Outils linguistiques

Marianne BUISSON
L3 Sciences du langage

11/12/2015

Contexte

Afin de nous approprier les instruments de constitution, de traitement et d'analyse linguistique d'un corpus, nous avons mené une étude nous révélant les tenants et les aboutissants des connaissances théoriques préalablement abordées sur la linguistique de corpus et ses outils.

Le travail a porté sur deux constructions quasi synonymes (sous-entendu qu'il n'existe pas de synonymes parfaits) : *sorte de* et *espèce de*. Par conséquent, nous avons cherché à savoir ce que ces deux formes avaient en commun et ce qui les différencie, afin de leur attribuer un certain degré de synonymie. Pour ce faire, nous avons procéder en premier lieu à une analyse purement instinctive, dont nous verrons vite les limites, pour exploiter, en second lieu, un corpus. Cette analyse de corpus nous a montré que l'on peut extraire des indications précises sur les fonctions sémantiques de certaines unités à travers leur distribution.

« Sorte de » et « espèce de »

La première approche de ces deux formes s'est faite sans support, soit de manière **intuitive** et sur la base de nos connaissances syntaxiques et sémantiques. De cette façon, nous avons pu déterminer que *sorte de* et *espèce de* sont deux moyens de spécifier un référent à une catégorie à laquelle il appartient, comme dans les exemples « espèce de poisson » ou « sorte de poisson ». La catégorie est définie par l'élément associé à cette construction (ici, « poisson »).

Néanmoins, nous avons observé que leur apparition dépend du contexte et du cotexte. Cette constatation s'est établi sur l'exemple « espèce de con » où « espèce » apparaît ici dans un contexte dans lequel se réalise l'exclamation qui sert à insulter. Par contre, on ne trouve pas la forme **sorte de con*. De là, notre travail s'est alors appuyé sur **l'hypothèse distributionnelle de Harris**, selon laquelle la distribution d'une unité linguistique va nous en dire sur sa sémantique.

Les quelques distributions analysées nous ont menés à émettre l'hypothèse que *espèce* aurait plus tendance à se comporter comme un nom que *sorte*, bien qu'ils soient tous les deux issus de cette catégorie syntaxique des noms. *Espèce* contiendrait plus de caractéristiques descriptifs d'une catégorie lexicale (c'est-à-dire qui fait référence à une entité, une action, une propriété du monde) alors que *sorte* jouerait un rôle plus grammatical (comme spécificateur des relations des autres mots entre eux) dans son usage. Il se comporterait comme un nom qui sert à modifier un autre nom : un nom support. Pour en arriver à cette supposition, nous avons regardé les caractéristiques de chacun par rapport au faisceau descriptif d'un nom (accord en genre et en nombre, déterminants, etc.) de la manière suivante :

Nom ⇓ Nom support	*Exemples*	*Possibilité de pluriel, de déterminants, etc.*
	[espèce][de poisson]	OUI
	[espèce de][bouteille]	NON
	[sorte de][bouteille]	NON

On pourrait dès lors dire que *espèce* fait plus référence au monde et donc à une catégorie, et que *sorte* sert à faire des opérations sur le référent comme marqueur d'approximation (c'est plus ou moins ce référent). Une *espèce de poisson* appartient à la catégorie des poissons, mais une *sorte de poisson* ne rentre pas dans la catégorie poisson, il s'en rapproche seulement (l'animal ressemble à un poisson).

Grammaticalisation, javellisation sémantique et coalescence

La grammaticalisation est un processus diachronique par lequel une unité à l'origine lexicale va perdre de son sens pour devenir une unité à fonction plutôt grammaticale. En latin, *amare habeo* était utilisé pour exprimer la nécessité d'aimer (*je dois aimer),* mais implique que dans le futur cette nécessité se réalise *(je vais aimer)*. Finalement, c'est donc plus le futur que la nécessité qui est exprimé comme lorsque l'on dit « je dois aller travailler ».

La grammaticalisation implique une **javellisation sémantique**, par laquelle un élément lexical est blanchi ; en d'autres termes, il perd sa coloration sémantique. Cette javellisation sémantique semble aller de pair avec une **érosion phonologique** puisque quand un élément perd ainsi sa signification lexicale pleine, il a tendance à être prononcé de façon plus rapide, à devenir plus léger au niveau du corps phonique pour aller jusqu'à se coller à son contexte *(amare habeo -> amaro -> amero -> amerai).* L'élément prend une valeur plus morphologique et va se coller à la structure qui l'héberge. Cette fusion porte le nom de **coalescence** et peut s'observer à travers des corpus.

Dans notre étude, on considère donc que *sorte* est en grammaticalisation et que *de* viendrait se coller à *sorte.* Une démarche intuitive s'appuie sur un raisonnement direct sans recours à l'expérience, c'est-à-dire non vérifié, c'est pourquoi elle ne nous permet que de façonner des hypothèses et non d'atteindre des conclusions fondées. Or, nous cherchons à valider nos hypothèses et pour inspecter et vérifier notre intuition, nous devons avoir recours aux corpus.

Notre corpus

Pour examiner nos suppositions, nous nous sommes basés sur un corpus du français déjà existant d'environ dix milliards de mots, le FrenchTenTen, auquel nous accédons par le biais de Sketch Engine. Le travail a donc été effectué selon la tokenisation (découpage des unités linguistiques), la lemmatisation (forme base sur laquelle on reconduit toutes les formes morphologiques d'un même mot) et l'étiquetage (annotation de token et des lemmes) proposés par Sketch Engine.

Fréquence d'apparition

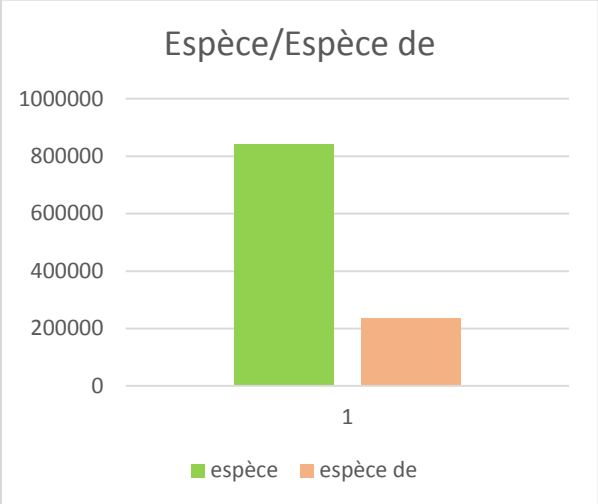
Le premier aspect inspecté dans notre travail a été la fréquence d'apparition. Si on fait l'hypothèse que *sorte de* est plus grammaticalisé et son sens plus javellisé que dans le cas de *espèce de*, on s'attend à le trouver plus fréquemment dans notre corpus.

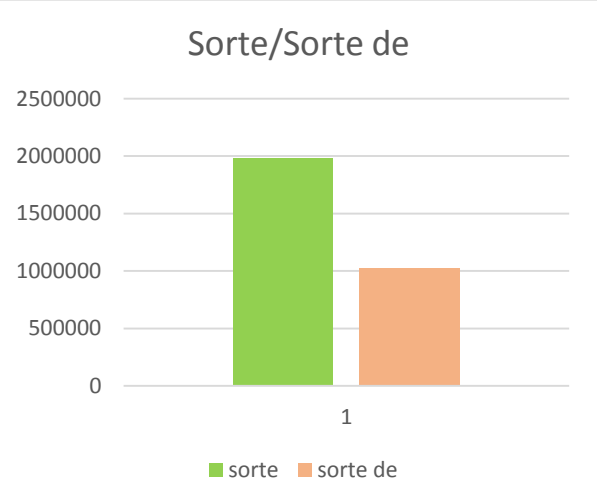
Les chiffres trouvés sont de 1,978,247 (17,285 per million) occurrences du lemme *sorte* et 840,529 (7,344 per million) pour *espèce*. Toutefois, les occurrences du verbe sortir apparaissent aussi dans les occurrences de *sorte,* et ne rentre pas dans le cadre de notre objet d'étude.

En conséquence, afin d'écarter les formes verbales, nous avons utilisé des expressions régulières sous forme de notation algébrique qui permettent de définir de manière formelles des patterns des séquences de caractères. De plus, elles nous ont permis de chercher plusieurs choses à la fois. Notre nouvelle requête s'est alors faite dans « CQL » ou « corpus query langage » (concordance > CQL) sous l'expression [lemma="sorte"][lemma="de"] qui permet au corpus de sortir toutes les occurrences de *sorte* suivies de *de.* Toutefois, après cela, quelques verbes apparaissent toujours et c'est là une des limites d'un travail sur corpus. Leur présence à un nombre peu élevé n'aura cependant pas de réelle incidence sur notre recherche. A ce moment-là, nous sommes passés par « View option » pour activer les tags, ce qui nous a permis d'identifier des erreurs –rares- d'étiquetage, erreurs qui nous ont fait réaliser là encore certaines limites du corpus quoique celles-ci puissent être corrigées.

Avec CQL, nous avons obtenu 1,020,892 (8,920 per million) occurrences de la construction *sorte de* contre 236,165 (2,063 par million) pour *espèce de. Sorte de* est effectivement largement plus fréquent que *espèce de* ce qui motive notre hypothèse que *sorte de* aurait les propriétés d'un élément grammaticalisé. En outre, si les deux formes avaient été synonymes, on aurait eu approximativement le même nombre d'occurrences.

En ce qui concerne la coalescence, les ratios entre les formes *sorte* et *sorte de*, et *espèce* et *espèce de* ont été calculés.



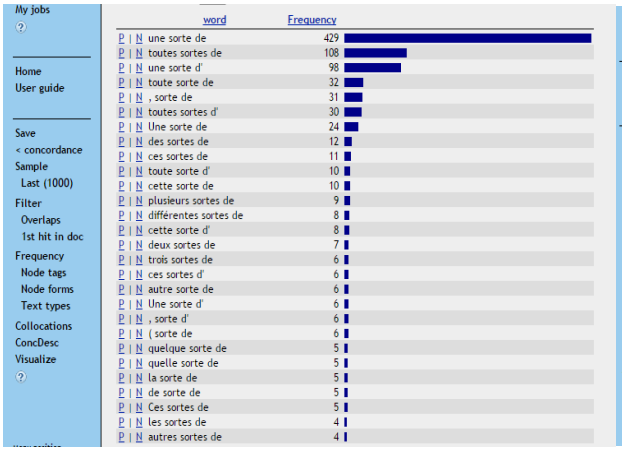


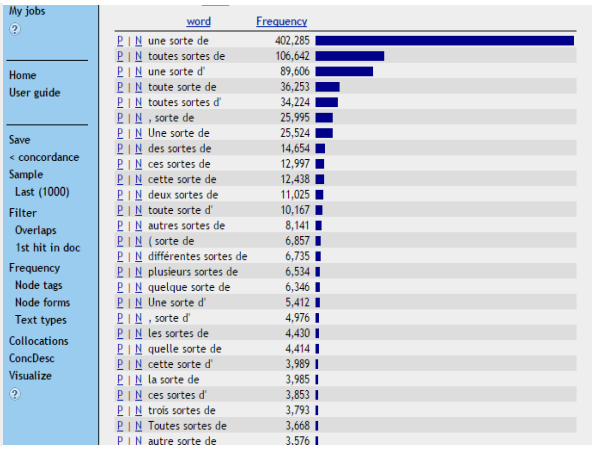
Ainsi, pratiquement une fois sur deux, *sorte* est collé à *de.* Sur le graphique, on voit en effet que *sorte de* représente à peu près la moitié des occurrences de *sorte*. A la différence, on ne trouve *espèce de* qu'une fois sur quatre occurrences de *espèce*. Ces informations nous font

encore dire que la construction *sorte de* est plus grammaticalisée que *espèce de,* car non seulement elle est plus fréquente, mais elle est aussi plus coalescente.

Echantillonnage

Dans « sample » sur Sketch Engine, nous avons réduit le corpus à un échantillon de 10 000 mots. Pour savoir si cet échantillon reste représentatif de l'ensemble, nous avons observé la liste de fréquence (« frequency »). Effectivement, les résultats sont proportionnels, autrement dit, l'échantillon est représentatif du corpus total. En effet, les courbes (imaginaires) de fréquence sont extrêmement similaires entre les deux. Notre travail s'est, de là, effectué sur un échantillon pour que le traitement soit plus rapide.





Echantillon Corpus complet

Pour assembler les formes telles que *sorte de* et *sorte d'*, dissociées sur les images ci-dessus, il faut demander « lemma » dans « frequency » plutôt que « word ».

Déterminants

L'étude de la nominalité de nos unités linguistiques est passée par celle des déterminants. Nous avons observé :

- *Toutes sortes :* « hier, j'ai mangé toutes sortes de gâteaux ». On comprend « toutes les sortes qui existent, tous les éléments ».
- *Toute sorte :* « je refuse toute sorte de gâteau ». On comprend n'importe quel gâteau.
- *Une sorte :* « j'ai mangé un sorte de gâteau ». On comprend un gâteau qui n'était pas prototypique, un élément qui se rapproche de l'ensemble, qui se situe à la limite de l'ensemble, qui s'en rapproche.

Dans ces trois cas, nous sommes face au même type d'opération, consistant à dire quels types de relation ce référent a avec un ensemble auquel il appartient, ou presque. Ils peuvent donc rentrer dans le cadre de notre étude. Nous avons donc décidé à ce moment-là de conserver ces constructions mais ce choix est arbitraire. Par contre, il a fallu supprimer *faire en sorte de.* Sur sketch engine, nous demandons l'exclusion de toutes les formes de

sorte de précédées de *en* sous une forme algébrique avec « ! » : [lemma != "en"] [lemma="sorte"] [lemma="de"], ce qui nous mène à 34 000 occurrences de moins que pour *sorte de.*

Par la suite, nous avons préféré éliminer les *toute* et *toutes* déterminant *sorte* pour savoir ce que l'on retrouve le plus fréquemment à gauche de *sorte* et sous quelle forme. La manipulation d'exclusion de ce lemme est la même que pour *en* mais avec l'utilisation du « & » de la manière suivante : [lemma != "en" & lemma != "tout"] [lemma="sorte"] [lemma"de"]. Puis, nous sommes retournons dans « frequency » en sélectionnant lemma pour ramener les formes d'un même mot à sa forme de base. La forme la plus fréquente, avec 6642 occurrences, est *un/e sorte de.*

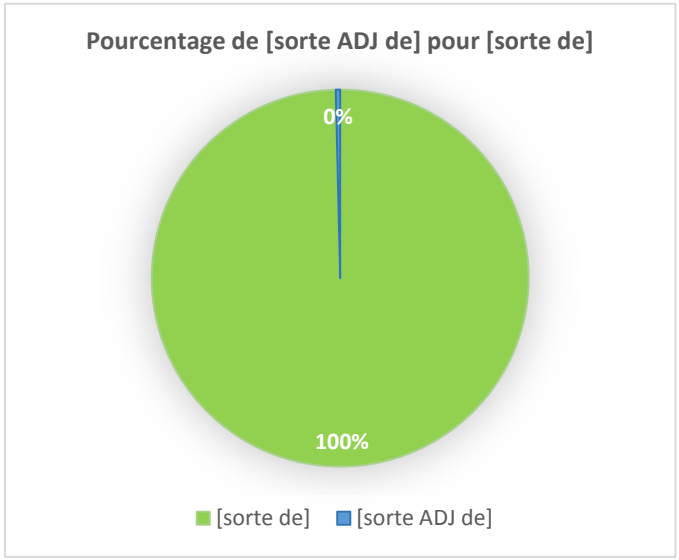
Avec *espèce de,* on sait que le déterminant est parfois accordé avec le nom après *de* plutôt qu'avec *espèce* : *un espèce de con.* Est-ce aussi le cas avec *sorte de* ? Pour répondre à cette question, nous sommes retournés dans « frequency » en demandant « word » pour récupérer les tokens. Les résultats est de 35 occurrences de *un sorte de.* L'existence de ces occurrences forme une contradiction à notre intuition puisqu'elles ramènent un peu *sorte de* vers un caractère nominal. Néanmoins, rien n'est jamais totalement défini dans l'utilisation de la langue, et un processus comme celui de la grammaticalisation se fait très lentement. De plus, le travail du linguiste est d'observer, il ne doit pas se poser de questions sur les erreurs possibles.

Double prédication

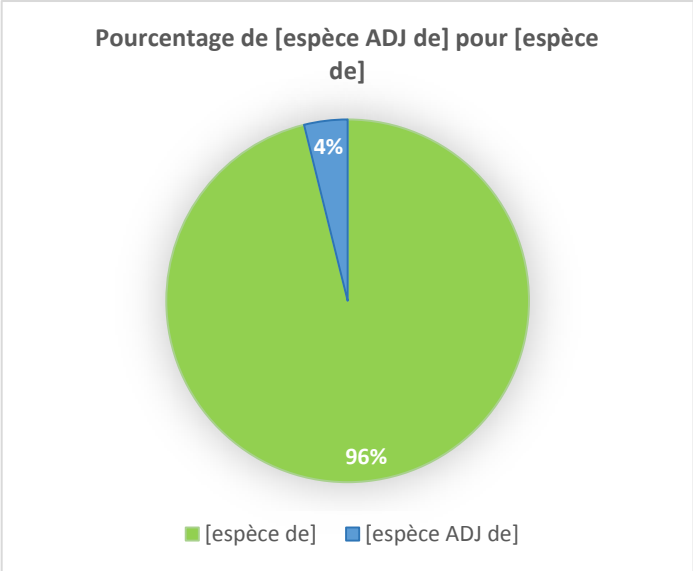
Notre intuition nous a menés à dire que *espèce de* pouvait se trouver en initial de phrase et que ce ne serait pas le cas pour *sorte de.* Pourtant, nous avons pu observer dans « frequency » que *sorte de* apparaît après des points, soit en début de phrase. Exemple tiré du corpus : ./SENT Sorte /NOM de /PRP mentor, Vinicius a produit son dernier (…). D'un point de vue syntaxique, cette structure s'appelle une double prédication ; c'est une forme précise dans laquelle une construction nominale sert à introduire ensuite une prédication, elle sert d'introduction à un verbe. Autrement dit, on a un nom à propos d'un autre nom.

Adjectifs

Toujours pour déterminer le caractère nominal d'un mot, il est aussi judicieux d'observer s'il est modifié par un adjectif. Ce dernier peut être épithète, mais il peut aussi qualifier un nom à distance. Pour raccrocher les noms à leurs modificateurs, nous devrions utiliser la dépendance qui nous permettrait d'identifier toutes les occurrences d'un adjectif se rapportant, dans notre cas, à *sorte de*, et ce à n'importe quelle distance. Cependant, l'outil que nous utilisons, Sketch Engine, n'est pas annoté pour la dépendance et nous limite alors à un critère positionnel, c'est-à-dire uniquement aux adjectifs qui suivent directement notre nom.



Il nous a fallu insérer un nouvel élément, un tag, à notre formule pour cette nouvelle recherche.



A l'aide de « tagset summary », nous avons pris connaissance du code à utiliser pour les adjectifs : ADJ. L'expression [lemma!="en" & lemma !="tout"] [lemma="sorte"] [tag="ADJ"] [lemma="de"] nous propose 3,845 (0.34 per million) occurrences. On s'attendait, d'après nos hypothèses, à trouver plus d'adjectifs du côté de *espèce.* En effet, pour espèce de ou [lemma="espèce"][tag="ADJ"][lemma="de"], le résultat est de 9,500 (0,83 per million). Il apparaît que *espèce* se trouve bien plus souvent modifié par un adjectif que *sorte*. Ceci se confirme en établissant le rapport de proportion entre les formes avec adjectifs et celles de bases, présenté sur les diagrammes.

Enfin, pour étudier les adjectifs placés entre *sorte* et *de* et entre *espèce* et *de*, nous avons aussi utilisé les expressions [lemma= "espèce"] [] {0,1} [lemma="de"] et [lemma != " tout" & lemma != " en"] [lemma= "sorte"] [] {0,1} [lemma="de"] desquelles sont sorties 261,135 (22.82 per million) occurrences pour *espèce* et 815,079 (71,22 per million) pour *sorte.*

De ce point de vue syntaxique, l'élément le plus nominal est toujours *espèce,* poussant encore *sorte* vers sa fonction d'opérateur sur les autres mots, d'approximant, en perte de sons sens référentiel.

Toutefois, le corpus n'étant pas annoté pour la dépendance, nous avons utilisé un critère positionnel ce qui signifie que les adjectifs trouvés n'ont pas forcément un rapport de dépendance avec *sorte de.*

Nous sommes ensuite revenus à notre première recherche sans adjectif, puis dans « frequency » nous avons sélectionné « 1R » pour s'intéresser à l'élément qui suit le nœud (l'élément recherché). Ce que l'on peut observer en premier, ce sont des guillemets ce qui veut dire que l'on est dans du métalangage.

Synonymie

Notre dernière observation s'est tout simplement portée sur la synonymie via Thesaurus, qui donne les unités qui ont les mêmes propriétés distributionnelles. Pour ce faire, Sketch Engine calcule le nombre de contextes d'apparition similaires.

Visiblement, *espèce* est plus ancré que *sorte* dans son sens nominal, et plus précisément dans la spécification d'un type biologique (animaux, oiseaux…). Pour *espèce, sorte* apparaît comme premier synonyme. Pour *sorte, espèce* apparaît comme cinquième synonyme ce qui nous permet de dire que *sorte* est plus général.

Conclusion

Cette étude, menée sur les contextes d'apparition de deux constructions et motivée par l'hypothèse distributionnelle de Harris, nous a menés à la thèse plus générale que, quand on étudie un corpus, on peut avoir des informations sur la langue que l'on n'atteint pas lorsque l'on creuse uniquement notre intuition.

Nous arrivons à la conclusion que, principalement par sa fréquence et son contexte d'apparition au sein du corpus, *sorte de* est plus grammaticalisé que *espèce de*. En réalité, bien que les deux unités se comportent différemment, elles ont toute deux tendance à évoluer vers la fonction de marqueur d'approximation mais le processus évolue plus rapidement en ce qui concerne *sorte*.

Ainsi, l'étude d'un corpus n'est pas que celle d'un état de langue, mais aussi celle de sa dynamique interne. La langue est en activité, elle bouge, elle n'est pas homogène et cela échappe à notre intuition.