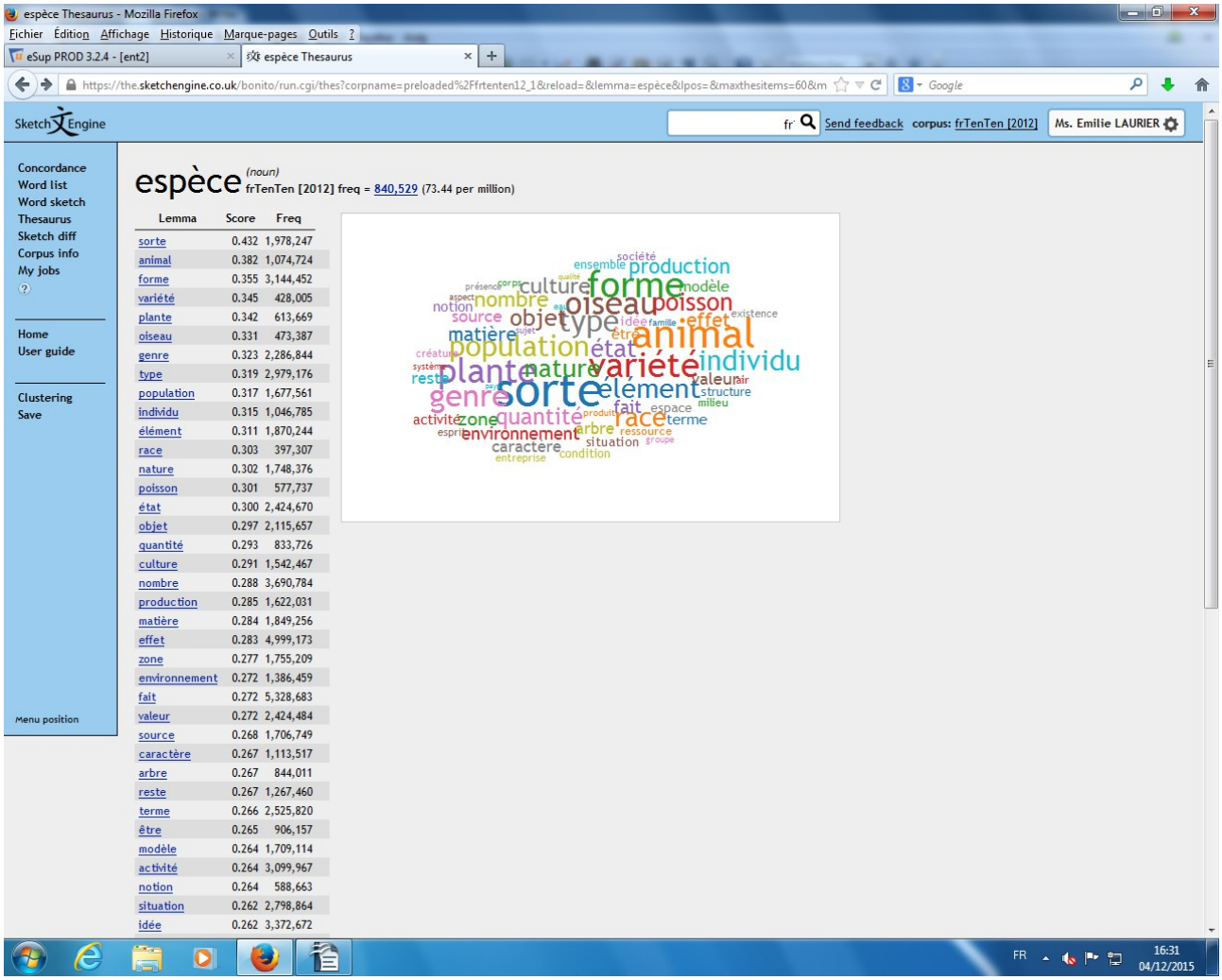
[lemma!="en"&lemma!="tout"][lemma="sorte"][]{0,1}[lemma="de"] → 815 079

[lemma!="en"&lemma!="tout"][lemma="sorte"][tag="ADJ"][lemma="de"] → 3845 = 0,47 %

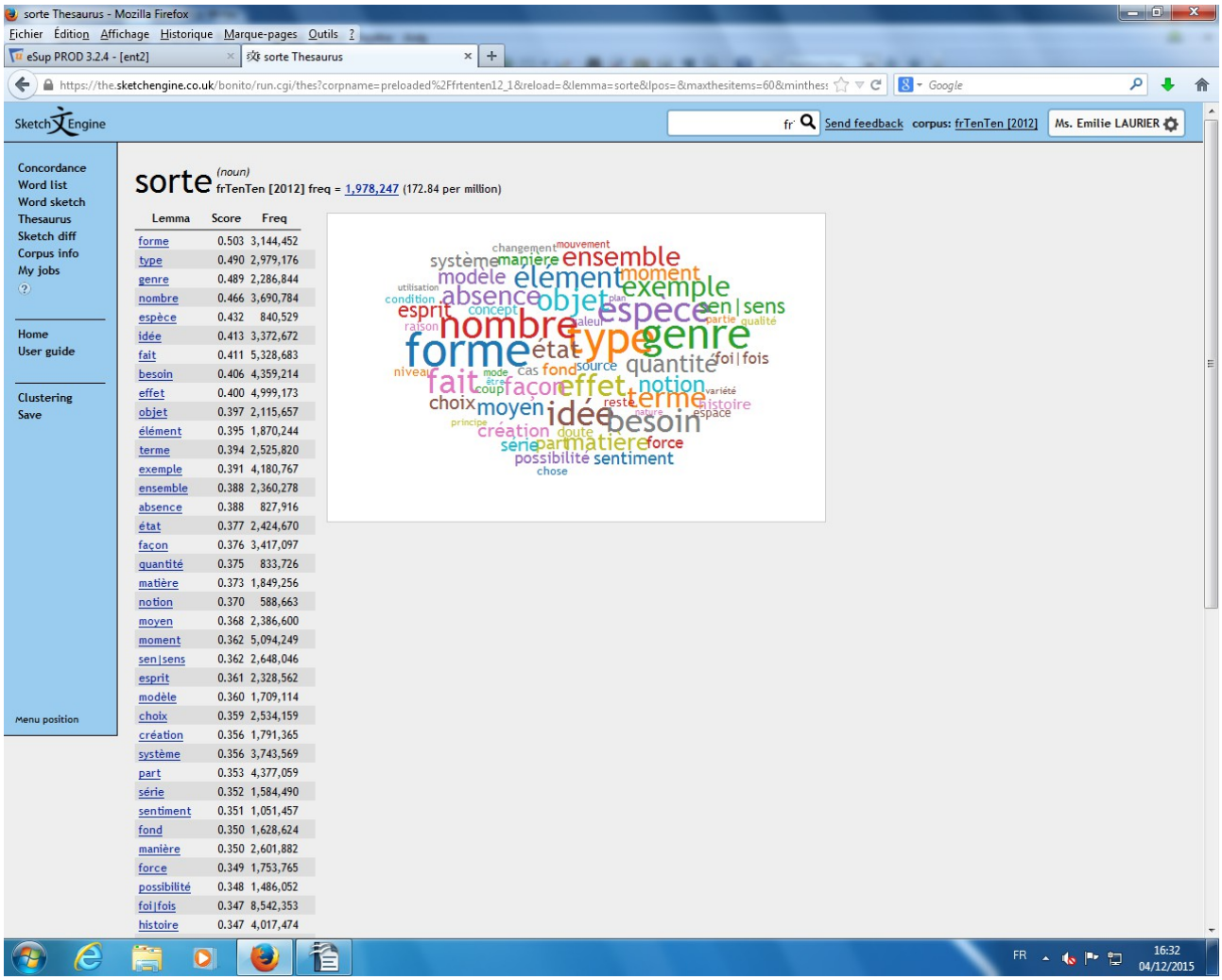
Interprétation des résultats → « espèce de » a un caractère plus nominale (type catégorielle) au niveau syntaxique et sémantique que « sorte de ».

Grâce à l'outil Thesaurus, on recherche les mots qui ont les même propriétés distributionnelles que « espèce de ». On utilise les sketchs lexicographiques pour voir leur similarité distributionnelle.

Résultats → on trouve quels mots sont similaires,dans quels contextes, et pourquoi cela a été calculé comme ça.




On va voir si « espèce » et « sorte » ont les mêmes mots similaires




Mots ayant les mêmes propriétés distributionnelles que « sorte » et « espèce » :

• Espèce → variété, animaux, oiseaux, biologie. Le mot « espèce » est encré dans son sens nominal, et en particulier sons sens d'indication de type biologique.

• Sorte → forme, type, genre, nombre, Moins spécifié.

Cette analyse nous permet de voir que la synonymie entre « espèce de » et « sorte de » n'est pas une synonymie parfaite. « Sorte » sera le premier lemme à avoir les mêmes propriétés distributionnelles que « espèce » alors que le premier lemme qui possède les mêmes propriétés distributionnelles que « sorte » est « forme ».

## Conclusion

Quand on étudie un corpus, on étudie pas seulement un état de langue, on étudie la dynamique d'un état de langue. De par nos observations, nous pouvons conclure que les deux unités ont tendances à évoluer vers un marqueur d'approximation sauf que ce processus est beaucoup plus rapide pour « sorte de » que « espèce de » . Ce type d'analyse nous permet de voir comment les processus diachroniques sont constamment à l'œuvre dans les langues. Ce phénomène échappe à notre intuition, puisque nous possédons l'intuition d'un individu et non celle d'une communauté.