Santi
Paloma

Outils linguistiques
Etude de cas : « sorte de » & « espèce de »

Cette étude s'inscrit dans le domaine de la linguistique de corpus, qui s'est fortement accru en France ces dix dernières années.
L'objectif de notre recherche est de comparer les deux constructions langagières suivantes : « espèce de » et « sorte de ». Celles-ci sont quasi synonymes, il est donc intéressant d'étudier ce en quoi ces formes différent, d'un point de vue sémantique et distributionnel.
Pour mener notre analyse, nous nous sommes appuyés sur la théorie du langage qui fut notamment développée dans les années 50 par Z.Harris : la sémantique distributionnelle. Elle se veut de montrer une corrélation évidente entre la sémantique d'un objet et sa distribution contextuelle. Si deux éléments ont la même sémantique, on s'attend à les retrouver dans les mêmes contextes linguistiques, or, la synonymie pure n'existe pas.
Si deux constructions telles que « sorte de » et « espèce de » se trouvent dans une relation de quasi synonymie et apparaissent dans des contextes différents, il va de soi que l'identification de ces contextes nous permettra d'en faire ressortir les nuances sémantiques.
Tout d'abord, nous avons émis des hypothèses sur les différences entre « espèce de » et « sorte », à partir de notre simple intuition. Pour vérifier la véridicité ou non de ces hypothèses, nous avons exploité un logiciel d'analyse de corpus nommé Sketch Engine. Nous allons démontrer la manière dont nous avons utilisé ce support pour enrichir nos connaissances linguistiques ainsi que les différentes étapes de notre travail.

Hypothèses intuitives sur les différences entre les deux constructions

- ♣ Des formes telles que « espèce de con ! » sont possibles, ce qui n'est pas le cas avec « sorte de » : « *sorte de con ! ». Dans ce contexte, « espèce de » est suivi par un terme péjoratif et a une valeur d'insulte. Ainsi, [espèce de] pourrait être utilisé dans un acte illocutoire d'exclamation sans que [espèce de] soit régi par un verbe. Dans cette configuration, « espèce de » serait toujours suivi d'un terme péjoratif.
- ♣ Selon une deuxième hypothèse, « espèce de » peut servir à identifier une catégorie biologique d'animaux ou de végétaux. Par exemple « une espèce de poisson » tandis que « une sorte de poisson » ne renvoie pas à ce sens de type, catégorie.
- ♣ « Sorte de » serait une construction moins nominale que « espèce de », c'est-à-dire que « espèce » conserverait toutes les propriétés d'un nom tandis que « sorte » serait plus grammaticalisé et perdrait de ces propriétés dans certains types de constructions.

Avant de passer à la démonstration de nos analyses, il est important de redéfinir quelques termes clés que nous abordons au sein de notre travail.
Un mot a un sens dit lexical s'il fait référence à une entité, action, propriété du monde.
Le sens grammatical, quant à lui, sert à spécifier des caractéristiques d'un autre mot qui fait référence au monde ou à spécifier les relations entre les mots lexicaux.

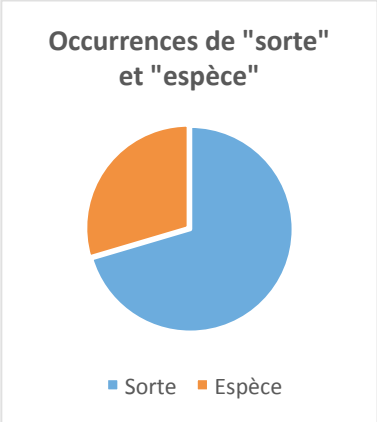
Analyse

Nous avons utilisé, tout au long de nos recherches, un corpus très large déjà existant, « Fr Ten Ten » sur Sketch Engine. Il comporte dix millions de mots. Pour notre recherche, seules quelques centaines d'occurrences suffirent pour valider nos hypothèses. Cependant on doit pouvoir avoir une immense variété de données pour être certains de retrouver les différents contextes d'emplois des constructions « sorte de » et « espèce de ».

Si notre hypothèse selon laquelle « sorte de » est une construction plus grammaticale que « espèce de » alors il sera question de vérifier les trois conditions suivantes :

- « sorte » est plus fréquent que « espèce »
- « sorte de » est une construction plus coalescente que « espèce de »
- le statut nominal de "sorte" doit être plus faible que celui de "espèce". A savoir que le statut nominal d'une unité se mesure en fonction de la modification par un adjectif et de l'accord avec un déterminant.

Tout d'abord, on a regardé la fréquence d'apparition de « sorte » et de « espèce ».
Les résultats démontrent que « sorte » apparaît plus fréquemment que « espèce » : deux millions d'occurrences pour « sorte » (dont 172 000 par million) contre 840 000 occurrences pour « espèce » (dont 73 000 par million).



« Sorte » apparaît plus fréquemment que « espèce ». Cette observation a constitué une première avancée vers la validation de notre hypothèse selon laquelle « sorte de » serait plus grammatical que « espèce de ».

Ensuite, on s'est intéressé aux constructions « sorte de » et « espèce de ».
Au sens littéral, « sorte de » et « espèce de » renvoient à la même chose. Une espèce correspond à une catégorie, un type. De même pour une sorte. Cela correspond à leur sens lexical plein.
Cependant, prenons un exemple où cela n'est pas le cas : « une espèce de poisson » fait en effet référence à un type, un groupe précis de poisson tandis que « une sorte de poisson » ne fait pas référence à un groupe mais à un animal qui se rapproche plus ou moins du poisson prototypique.

Ici, "sorte de" sert à faire des opérations sur le référent poisson, c'est un approximant, il serait donc grammaticalisé.
Rappelons que le processus de grammaticalisation est le fait pour une unité linguistique ayant un sens plein, d'obtenir, diachroniquement, un sens plus grammatical.

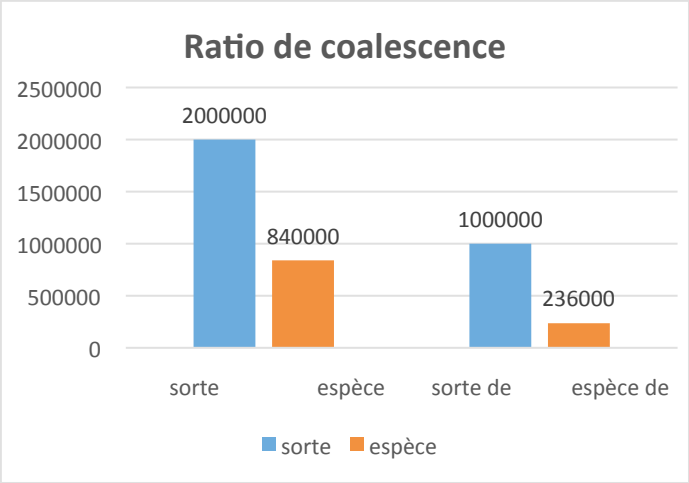
Nous avions illustré cela en classe par l'exemple suivant : « amare habeo », forme latine qui signifie « je dois aimer ». Or, la nécessité d'aimer implique que dans le futur cette nécessité se réalise. « Amare habeo » a à la fois la signification de "je dois aimer" et "je vais aimer" (je le ferai). On a la fois l'interprétation de nécessité et de futurité. Quand un élément perd sa signification lexicale pleine, il a tendance à subir une érosion phonologique, à être prononcé plus rapidement). En effet, l'expression « amare habeo » est devenue diachroniquement « amero » (« j'aimerai »). « Habeo » est devenu un morphème –o qui s'est accolé à l'élément « amare ».
Le **processus de grammaticalisation** correspondrait ainsi à une ***javellisation sémantique*** qui entraîne une ***érosion phonologique*** des unités. Celles-ci deviennent plus légères en substance phoniques et ont tendance à s'accoler, comme illustré dans l'exemple ci-dessus. Cela correspond au processus de coalescence : des éléments fusionnent. C'est un processus qui s'opère sur des centaines d'années.
Ainsi, « sorte de », qui selon nos hypothèses serait en javellisation sémantique, devrait par la même être une forme de plus en plus coalescente et apparaître plus souvent que « espèce de ».
Pour cela, nous sommes allés sur Sketch Engine et avons utilisé la notation d'expressions régulières: pattern matching. C'est une notation algébrique qui nous permet de définir de manière formelle des patterns, des schémas de séquences de caractères.
On sélectionne la fonction CQL qui va nous permettre de demander au corpus de sortir toutes les séquences de lemmes. Chaque élément correspond à un une étiquette, un lemme.
La formule permettant de vérifier cela: [lemma = "sorte"] [lemma = "de"]. Nous avons réitéré l'opération pour « espèce de ».
Les résultats sont les suivants : **1 million d'occurrences** (89 000 par million) pour « sorte de » contre **236 000** (20 000 par million) pour « espèce de ».

Il est important de préciser qu'à la suite de cette manipulation, nous avons retrouvé des occurrences de « sorte » avec le statut de verbe. Nous avons alors regardé dans « view option » pour savoir comment le corpus avait été étiqueté.
On a remarqué que pour certaines occurrences, Sketch Engine comprenait « sorte » comme un nom alors qu'il s'agissait d'un verbe, par exemple « qu'il sorte ». Ceci illustre certaines limites des corpus électroniques qui, comme nous, ne sont pas infaillibles. Cependant, ces erreurs sont moindres, pas significatives sur un aussi grand corpus, et ne viennent ainsi pas fausser nos résultats.

Nous avons ensuite observé la **coalescence** de « sorte » et « de » en comparaison avec celle entre « espèce » et de « de ». Pour ce faire, il suffit de comparer le nombre d'occurrences de « sorte » avec celui de « sorte de », de même pour « espèce » et « espèce de » :

- Nous avons trouvé deux millions d'occurrences pour « sorte » et un million pour « sorte de ». Le ratio est ainsi de ½. Autrement dit, une fois sur deux « sorte » et « de » apparaissent ensemble.

- Pour ce qui est de « espèce » (840 000 occurrences) et «espèce de » (236 000 occurrences), le ratio est de ¼. Une fois sur quatre, « espèce » apparaît avec « de ».



Ces indications chiffrées vont bien dans le sens de départ selon lequel « sorte de » est une forme plus grammaticalisée que « espèce de » car c'est une forme plus fréquente et plus coalescente. Il y a une tendance majeure pour « sorte » et « de » d'apparaître ensemble. Ces résultats valident la deuxième condition de notre hypothèse.

Nous avons ensuite sélectionné, de manière arbitraire, et comparé trois constructions dans lesquelles « sorte de » est précédée par un/plusieurs autre(s) élément(s) pour savoir si ces locutions ont le même sens : « toute sorte de », « toutes sortes de » , « une sorte de » « faire en sorte de » suivies par « gâteau(x) ».
La première, « **toute sorte** de gâteau », fait référence à n'importe quelle sorte de gâteau. Cette construction indique la présence de n'importe quel élément x, y ou n dans l'ensemble gâteau.
« **Toutes sortes** de gâteaux » fait référence à toutes les différentes sortes de gâteaux et indique la présence de tous les éléments x+y+n dans l'ensemble gâteau.
« **Une sorte** de gâteau » indique un élément pouvant se trouver dans l'ensemble « gâteau », ou non, mais à la limite de cet ensemble. Ici, « sorte de » indique l'approximation à un groupe prototypique. C'est une construction plus grammaticale que les deux précédentes. Cette occurrence de « sorte » est moins nominale car elle ne peut pas être remplacée par « type » par exemple et a un rôle d'approximant. Ici, « sorte » est [-NOM] tandis que dans les constructions « toute(s) sorte(s) de _ » il est [+NOM].
Ces 3 types de constructions sont toutefois en lien car elles constituent toutes des opérateurs sur un ensemble. Soit elles permettent de choisir un élément de l'ensemble, tous les éléments de l'ensemble ou permettent d'identifier un élément qui est proche de l'ensemble. Elles sont apparentées d'un point de vue sémantique.
«* Faire **en** sorte de gâteau ». La construction « faire en sorte de » n'a rien à voir avec la notion d'ensemble. On s'éloigne complètement de l'idée d'opération sur un ensemble. Il faut ainsi l'exclure car elle sort de l'objet de notre étude.
Ainsi, on a regardé le nombre d'occurrences de "sorte de" en excluant le « en » pouvant la précédé. Voici la formule adéquate tapée sur Sketch Engine:

[lemma!="en"][lemma="sorte"][lemma="de"]
On remarque qu'il n'y a que des dizaines de milliers d'occurrences de « sorte de » avec "en", plus précisément 34000, soit environ 3% du total d'occurrences de « sorte de ».

Nous avons vérifié et validé les deux premiers critères quant à la grammaticalité de « sorte de » en comparaison avec « espèce de». Notre intérêt se porte maintenant sur la question de la nominalité de « sorte de » par rapport à celle de « espèce de », qui dépend de sa modification par un adjectif et son accord ou non avec un déterminant.
Le degré de grammaticalisaté est inversement proportionnel au statut de nominalité des constructions "sorte de" et "espèce de". D'après les résultats obtenus jusqu'à maintenant, « sorte de » devrait avoir ainsi un faible statut nominal.
Il nous faut attester cela à travers des données chiffrées.

« Espèce de » - accord avec un déterminant/

Pour vérifier ce par quoi cette construction peut être précédée, nous avons utilisé un échantillon de 10 000 mots (largement suffisant pour avoir des résultats signficatifs), et regardé la fréquence d'apparition des déterminants. Ce ne sont pas les seuls éléments qui précèdent « espèce ». En effet, on trouve différentes occurrences telles que: « une espèce de », « l'espèce de », « toute espèce de », «**.** Espèce de », « cinq espèces de ».
D'après Sketch Engine, 64 fois sur 100000« espèce de » est précédé par un point. Cela nous indique que la forme « espèce de » peut apparaître en début de phrase sans nécessairement être précédée par un déterminant.
Nous nous sommes intéressés, dans ce cas précis où « espèce de » est tête de phrase, si s'ensuit toujours un terme péjoratif, comme nous l'avions proposé dans nos hypothèses. Ceci peut être vérifié en sélectionnant « lemma » dans « attribute », au sein de la fréquence. Les résultats montrent que « espèce de » n'est pas toujours suivi par une insulte car beaucoup de noms biologiques furent relevés, cela vient ainsi à l'encontre de notre intuition.
Ensuite, nous avons étudié le genre des déterminants précédant « espèce de » pour voir si l'accord était présent. Pour cela, il faut chercher le « word » qui précède « espèce » et non pas le « lemme » et nous avons obtenu: une > des > l' > un.
3375 lemmes de « un » dont 433 occurrences à la forme masculine « un » furent relevées, par exemple «un espèce d'ordinateur ». Il est intéressant de faire remarquer cela puisque « espèce », étant un nom de genre féminin, le déterminant le précédant est censé l'être également. Ces résultats nous permettent ainsi de dire que « espèce » présente un statut nominal affaibli en ce qui concerne l'accord avec le déterminant.
Contrairement aux attentes, "espèce" est précédé par des déterminants qui ne s'accordent pas et cela constitue un contre-exemple de la plus grande nominalité de « espèce de ».
Cependant cela pourrait s'expliquer par le fait qu'il y aurait une sorte de neutralisation du genre dans la plupart des contextes avec « espèce » à l'oral. Nous avons décidé de ne pas développer le point suivant car il s'écarte de notre objet d'étude.

Espèce de – modification adjectivale/

Pour répondre à la question de l'accord avec l'adjectif, il nous est nécessaire d'avoir accès un

« Treebank » annoté en dépendances, dans lequel pour chaque mot est précisé quel(s) élément(s) dépend(ent) de celui-ci. Or, Sketch Engine ne permet pas d'avoir accès à toutes ces dépendances. Cela constitue une nouvelle limite du logiciel, mais celle-ci est encore minime. On peut s'approximer de l'information en cherchant toutes les occurrences de « espèce » suivies par un adjectif, lui-même suivi par « de ».

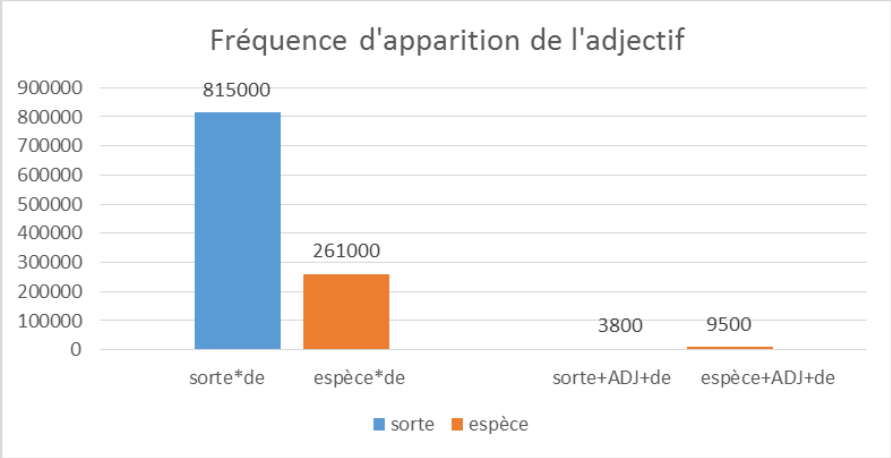
Rappelons que plus une forme sera modifiée par un adjectif plus celle-ci sera nominale, comme nous l'avions proposé précédemment.
Tout d'abord, nous avons regardé combien de constructions "espèce-adjectif-de" il y avait par rapport à la construction "espèce * de" («espèce » suivi de rien ou suivi par un élément autre qu'un adjectif puis « de »).Voici la formule correspondante :
[lemma= "espèce "][]{0,1}[lemma="de"]
Résultat : dans 261 000 occurrences, « espèce » est suivi par un élément autre qu'un adjectif.
Ensuite, on a regardé toutes les occurrences où un adjectif était présent entre « espèce » et « de » sur le total de 261 000 : [lemma="espèce"][tag="ADJ"][lemma="de"].
Résultat : 9500 soit 3,6% du total énoncé.

« Sorte de » - accord avec un déterminant/

Pour « sorte de », nous avons eu affaire à une difficulté supplémentaire. Nous avions retenu les deux constructions « toute sorte de », « toutes sortes de », où «sorte » est [+NOM].
Or, si l'on veut comparer « sorte de » et « espèce de », on doit éliminer « toute(s) », de la même manière que nous avions éliminé « en sorte de ».
Est-il possible d'enlever toutes ces formes d'un coup ? Oui, à l'aide de la formule suivante :
[lemma != « en»][lemma != « tout »][lemma="sorte"][lemma="de"]
Nous avons ainsi obtenu 789000 occurrences dont 6000 lemmes avec « un » suivi de « sorte de ».

« Sorte de » - modification adjectivale/

Nous avons opérer de la même manière qu'avec « espèce de », en pensant à enlever « en » et « toute(s) » :
[lemma !="en" & lemma !="tout"][lemma="sorte"][]{0,1}[lemma="de"]
Le résultat est de 815 000 occurrences pour des constructions avec des éléments autres que des adjectifs.
[lemma !="en" & lemma !="tout"][lemma="sorte"][tag="ADJ"][lemma="de"]
Il est de 3800 pour les cas où « sorte » est modifié par un adjectif soit 0,4% du total.



Bilan : Il y a ainsi une coalescence plus importante entre « un » et « sorte de » que pour « un » suivi de « espèce de » (6000 occurrences contre 3375). Cet élément vient de nouveau renforcer notre hypothèse quant à la plus grande grammaticalité de « sorte de ».
De plus, l'adjectif vient modifier beaucoup plus souvent "espèce de" que « sorte de », et cela de manière significative (3,6% contre 0,4%). Cela rejoint directement notre hypothèse selon laquelle "espèce de" est plus nominal que "sorte de".

Une fois l'analyse distributionnelle terminée, nous avons effectué une analyse sémantique plus poussée en se servant de l'outil **« Thesaurus »** de Sketch Engine. Cette application nous permet d'observer toutes les unités linguistiques qui sont proches de « espèce » et « sorte » d'un point de vue distributionnel, et qui font d'elles des synonymes potentiels.
Les mots qui apparaissent le plus souvent dans le même contexte que « espèce » sont : « sorte », « animal », « forme ».
En revanche, les mots ayant la même distribution que sorte de manière la plus fréquente sont : « forme », « type », « genre », « espèce ». On remarque que dans ce cas-là « espèce » n'apparaît que 4ème sur la liste.
Le fait que « espèce » et « sorte » ne partagent pas les mêmes synonymes implique et valide le fait que « sorte » et « espèce » ne sont pas de parfaits synonymes.
On observe finalement que « espèce » est plus nominal et a une prédilection à apparaître dans des registres de langue dans lesquels on parle de biologie.

Pour conclure, nous pouvons tout d'abord souligner le fait que notre intuition nous a permis de formuler des hypothèses valides quant aux différences entre « sorte de » et « espèce ». Cependant, la seule intuition ne suffit pas pour étudier de manière assez fine les différences sémantiques et distributionnelles de ces deux expressions. C'est pourquoi nous avons eu recours à de larges bases de données linguistiques permises par l'existence de corpus électronique tels que Sketch Engine.
C'est dans le cadre du cours d' « Outils linguistiques » que nous avons appris à manipuler ce type de logiciel, et utiliser la méthodologie adaptée pour répondre à nos questions de départ.

Ainsi, nous avons étudié trois choses principales : la fréquence, l'accord avec un déterminant et la modification par un adjectif, qui sont des indices majeurs sur le niveau de grammaticalité d'une unité.
L'ensemble de nos résultats nous ont permis de valider toutes nos hypothèses initiales et d'arriver à la conclusion que « sorte de » est une construction plus grammaticale que « espèce de ». « Espèce de » peut également présentée des aspects grammaticaux mais garde toutefois son sens plein dans la presque totalité des cas observés. Quant à « sorte de », c'est une construction qui sert d'approximant dans la grande majorité des cas et cela peut s'observer d'un point de vue diachronique. Ces deux formes n'ont ainsi pas le même mouvement vers la grammaticalité.
Ce travail a été réalisé dans le but d'une prise de conscience de l'existence d'outils informatiques fiables nous permettant d'augmenter nos connaissances linguistiques, sur la base de corpus immenses. La linguistique de corpus permet justement d'étudier la langue telle qu'elle est représentée chez la presque totalité des locuteurs de la communauté linguistique, et non pas la seule compétence d'un locuteur unique.