La productivité syntaxique à l'aune de la sémantique distributionnelle

Peter Lauwers

Université de Gand (Universiteit Gent)

Peter.lauwers@ugent.be

Résumé. La *productivité* peut être définie comme le domaine d'application lexical (potentiel) d'une règle morphologique ou d'une construction syntaxique. Après avoir introduit la notion et son application en syntaxe, nous montrerons comment ses différentes facettes peuvent être mesurées au moyen de comptages dans de grands corpus. Puis, nous rappellerons les principes de la sémantique distributionnelle (ou sémantique vectorielle), pour illustrer comment le calcul de la similarité sémantique enrichit l'étude de la productivité. Le tout sera illustré par une étude de cas portant sur les 'minimiseurs' en français, éléments renforçant la négation phrastique (p.ex. *V pas un iota*).

Abstract. Productivity can be defined as the (potential) lexical scope of a morphological rule or syntactic construction. After having introduced the concept and its application to syntax, I show how its various facets can be measured by means of corpus counts. I then review the principles of distributional semantics (or vector semantics), to illustrate how the calculation of semantic similarity enriches the study of productivity. This will be illustrated with a case study on French 'minimisers', elements that reinforce phrasal negation (e.g. *V pas un iota*).

Introduction[1]

La *productivité* d'une règle ou, de manière plus générale, d'un schéma (angl. *pattern*) grammatical, peut être définie comme son domaine d'application lexical (potentiel). Ainsi, la suffixation en *-able* est un schéma productif, qui permet de générer une longue liste d'adjectifs (déverbaux ou dénominaux) bien formés, attestés ou non dans les dictionnaires (Grabar et al. 2006):

> regard**able**, écout**able**, oubli**able**, tax**able**, liquid**able**, déclin**able**, surpass**able**, sélectionn**able**, opér**able**, médaill**able**, éduc**able** …

De même, en syntaxe, le renforcement de la négation phrastique par un *minimiseur*, un nom dénotant une petite quantité, un petit objet, une petite distance ou une petite valeur (Bolinger 1972 : 120 ; Hoeksema 2002) dans la construction [prédicat *pas un* N_{min}], s'avère être un schéma productif:

> Votre contrôle technique automobile ne vous coûtera pas un **centime** (Fr10[10])
>
> Pendant le match, il n'y avait pas un **chat** dans les rues (Fr10[10])
>
> Quand l'un d'eux croise mon regard à plusieurs reprises, je prends bien soin de ne pas bouger d'un **cil** et de me cacher derrière mon Canon (Fr10[10])

Dans le *French Ten Ten corpus 2017* (désormais *Fr10[10]*), un corpus web, Van den Heede (en prép.) en a trouvé pas moins de 124, rien que pour le domaine hexagonal (*.fr*), qui totalise déjà 2,9 milliards de mots. En outre, chacun de ces minimiseurs se combine avec une série de prédicats (296 au total, pour 3102 occurrences), mais pas au même titre. Les uns se combinent avec un grand nombre de types différents et s'étendent facilement à de nouveaux items lexicaux (p.ex *centime*), alors que d'autres sont caractérisés par une ouverture lexicale restreinte (p.ex. *chat*, *cil*). En termes constructionnels, chaque minimiseur se présente donc comme une construction lexicale du type [prédicat *pas un* N_{min}], dont la position prédicative est caractérisée par une productivité variable.

Comme les minimiseurs constituent un paradigme très riche, ils nous fournissent un banc d'essai idéal pour des comparaisons de toute sorte en matière de productivité et d'ouverture sémantique. Ils nous permettront

notamment d'illustrer les grandes questions de ce champ de recherche, ainsi que les méthodes que l'on peut proposer pour les traiter. Cet article se veut donc avant tout un aperçu de la problématique, ainsi qu'un programme de recherche. Pour une analyse plus circonstanciée et plus technique de l'étude de cas des minimiseurs, nous renvoyons le lecteur à Lauwers & Van den Heede (en prép.).

Quelles sont alors ces grandes questions ?

Tout d'abord, on a beau avoir des intuitions sur des différences de productivité, il reste à savoir sur quelle base quantitative on peut la mesurer et comment les différentes mesures proposées se rapportent les unes aux autres et quelles dimensions elles font apparaitre. Ces questions feront l'objet de la section 2, qui fait suite à une première section qui se propose d'initier le lecteur à la problématique de la productivité.

Puis, si la diversité lexicale d'une construction est certainement un indice important de son ouverture, on ne saurait pas non plus négliger sa portée ou, encore sa diversité sémantique, qui semble être une autre facette de l'ouverture d'une construction. Si les deux sont souvent confondues, il nous semble crucial de les dissocier sur le plan méthodologique, afin d'être en mesure d'examiner de manière critique leur relation. Nous verrons ainsi que les deux ne vont pas toujours de pair. Mais, avant toute chose, il faudra donc trouver un moyen de faire entrer en jeu la sémantique, ou, du moins, une certaine approche – quantifiable – de la sémantique, et notamment d'opérationnaliser le concept de *similarité sémantique*. Cette approche quantifiable nous est offerte par la sémantique distributionnelle (ou vectorielle), qui sera au centre de la section 3.

Munis de tous ces outils quantitatifs, nous montrerons comment on peut s'en servir pour s'attaquer à quelques-unes des grandes questions qui se posent concernant les **rapports entre productivité et sémantique**. Ce sera l'objet de la section 4, où nous illustrerons l'usage de notre boite à outils en l'appliquant aux minimiseurs, avant de conclure par un bilan provisoire et une mise en perspective.

1 De la morphologie à la syntaxe

Le concept de productivité (graduelle) provient de la morphologie (1.1), mais a pu se frayer un chemin en syntaxe depuis une petite vingtaine d'années (1.2). Disons d'emblée que le concept est entouré d'un flou artistique, qui a conduit à des usages variés qui se chevauchent en partie (voir Barðdal 2008, chap. 2, pour un aperçu critique).

1.1 En morphologie

A l'origine, le concept de productivité fut utilisé pour distinguer les règles morphologiques capables de générer de nouveaux lexèmes construits des règles morphologiques 'mortes'. Ainsi, le francisant danois Nyrop reconnaît dans sa grammaire historique entre deux types de suffixes :

> Dans *feuillage* et *plumage* tout le monde reconnaît le suffixe *-âge*, comme dans *passage, lavage, tirage;* cependant le *-âge* des premiers mots, qui a un sens collectif, n'est plus productif: on ne forme plus de mots nouveaux en ajoutant *-âge* à des substantifs; il ne s'ajoute dans la langue moderne qu'à des verbes. Donc le type de dérivation représenté par *feuille—feuillage* est mort (il serait absolument impossible de tirer un *fleurage* de *fleur*), mais celui de *passer — passage* est resté vivant et productif: *boycotter — boycottage, lyncher — lynchage* » (Nyrop, 1908, t. 3, § 37).

Avec l'avènement des grands corpus électroniques, cette conception binariste de la productivité a cédé le pas devant une approche graduelle basée sur la quantification (voir les aperçus dans e.a. Dal 2003, Zeldes 2012 et Barðdal et al., sous presse). Ainsi, si le suffixe *-able* est très productif, force est de constater que le suffixe *-ité* s'avère encore plus productif (*fragilité, productivité, bouddhéité, …*). Après avoir parcouru un corpus de 10 millions de mots (occurrences), Grabar *et al.* (2006) en arrivent à un inventaire d'environ 500 types lexicaux différents, alors que *-able* plafonne à 250. Désormais, on est donc amenés à distinguer la « disponibilité » de la règle, c'est-à-dire son aptitude à créer de nouveaux mots (types), de sa « rentabilité » (Corbin 1987 : 42), i.e. sa productivité réalisée (Dal 2003). Rappelons toutefois que la productivité se définit

toujours à l'intérieur des limites imposées par les règles, c'est-à-dire les contraintes imposées aux bases (cf. déjà Nyrop, *supra*). Ainsi, pour former des adjectifs dénominaux en *-u* signifiant 'qui a N', il faut que la base nominale soit monosyllabique et qu'elle désigne une partie inaliénable (Aurnague & Plénat 1997 : 23-24) :

N + *u* : *poilu, barbu, ventru, … briochu, culu, cuissu, ….*

Si le domaine d'application de cette règle est assez circonscrit, on constate qu'elle se prête facilement à des créations inédites ; elle est donc productive.

1.2 En syntaxe

Le concept de productivité a ensuite été étendu à la syntaxe, notamment dans le cadre *usage-based* des Grammaires des constructions d'obédience cognitive. En effet, on peut facilement réinterpréter les règles morphologiques comme des schémas ou constructions (Booij 2010). Ainsi, la règle de la suffixation en *-ité* (*agilité, absurdité, productivité…*), qui décrit une *procédure* à suivre à partir d'un input de type adjectival:

Adj → Adj - *ité*$_N$

peut être réécrite dans un style de notation *déclaratif* (Jackendoff & Audring 2019 : 12) comme un schéma constructionnel contenant une position ouverte (angl. *slots*) qui répond à certaines spécifications :

[_____ $_{Adj}$ - ité]$_N$

De manière cruciale, les positions (semi-)ouvertes de ces schémas constructionnels peuvent être remplies par une série de lexèmes (ou types lexicaux), appelés *fillers* en anglais, qui en déterminent la productivité. De là aux schémas syntaxiques il n'y a plus qu'un pas :

> La construction transitive: [sujet ___$_V$ COD] : *Pierre* ***mange*** *sa soupe*, *Comment continuer à* ***binge-watcher*** *quand on devient parent ?* (journal *20 minutes*, 16 octobre 2017, page 22 ; Wiktionnaire, s.v. binge-watcher[2])

> Les locutions prépositives du type [*sous le* ____ $_N$ *de* SN] : *sous le* ***contrôle*** *de, sous la* ***direction*** *de, sous la* ***baguette*** *de …* (Lauwers 2014)

A leur tour, toutes ces constructions syntaxiques comportent au moins une position (semi-)ouverte qui accueille un certain nombre de types lexicaux qu'on peut compter. Il s'ensuit que toute construction syntaxique a son propre domaine lexical (potentiel). Toute construction a une productivité plus ou moins grande (Barðdal 2008 ; Zeldes 2012) et à comparer la productivité de constructions sœurs au sein d'une même famille, on voit apparaitre des différences considérables. Ainsi, pour les verbes attributifs (Van Wettere 2018: 639), on voit que les courbes du nombre de types et d'hapax (= types qui n'apparaissent qu'une seule fois) divergent pas mal, notamment si le nombre d'occurrences vues (en abscisse) augmente :

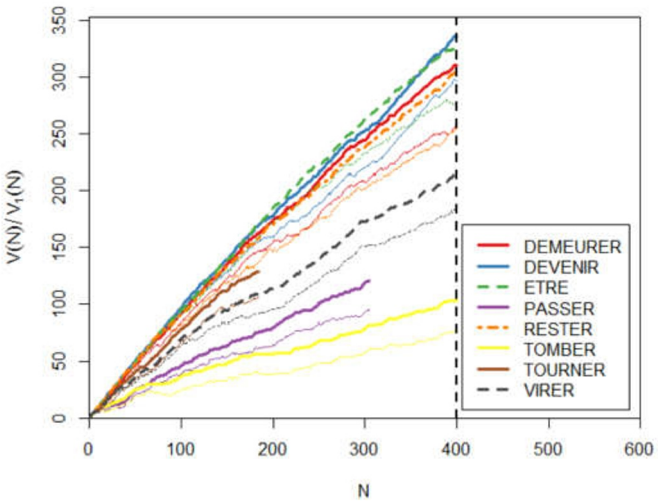


Figure 1. Courbe d'accroissement du nombre de types (continu) et d'hapax (pointillé)

Certes, pour certains, le problème de la productivité en syntaxe peut paraitre un faux problème. Notamment en grammaire générative, les règles syntaxiques sont censées être régulières, donc productives par définition, les exceptions et idiosyncrasies étant stockées dans le lexique (mental) (Pinker 1999: 19). Cette idée de la productivité/régularité des règles remonte à Chomsky, qui a dès le départ mis l'accent sur la capacité générative, donc « productive » - qu'il qualifie aussi de « créativité » - des règles de la grammaire (Chomsky 1965 : 6). Or, les recherches sur corpus et notamment dans le cadre des modèles "usage-based" (p.ex. Bybee 2006), selon lesquels l'exposition à l'usage façonne la compétence du locuteur (ses représentations mentales), ont montré que souvent la productivité absolue n'existe pas. On se rappellera à cet égard l'étude de Goldberg (1995 : ch. 5) sur la construction ditransitive en anglais, qui aboutit à la reconnaissance de micro-classes de verbes, obtenues à partir de généralisations locales, à productivité variable. Zeldes précise l'enjeu des études sur la productivité en syntaxe en ces termes :

> if all syntactic rules can generate an infinity of utterances from a limited vocabulary of morphemes, words or constructions, why is it that we are inclined to use some rules more often (but not exclusively) with familiar material, which is presumably already stored in memory, while other syntactic patterns tend to fill out their empty slots with items the speaker has never before used or heard in that position? Do speakers know when they are 'allowed' to make use of the infinitely productive facility of syntax and when they 'should' repeat familiar forms? (Zeldes 2012 : 1)

Notons que la même difficulté de circonscrire la productivité en termes de contraintes abstraites se pose également en morphologie:

> il existe des procédés peu contraignants envers les bases et pourtant inaptes à former de nouveaux mots (Aronoff 1976 et Scalise 1984 : 157 citent le cas du suffixe *-ous* anglais); inversement, un suffixe comme *-u* a déjà montré qu'un procédé peut contraindre fortement les bases qu'il sélectionne tout en étant apte à former de nouvelles unités lexicales (Dal 2003 : 11)

On se rappellera aussi la discussion dans Baayen & Renouf (1996) : la productivité n'est pas simplement la résultante du nombre de contraintes pesant sur la règle.

2 Mesurer la productivité

Une fois que la productivité est apparue comme une notion gradable, les morphologues ont cherché à la quantifier. Nous nous limiterons ici à signaler les travaux fondateurs de Harald Baayen (et collègues) (Baayen 1992, 1993, 2009), qui ont mis en place les principales mesures quantitatives basées sur l'examen de grands corpus électroniques (voir Dal 2003 pour un aperçu).

2.1 Les mesures : aperçu

La mesure la plus intuitive consiste à compter le nombre de types différents auxquels une construction a été appliquée, la *fréquence-type* (angl. *type frequency*), qui se distingue de sa *fréquence-token* (angl. *token frequency*), c'est-à-dire le nombre d'occurrences de la construction. Comme le nombre de types dépend en grande partie du nombre d'occurrences vues, la fréquence-type est souvent remplacée par le ratio *TypeToken*, le nombre de types (différents) n par nombre d'occurrences (= *tokens*) m^3. Il s'agit donc du nombre de types réalisés (« realized productivity »), qui quantifie l'étendue lexicale (« extent of use », Baayen 1993: 181). Or, l'usage attesté comporte aussi des indices sur le comportement futur des locuteurs : une construction qui a été observée avec beaucoup de types différents donnera aux locuteurs la confiance de l'étendre à des types qu'ils n'ont jamais entendus ou lus (Goldberg 2019).

Toutefois, cette dimension potentielle de la productivité, c'est-à-dire l'extensibilité de la construction, est mieux captée par une autre mesure, à savoir le ratio *HapaxToken*, le nombre de types 'nouveaux' dans le corpus, à fréquence 1, donc d'hapax, par n occurrences. Traduit en termes probabilistes, c'est la probabilité avec laquelle de nouveaux types seront réalisés (Baayen and Renouf 1996 : 73-74, d'après une idée avancée par Bolinger). Dans cette optique, les hapax sont la preuve la plus directe du potentiel de la construction - même si cela reste, tout compte fait, un 'proxy', puisqu'on extrapole le possible à partir du réalisé - dans la mesure où tout hapax est supposé être le résultat de l'application du schéma constructionnel, 'in situ' pour ainsi dire. En effet, ces réalisations rares ont peu de chances d'avoir été stockées en mémoire. Si un tel schéma constructionnel se trouve souvent activé de cette manière, il a de fortes chances de générer encore de nouveaux types non attestés à l'avenir.

Enfin, le ratio *HapaxType* est le rapport entre le nombre d'hapax et le nombre de types. Il est moins souvent utilisé car, comme le signale Van Wettere (2022 : 169), certains chercheurs le trouvent redondant puisqu'il s'appuie sur les deux autres quantités ; d'autre part, son interprétation linguistique parait moins évidente.

Ces trois mesures captent déjà différentes facettes de l'ouverture lexicale de la construction. Or, la problématique peut tout aussi bien être approchée par son versant négatif, c'est-à-dire par le biais de la conventionnalisation, qui passe pour une entrave à la diversité lexicale et qui est donc signe 'd'anti-productivité'. Dès lors, les trois autres mesures que nous allons proposer concernent toutes la 'hauteur' ou le 'poids' relatif du sommet de la distribution fréquentielle des types. Ainsi, on pourrait inclure la fréquence du prédicat le plus fréquent (*FrTop1*), la fréquence moyenne des trois types les plus fréquents (*MoyenneFrTop3*) et leur écart-type (*DSTop3*), cette dernière mesure étant un indice du rythme avec lequel la fréquence des types-phare diminue. Notons que cette dernière mesure est très proche de la mesure « Zipf Alpha », qui modélise le spectre de fréquence *dans son ensemble* (Van den Heede & Lauwers 2023). Elle correspond à la pente de la fonction zipféenne log-transformée entre les fréquences et les rangs. Comme la qualité de l'ajustement linéaire obtenu à partir des données brutes varie d'un minimiseur à l'autre et que cette mesure est de toute manière fortement corrélée à la fréquence des types les plus fréquents (Van den Heede & Lauwers 2023), nous ne l'incluons pas dans notre étude de cas.

En revanche, nous incorporons à l'analyse encore une dernière mesure, également basée sur la forme de la courbe fréquentielle : l'entropie. Dans le domaine de la théorie de l'information, l'entropie Shannon correspond à "the average amount of uncertainty of a random variable: the larger H or Hrel, the more random a distribution is » (Gries 2010 : 272). L'entropie (et donc l'incertitude, la surprise)[4] est maximale si la distribution des types est uniforme, c'est-à-dire si chaque type a une probabilité d'occurrence égale. Par contre, l'entropie est nulle si un seul type absorbe toutes les occurrences de la construction ; l'incertitude ou la surprise est alors réduite à zéro. Il s'ensuit que l'entropie peut être considérée comme une mesure de

productivité (e.a. Gries 2015 : 514 ; voir Pankratz *et al.* 2022 pour d'autres références) : si les occurrences se regroupent sur quelques types hautement fréquents, l'entropie sera réduite. Comme l'entropie est une quantité non bornée, elle augmente aussi avec le nombre de types. Elle est uniquement bornée par la taille constante de nos échantillons (100 occurrences), qui plafonne le nombre maximal de types à 100 (et donc les probabilités minimales à 0.01), ce qui donne une entropie maximale de 6,64 (= log2(n), n correspondant à 100 types ici). Cela nous offre donc une autre base intéressante pour la comparaison de la productivité des minimiseurs.

2.2 Les rapports entre les mesures

On s'accorde en général à dire que la fréquence-type et le ratio *HapaxToken* correspondent à deux dimensions différentes du « complexe de productivité », qui, en morphologie, peuvent facilement donner lieu à des résultats différents si on cherche à classer des affixes en termes de leur productivité (Baayen & Lieber 1991 ; Zeldes 2012 : 86-89). Il reste à voir ce qu'il en est en syntaxe. Ainsi, dans l'étude de cas discuté dans Zeldes (2012 : 124), on voit que les deux mesures sont étroitement corrélées, tout comme chez Barðdal *et al*. (sous presse). Le ratio *HapaxType*, quant à lui, semble faire bande à part dans le domaine des semi-copules (Van Wettere 2021), mais pas dans le domaine des auxiliaires inchoatifs espagnols (Van Hulle, Enghels & Lauwers, s.p.p.). Son statut reste donc incertain. Puis, les mesures d'anti-productivité, quant à elles, seront probablement étroitement liées entre elles, et anti-corrélées au ratio *TypeToken*. Enfin, il reste à voir jusqu'à quel point l'entropie leur emboîte le pas. Dans ce qui suit, nous allons donc examiner les corrélations entre les différentes quantités mesurées pour la famille des minimiseurs.

2.3 Illustration : les minimiseurs

Pour appliquer ces mesures aux minimiseurs, nous devons nous limiter aux 19 minimiseurs qui comptent au moins 100 attestations dans le corpus. Il faut en effet que les minimiseurs soient suffisamment enracinés cognitivement parlant (angl. *entrenched*) pour fonctionner comme constructions et qu'ils comptent un nombre suffisant d'occurrences pour autoriser un traitement quantitatif. En outre, il fallait en tirer des échantillons de taille égale (n = 100) afin d'aboutir à une comparaison équitable de la productivité (Gaeta & Ricca 2006 ; Zeldes 2012)[5]. Regardons à titre illustratif le dégradé des ratios *TypeToken* et *HapaxToken* :

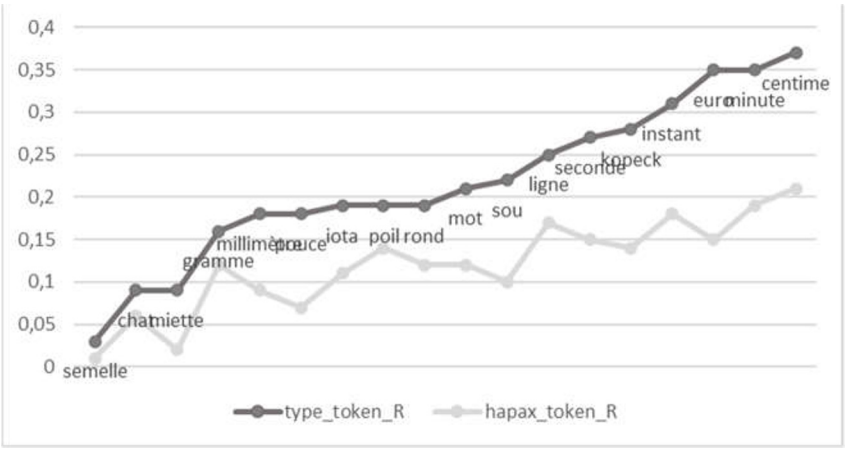


Figure 2. Ratios *TypeToken* et *HapaxToken* des minimiseurs

Afin d'examiner les corrélations entre les différentes mesures et de comparer entre eux les minimiseurs, on peut mener une analyse en composantes principales (ACP ; Kassambara 2017] ; R FactoMineR [Lê et al. 2008] et factoextra [Kassambara et Mundt 2020]), à l'instar de Desagulier (2016) et Van Wettere (2021).

La Figure 3 présente le cercle de corrélation dans lequel les différentes mesures sont représentées comme des vecteurs dans un espace réduit à deux dimensions :

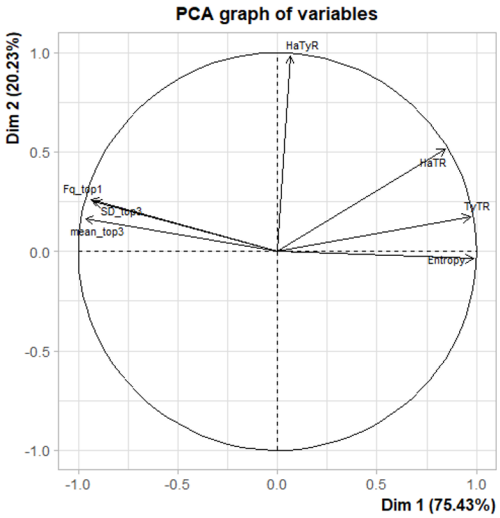


Figure 3. Les corrélations entre les mesures de productivité (Analyse en Composantes Principales)

Ces deux macro-dimensions latentes, appelées « composantes principales », expliquent à elles seules plus de 95 % de la variabilité. La position de la pointe des flèches visualise à quel point les mesures sont associées aux deux composantes principales. Si deux vecteurs sont proches, cela signifie que les mesures sont étroitement corrélées.

Comme on pouvait s'y attendre, les mesures de conventionnalisation sont très proches les unes des autres ; elles se positionnent dans la moitié gauche du cercle. Elles sont positivement corrélées avec la première dimension et anti-corrélées[6] avec le ratio *TypeToken*, l'*entropie* et le ratio *HapaxToken*, qui sont situées à droite du cercle, le long du même axe. Ces deux faisceaux de mesures pointent donc vers la même dimension latente : le degré d'ouverture lexicale (ou productivité).

Notons que le ratio *HapaxToken* suit de très près le ratio *TypeToken* (r = 0,92) : si une construction compte beaucoup de types différents, cela se traduit aussi par un grand nombre d'hapaxes. Il est également à noter que l'*entropie* coïncide presque avec le ratio *TypeToken* (r = 0,96). On peut donc dire que l'entropie vient compléter la liste de mesures qui mettent en évidence l'étendue lexicale de la construction, tout comme le ratio *TypeToken* et les mesures d'*anti-productivité* [r = -0,90 (FqTop1), -0.99 (FqTop3), -0,87 (SDTop3)].

La deuxième macro-dimension latente, qui recueille les variables qui sont maximalement différentes de celles liées à la première dimension, est entièrement vouée au ratio *HapaxType*. On comprend que cette variable composite est tiraillée entre les deux autres quantités, qui constituent son numérateur et son dénominateur, respectivement, qui lui donnent un profil différent. Toutefois, il est intéressant de faire remarquer que le ratio *HapaxType* suit de plus près le nombre d'hapax que le nombre de types (r = 0,54 vs 0,21). Cette mesure est donc davantage liée à l'ouverture potentielle de la Cx. En effet, on pourrait y voir un autre aspect de la productivité, dont l'importance est sans doute sous-estimée. Le ratio *HapaxType* focalise la dichotomie entre hapax et non-hapax au sein du pool de types et annule presque l'effet du poids des types très fréquents (Van Wettere 2021 : 410). En voici un exemple : *pas un poil* compte encore 14 hapax parmi les 16 types restants « derrière » un top 3 particulièrement lourd, qui absorbe à lui seul 81 des 100 occurrences de l'énchantillon) :

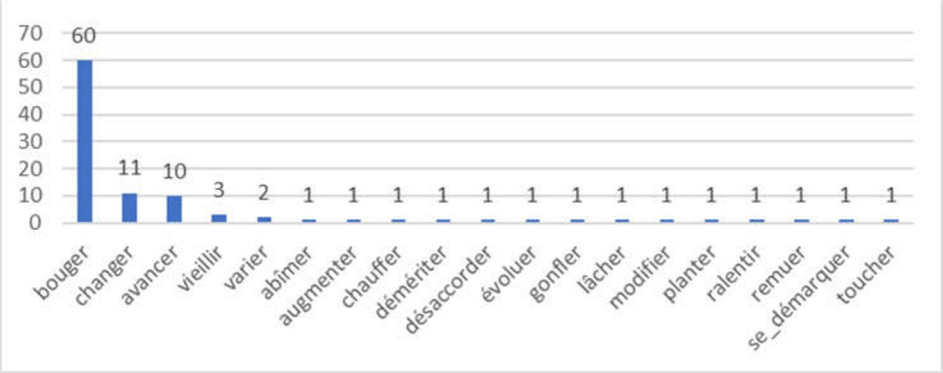


Figure 4. Fréquences des types de *pas un poil*

Cela donne des ratios *TypeToken* (0,19) et *HapaxToken* (0,14) plutôt bas ou moyens, alors que le ratio *HapaxType* est le deuxième ratio le plus élevé des 19 minimiseurs (0,74). Malgré son sommet particulièrement lourd, cette construction présente encore une remarquable diversité dans la partie la plus innovante du spectre fréquentiel, à savoir dans la queue du peloton des types. On pourrait parler ici d'une espèce de « productivité secondaire » que l'on risque de voir passer sous les radars si on ne prend pas en considération le ratio *HapaxType*. On peut en conclure qu'une construction qui se retrouve souvent avec les mêmes types, ne cesse pas nécessairement d'être innovatrice dans le reste de son spectre, comme nous l'avons aussi constaté pour les minimiseurs en néerlandais (Van den Heede & Lauwers 2023). Cela contredit la thèse de Bybee (Bybee 1995 : 433-435) et une série d'autres auteurs (mentionnés dans Barðdal 2008 : 49) , selon laquelle la haute fréquence est une entrave à la productivité. Dans l'optique de Bybee, sur le plan cognitif, cela correspond à l'idée qu'une réalisation fortement conventionnalisée d'une construction est fortement ancrée en mémoire (angl. *entrenchment*), au point de devenir une construction lexicalisée autonome, détachée de la construction-mère. Il s'ensuivrait que ces occurrences ne contribuent plus à la schématisation de la construction-mère (Barðdal 2008 : 49) et donc à sa productivité. Toutefois, le fait de « ne plus compter pour la cx-mère » pourrait tout aussi bien être interprété autrement : si une sous-construction qui tend à se lexicaliser est mentalement détachée du reste du spectre, pourquoi aurait-elle encore une influence sur les types restants de la construction-mère (cf. également Feltgen 2017 : 261-262)? Celle-ci pourrait toujours se montrer innovatrice, extensible donc, et le ratio *HapaxType* pourrait en rendre compte.

On peut maintenant projeter les minimiseurs sur cet espace bidimensionnel :

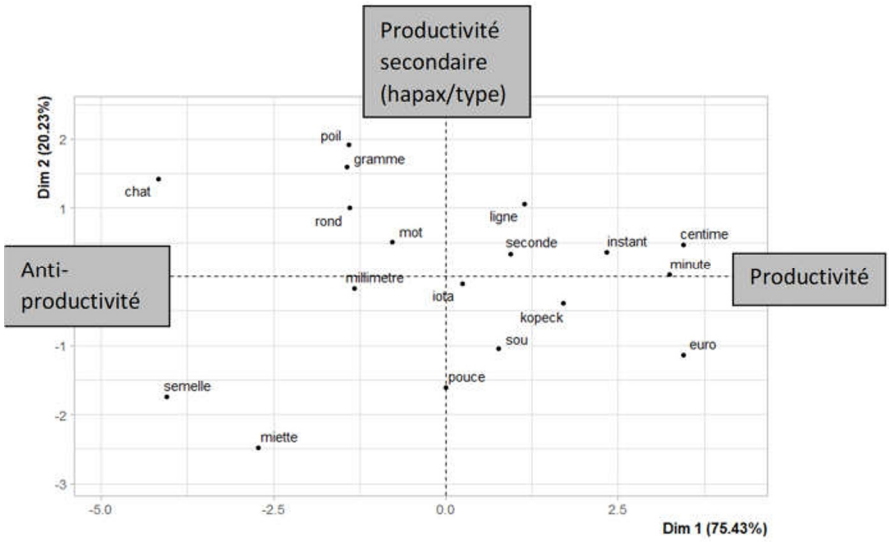


Figure 5. Positionnement des minimiseurs dans l'espace bidimensionnel des deux composantes principales

Les minimiseurs se positionnent le long de l'abscisse, qu'on peut interpréter comme l'axe de l'ouverture lexicale (= la productivité), avec à gauche, *chat*, le moins productif, et à droite, le trio *centime, minute* et *euro*, les plus productifs. L'ordonnée, quant à elle, reflète essentiellement la variation en termes de productivité secondaire (le ratio *HapaxType*), avec *poil* et *gramme* comme premiers de classe.

3 Quantifier la sémantique : entre la sémantique distributionnelle

La productivité a pendant longtemps été une affaire de comptages de mots. Face à cet empressement quantitatif, la sémantique, dont on avait petit à petit reconnu la pertinence pour la problématique de la productivité (Bybee & Eddington 2006 ; Barðdal 2008), faisait plutôt pâle figure. Toutefois, ces dernières années, et notamment depuis la percée de la sémantique distributionnelle en linguistique descriptive (*sémantique vectorielle, word-embeddings*, etc. ; Erk 2012 ; Jurafsky & Martin 2024), des concepts comme la *similarité* et la *densité sémantiques* sont entrés de plein pied dans le domaine des études sur la productivité (Suttle & Goldberg 2011 ; Perek 2016 ; Goldberg 2019 ; Van den Heede & Lauwers 2023).

3.1 Sémantique distributionnelle : principes et chaine de traitement

A la base de l'approche réside l'hypothèse distributionnelle, dont on connait l'importance dans le domaine de la syntaxe: "you shall know a word by the company it keeps" (Firth 1957). Dans son application au lexique, elle stipule que les mots à sens similaire apparaissent dans des contextes phrastiques similaires, c'est-à-dire avec les mêmes mots contextuels, les mêmes co-occurrents (angl. *collocates*). Connue depuis longtemps en TAL, la méthode a désormais aussi fait ses preuves dans la recherche fondamentale. On a pu montrer que les valeurs de similarité sémantique obtenues au moyen de la méthode distributionnelle suivent d'assez près les intuitions d'annotateurs humains (voir e.a. Mandera *et al*. 2017 ; Joubarne & Inkpen 2011).

Les mesures sémantiques que nous utiliserons ici sont basées sur des espaces vectoriels générés à partir d'un autre grand corpus web, le *French COW Corpus* (Schäfer 2015 ; Schäfer et Bildhauer 2012), un corpus qui peut être téléchargé librement. Pour nos besoins, nous n'avons utilisé qu'un tiers du corpus, totalisant 47 millions de phrases, soit 994 millions d'occurrences lemmatisées. La partie 'TAL' de la chaine de traitement a été exécutée entièrement sur R Studio à l'aide du package Quanteda (Benoit et al. 2018), même s'il a d'abord fallu extraire les colonnes pertinentes du corpus à l'aide de quelques lignes de code Python. Le corpus a été tokenisé et considérablement nettoyé et élagué. En effet, il fallait éliminer les lemmes

incertains (en partie à l'aide du *Lexique* de B. New), les caractères spéciaux (*ù, $,* * etc.), nombres et lettres, les mots-outils et les noms propres (et leurs dérivés) rares. Pour une analyse distributionnelle du sens, ces éléments contextuels ne sont que du bruit en ce qu'ils ne révèlent rien ou peu sur le sémantisme des mots qu'ils entourent. Une fois le corpus élagué, nous avons procédé au comptage des co-occurrents, utilisant différentes fenêtres (3 ou 5 mots avant le mot pivot, ou encore, la phrase entière), ce qui a donné lieu à différentes matrices de co-occurrence. La méthode choisie, appelée « Bag of Words », est très simple, sans pondération en fonction de la distance par rapport au pivot, sans rembourrage des positions laissées vides suite à l'élagage (*padding*), et surtout, sans étiquetage syntaxique. En voici un extrait d'une telle matrice de co-occurrence générée à l'aide de Quanteda :

features	ami	mûr	aboyer	miauler	jardin	bâtiment	mordre	morsure	étudier	crotte
chat	468	6	47	135	289	12	88	47	37	19
chien	715	4	1028	17	306	40	684	248	50	303
banane	16	106	0	0	14	2	2	0	2	0
université	383	10	0	0	195	722	3	1	2501	2

Figure 6. Extrait de la matrice des co-occurrences brutes

Dans cette matrice, chaque mot du lexique (lemme) correspond donc à un vecteur composé des fréquences de cooccurrence avec 50 000 cooccurrents, le nombre de coocurrents (les plus fréquents) étant un autre paramètre à ajuster (Jurafsky & Martin 2024: 9-10). Ces fréquences brutes peuvent être considérées comme des coordonnées dans un espace sémantique de 50 000 dimensions.

Cette matrice va servir de base aux calculs de similarité - ou, son complément, de distance - sémantique. On peut visualiser les distances entre les mots dans un système de coordonnées cartésiennes. Si on se limite à deux dimensions, en l'occurrence les cooccurrents *jardin* et *bâtiment*, on voit que *chat* et *chien* se rapprochent :

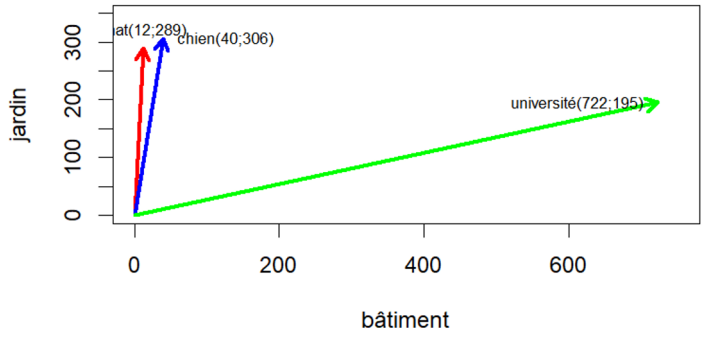


Figure 7. Visualisation de la distance entre *chat, chien* et *université* en fonction de deux mots contextuels, *jardin* et *bâtiment*

La distance géométrique entre deux mots correspond alors au cosinus de l'angle entre leurs vecteurs respectifs, ce qui donne une valeur entre 1 (les vecteurs coïncident, l'angle est égal à zéro, donc cosinus = 1, similarité maximale) et 0 (les vecteurs sont orthogonaux, le cosinus de 90° = 0, donc similarité nulle). En termes algébriques, la distance entre deux vecteurs correspond à leur produit scalaire, divisé par le produit de leurs longueurs respectives afin de normaliser la longueur inégale des vecteurs (Jurafsky & Martin 2024 : 11) :

$$\text{cosine}(\mathbf{v},\mathbf{w}) = \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

Figure 8. Formule pour le calcul de la distance cosinus

Toutefois, pour donner aux cooccurrents les plus informatifs plus de poids, nous effectuons d'abord une pondération d'après leur PPMI (*Positive Pointwise Mutual Information Score*), qui évalue les fréquences de cooccurrence brutes par rapport à la fréquence globale du mot et du cooccurrent. Voici la formule pour calculer le score PMI pour un mot w et un co-occurrent c, P étant la probabilité d'occurrence (voir Jurafsky & Martin 2024 : 14, pour des références plus techniques):

$$\text{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$$

Figure 9. Formule pour le calcul du PMI (*Pointwise Mutual Information Score)*

Il s'ensuit aussi des scores d'attraction parfois négatifs, mais ceux-ci sont remplacés par des zéros (d'où « positive » PMI) (Heylen et al. 2015: 156 ; Jurafsky & Martin 2024 : 14).

À ce stade, les matrices de cooccurrence sont toujours très éparses, ce qui signifie qu'elles contiennent beaucoup de zéros. Pour réduire le bruit, on a aussi procédé à une réduction de dimensionnalité au moyen de la décomposition en valeurs singulières (qui, soit dit en passant, est également le mécanisme qui sous-tend l'ACP utilisée ci-dessus). Cela donne des vecteurs denses à 200 ou à 300 dimensions. Le R package Wordspace (Evert 2014) se révélé particulièrement utile ici. Le résultat est une matrice dense de 50 000 lignes et 200 ou 300 colonnes, selon le cas. Ces vecteurs denses sont plus performants dans la mesure où ils captent mieux les similarités entre les mots contextuels (voir e.a. Jurafsky & Martin 2024 : 17). Ainsi, *vélo* et *bicyclette* constituent deux dimensions séparées, mais il va de soi qu'ils s'inscrivent, avec d'autres items, dans une dimension latente plus abstraite. Plusieurs de ces modèles vectoriels ont ainsi été générés, d'après plusieurs paramètres. Ce n'est qu'à ce moment-là que les similarités cosinus peuvent être calculées (R package *svs*, Plevoets [2015]). Le résultat est une matrice de similarité (variant de 0 à 1), qu'on peut facilement convertir en matrice distancielle, en prenant le complément des quantités obtenues. Ainsi, une distance zéro égale une similarité parfaite, alors qu'une distance de 1 correspond à une distance maximale :

	clavier <dbl>	percussion <dbl>	banal <dbl>	reprendre <dbl>	d'abord <dbl>	humeur <dbl>	planter <dbl>	fleur <dbl>	rose <dbl>
clavier	0.0000000	0.3382645	0.6022244	0.5859063	0.6496648	0.6145732	0.6751219	0.7774339	0.6964206
percussion	0.3382645	0.0000000	0.6928148	0.6555286	0.7291226	0.6714578	0.7612903	0.7838160	0.7465066
banal	0.6022244	0.6928148	0.0000000	0.5058957	0.4937162	0.3434148	0.5836502	0.6180577	0.5512639
reprendre	0.5859063	0.6555286	0.5058957	0.0000000	0.1720796	0.4992277	0.5687507	0.6588395	0.6142459
d'abord	0.6496648	0.7291226	0.4937162	0.1720796	0.0000000	0.5474118	0.6042011	0.6287646	0.6171070
humeur	0.6145732	0.6714578	0.3434148	0.4992277	0.5474118	0.0000000	0.6347815	0.6136275	0.5597522
planter	0.6751219	0.7612903	0.5836502	0.5687507	0.6042011	0.6347815	0.0000000	0.2717842	0.3739351
fleur	0.7774339	0.7838160	0.6180577	0.6588395	0.6287646	0.6136275	0.2717842	0.0000000	0.1181079
rose	0.6964206	0.7465066	0.5512639	0.6142459	0.6171070	0.5597522	0.3739351	0.1181079	0.0000000

Figure 10. Extrait de la matrice distancielle (modèle 50K_W5_200)

Enfin, nous avons choisi le modèle le plus performant (50 000 lemmes ; une fenêtre de 2 fois trois mots contextuels ; 200 dimensions), c'est-à-dire le modèle qui donnait les meilleures corrélations (r = -0.73 ; p= 3.19e-12) avec les scores donnés par des évaluateurs humains pour une liste de 65 paires de lemmes (Joubarne & Inkpen 2011).

3.2 Sparsité (ou densité) sémantique

Maintenant que nous sommes en mesure de calculer les distances cosinus entre les types lexicaux d'une construction donnée[7], il reste à mettre cette information au service de la question qui nous occupe, à savoir la diversité sémantique des constructions et son rapport à la productivité. Il faudra donc en tirer des mesures opératoires.

Une première possibilité est de compter sur les bons services d'un algorithme de classification automatique (= clustering) de toutes les distances cosinus entre les types attestés, tous minimiseurs confondus. Ainsi on obtient des îlots (= clusters) de prédicats sémantiquement cohérents. Ensuite, on peut projeter les types attestés dans chaque construction sur cette carte onomasiologique pour déterminer le nombre de clusters (groupes) couverts par chaque construction (cf. Van den Heede & Lauwers 2023). Une construction à portée sémantique plus large est représentée dans un nombre plus grand de groupes sémantiques. Cette méthode n'est cependant pas optimale, étant donné que le nombre de clusters couvert par chaque construction suit de très près le nombre de types. En outre, le résultat dépend en partie du nombre de clusters choisis : en général, plusieurs options sont possibles et également bonnes sur le plan statistique, ou, plus souvent, également mauvaises (les distances entre les types étant petites, leur regroupement est parfois compliqué). C'est pourquoi nous ne l'utiliserons plus ici. Une manière plus directe de mettre à profit les distances cosinus entre les types d'une construction donnée consiste à calculer la densité ou encore, la sparsité moyenne d'une construction donnée. Pour cela, il suffit de prendre la moyenne des distances entre les types d'une construction (deux à deux). Plus la moyenne est élevée, plus les types sont éparpillés dans l'espace sémantique. Il s'ensuit un spectre peu dense, signe d'une grande diversité sémantique. En revanche, si la moyenne est basse, le spectre est dense, pointant vers une faible diversité sémantique. Nous avons déjà appliqué cette mesure aux données du néerlandais (cf. Van den Heede & Lauwers 2023), mais ici nous la perfectionnerons au moyen d'une procédure Monte Carlo (cf. *infra*). En outre, nous étendons son application au domaine des hapax pour calculer la densité/sparsité moyenne de leur entourage sémantique. Les résultats de ces mesures seront appliqués dans la section 4, où ils seront d'emblée confrontés aux indices de productivité.

4 Productivité et sémantique : application à l'étude de cas

Dans ce qui suit, nous appliquerons donc nos outils au cas des minimiseurs en français pour répondre à une série de questions relatives aux rapports entre productivité et sémantique (4.1). Comme il s'agit surtout d'illustrer la problématique, nous nous limiterons aux constats les plus importants de cette étude empirique (4.2). Pour une analyse plus élaborée et plus technique, notamment pour ce qui est des procédures Monte Carlo, nous renvoyons le lecteur à Lauwers & Van den Heede (en prép.).

4.1 Questions de recherche et opérationnalisation

Pour examiner les rapports complexes entre ouverture/diversité sémantique et productivité, nous essaierons de répondre à trois questions, auxquelles seront d'emblée associées trois hypothèses, basées sur la littérature. La première question concerne les types en général, alors que la deuxième et la troisième question visent les particularités des entourages hapaxiques. Rappelons que les hapax peuvent être considérés comme des poches d'innovation lexicale où la construction productive peut être observée lorsqu'elle est « à l'œuvre ».

Question 1 : Quelle est la relation entre la productivité réalisée (~ fréquence-type) et la diversité sémantique (~ sparsité sémantique) ?

Hypothèse : On s'attend à ce qu'il y ait un lien étroit entre les deux dans cette famille de constructions comparables. Les constructions les plus productives (en termes de fréquence-type) ont de bonnes chances d'être les plus diversifiées sur le plan sémantique. Tout d'abord, parce qu'une plus grande diversité sémantique (comme corollaire d'une plus grande schématicité, Langacker 1987, i.e. degré d'abstraction),

soit une plus grande diversité de situations d'emploi, implique a priori aussi une compatibilité avec un plus grand nombre de types lexicaux, à la fois dans le domaine des évolutions diachroniques (Trousdale 2008 : 170-171, *apud* Perek 2020 : 144), mais aussi en synchronie (Barðdal 2008).

Question 2 : Les extensions d'une construction (= hapax) sont-elles sémantiquement plus denses que le reste du spectre?
En d'autres termes, est-ce que la similarité sémantique avec les types environnants est plus importante dans les zones 'innovantes' du spectre ? Cela suggérerait que l'extension se réalise essentiellement par analogie sémantique, de proche en proche, avec ce qui existe déjà.

Hypothèse : oui, on s'attend à ce que ce soit le cas, d'après les travaux diachroniques de Perek (2016 : 182): les zones les plus denses (avec de nombreux types étroitement liés) ont plus de chances d'attirer de nouveaux types, c'est-à-dire des types qui apparaissent pour la première fois dans la construction, dans la période subséquente. Le mécanisme qui sous-tend ce constat est comme suit : c'est dans ces zones-là que ces nouveaux types rencontrent la plus grande 'couverture' (angl. *coverage*), d'après la terminologie de Suttle & Goldberg (2011 ; cf. aussi Goldberg 2019), c'est-à-dire les zones où, cognitivement parlant, les locuteurs peuvent le plus facilement créer une généralisation sémantique locale qui couvre et donc légitime le nouvel item.

Question 3 : les constructions qui ont moins de types (et souvent aussi moins d'hapax, cf. *supra* 2.3.) **doivent-elles nécessairement s'étendre de manière plus cohérente que les constructions avec beaucoup de types, qui, elles, ratissent très large sur le plan sémantique, y compris pour leurs hapax?** En d'autres termes, **l'entourage sémantique de leurs hapax est-il plus dense?**

Hypothèse : oui, d'après Barðdal (2008), la cohérence sémantique est une condition *sine qua non* à l'extensibilité pour les constructions qui ne comptent que peu de types attestés. Notre méthode cherche donc à opérationnaliser les prédictions de son modèle théorique.

A travers nos analyses, nous exploiterons le concept de *sparsité sémantique* de deux manières différentes :

> OBS_SPARS_GLOB : la sparsité globale (observée), soit la distance cosinus moyenne par paire entre tous les types de la construction
> OBS_SPARS_ENV_HAPAX : la sparsité moyenne des environnements hapaxiques (observée), soit la distance cosinus moyenne des hapax par rapport à leur(s) plus proche(s) voisin(s) dans la construction.

Ces moyennes absolues seront complétées par une procédure d'échantillonnage aléatoire de type Monte Carlo qui appréhende l'idée de sparsité en termes plus relatifs, i.e. par rapport à ce que l'on est en droit d'attendre étant donné un nombre donné d'hapax et de types (c'est-à-dire le nombre d'hapax et de types de la construction en question). De cette manière, nous cherchons à contrôler pour certains effets liés au nombre de types sur lesquels on prend la moyenne[8]. Concrètement, cette procédure consiste à extraire 100 ou 1000 échantillons aléatoires du pool des 285 types attestés dans le corpus, c'est-à-dire qui apparaissent au moins une fois avec l'ou au l'autre minimiseur. Elle aboutit à des quantités standardisées, notées en *score-z* (ou *z-score*), portant le même nom (pour des détails, voir Lauwers & Van den Heede, en prép.):

> Z_SPARS_GLOB
> Z_SPARS_ENV_HAPAX

Ainsi, Z_SPARS_GLOB compare la sparsité globale observée à 1000 échantillons de n types pris dans le pool des types. Z_SPARS_ENV_HAPAX fait de même pour 100 échantillons de n types dont on extrait ensuite un nombre m d'hapax égal au nombre d'hapax de la construction en question pour en calculer la sparsité des voisinages (1, 2, 3 voisins etc.). Dans les deux cas, on détermine à quel point les moyennes observées sont « spéciales » par rapport aux attentes qu'on peut avoir pour une construction avec n types (et m hapax).

4.2 Résultats

Dans ce qui suit, nous reprendrons les 3 questions dans l'ordre.

Question 1 : Quelle est la relation entre la productivité réalisée (~ fréquence-type) et la diversité sémantique (~ sparsité sémantique) ?

Examinons tout d'abord la diversité sémantique. La Figure 11 montre que les minimiseurs présentent un beau dégradé pour la *sparsité sémantique globale*, qui va de 0,34 à 0,60 (0 = similaire ; 1 = différent)[9]:

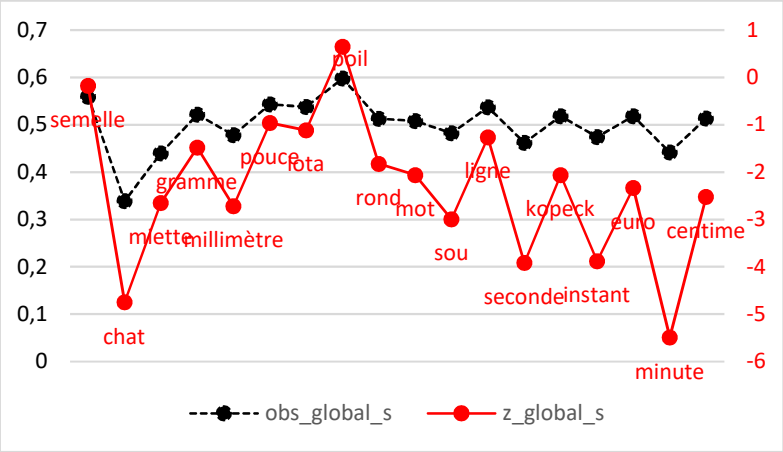


Figure 11. OBS_SPARS_GLOB et Z_SPARS_GLOB

Notons que la correction par z-score a un impact très réduit (r = 0,87).
Quant à la question de savoir si l'ouverture lexicale est corrélée à l'ouverture sémantique, telle que celle-ci est représentée par les taux de sparsité globale (Z_SPARS_GLOB), il suffit de projeter les ratios *TypeToken* de la Figure 2 sur la Figure 11 :

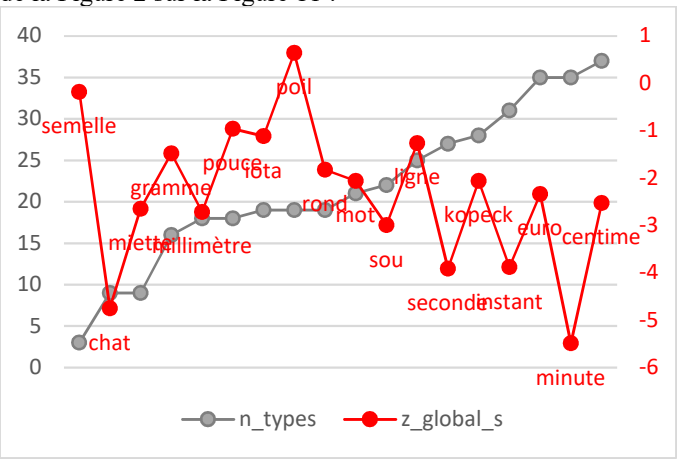


Figure 12. Productivité réalisée (Type_fq) et sparsité globale (Z_SPARS_GLOB)

Nous devons conclure que les deux dimensions, la diversité lexicale et la diversité sémantique, ne sont pas vraiment liées. Leur corrélation est faible et non significative (r = -0,36 ; p = 0,14). Notons que si l'on se limite aux valeurs observées (non z-scorées, donc OBS_SPARS_GLOB), on voit que toute corrélation a disparu (r = + 0,06). C'est dire que notre hypothèse initiale n'a pas été confirmée.

Question 2 : Les extensions d'une construction (= hapax) sont-elles sémantiquement plus denses (= plus cohérentes) que le reste du spectre?

Pour répondre à cette question, il faut confronter les densités du voisinage des hapax à celles des non hapax pour chaque construction. Nous calculons donc pour chaque hapax et pour chaque non-hapax de la construction la distance par rapport à son plus proche voisin (= NND ou *near-neighbor distance*) et prenons les moyennes par construction. Nous ferons de même pour les voisinages plus larges, c'est-à-dire les deux et trois voisins les plus proches. Ainsi, on voit par exemple que la distance moyenne des hapax par rapport à leur plus proche voisin varie de 0,21 à 0,38 (= ligne noire continue), abstraction faite d'une valeur aberrante, *semelle* (voir ci-dessous). Il va de soi que la sparsité des environnements augmente légèrement avec le nombre de voisins impliqués dans le calcul.

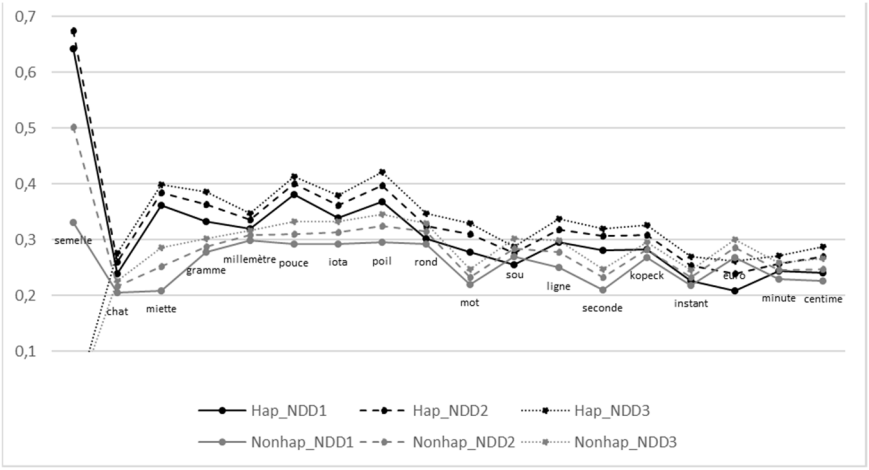


Figure 13. Sparsité des environnements hapaxiques vs non-hapaxiques

La Figure 13 permet de faire deux constats. D'abord, contrairement aux attentes, les hapax ne sont pas mieux entourés que les non-hapax ; on voit plutôt une tendance inverse. Voici les moyennes, sans *semelle*, valeur aberrante : NND1 = 0,29 vs 0,25 ; NND2 = 0,32 vs 0,27 ; NND = 0,33 vs 0,29. Cette différence est significative (R-base, Wilcoxon rank sum exact test, p = 0,03, 0,02 et 0,02 ; distribution non normale). Puis, comme nous avons placé les minimiseurs en ordre croissant de ratio *TypeToken*, on constate que l'écart se réduit presque à zéro pour les plus productifs (à droite), même s'il augmente à nouveau lorsqu'on prend en considération un voisinage plus large.

On peut donc conclure que quand les constructions s'étendent, elles ne le font pas de manière plus cohérente par rapport au reste de leur spectre de types. En d'autres termes, la similarité sémantique n'est pas plus importante pour les hapax, vecteurs d'innovation/d'extensibilité, que pour les autres types, plus fréquents. Bien au contraire, avec les minimiseurs les moins productifs, les hapax sont, en moyenne, même plus éloignés de leur voisinage que les types plus fréquents. Pour les plus productifs, les hapax se comportent plus ou moins comme les non-hapax (dans un spectre globalement plus dense). Notons déjà que le comportement divergent des constructions productives et moins productives sera confirmé ci-dessous.
La valeur aberrante de *semelle* est intéressante :

> En 50 ans, la Doc Martens ne s'est pas démodée d'une semelle. Sans cesse relookée et customisée, la chaussure mythique écoulée à plus de 100 millions de paires dans le monde entier (Fr10[10])

Cet hapax, *démoder*, qui est le seul hapax de *pas d'une semelle*, est très éloigné des deux autres types, archi-fréquents, du minimiseur, *quitter* et *lâcher*. Il s'ensuit un indice de sparsité hapaxique très élevé. Il est clair qu'il ne s'agit pas d'une extension naturelle de la construction [*pas* V *d'une semelle*], qui demande des prédicats de mouvement compatibles avec l'idée de 'la plus petite distance possible'. Nous avons ici un bel exemple de la façon dont la macro-construction [V *pas* + $N_{\text{petite valeur sur une échelle}}$], qui chapeaute l'ensemble

des minimiseurs à un niveau de schématicité (= d'abstraction) plus haut, peut être exploitée de manière ludique, en interaction avec la situation d'énonciation. Le minimiseur choisi ici par le locuteur correspond à une petite valeur pertinente, taillée à la mesure de cette situation très particulière, où il est question d'une marque de chaussures. On pourrait y voir un cas de coercition par la macro-construction, qui force *semelle* dans un emploi que seul son rapport méronymique avec le référent central de la situation (une chaussure) permet de motiver. Le même mécanisme de création – ludique, extravagant – de minimiseurs 'sur le tas', informés par la situation énonciative, se retrouve assez souvent avec des minimiseurs hapaxiques, qui n'apparaissent donc qu'une seule fois dans le French Ten Ten Corpus 2017 :

> Avant de passer à table, les paladins en chef **ne souriaient pas d'une moustache**
>
> les musiques électroniques peuvent être domestiquées et ne **pas vieillir d'un sillon**. D'ailleurs, ici, le disque en plastique […] est bien rayé par le temps mais tourne aussi rond.

Question 3 : les constructions qui ont moins de types doivent-elles nécessairement s'*étendre* de manière *plus cohérente* que les constructions avec beaucoup de types ?

Jusqu'ici, nous avons vu que productivité (réalisée) et diversité sémantique ne sont pas nécessairement liées. Qui plus est, s'il y a un lien, nos résultats militent plutôt pour une faible corrélation inverse, qui n'est cependant pas significative. Ce fut la réponse à la première question. La deuxième question faisait un zoom sur les zones du spectre lexical associées à l'innovation lexicale, matérialisée par des hapax. Rappelons à ce propos que la productivité est avant tout une affaire d'applications potentielles, que le locuteur n'a pas encore lues ou entendues (y compris des types néologiques). C'est ce que Baayen (2009) appelle la productivité dite « potentielle » d'une construction, ce qui correspond à son extensibilité (Barðdal 2008). Nos résultats suggèrent que ces zones ne sont pas globalement plus denses, relativisant le rôle de la proximité sémantique dans l'extension (= Question 2), et que l'extension des constructions les moins productives en termes de fréquence-type semble, qui plus est, se réaliser encore d'une manière moins cohérente. Il reste maintenant à vérifier ce dernier constat, à savoir si ces environnements hapaxiques globalement moins denses, ces « loci d'extensibilité », sont tributaires de différences entre les constructions en termes de fréquence-type. Rappelons que le modèle théorique de Barðdal (2008) prédit que les constructions comptant peu de types, si elles s'étendent, devraient le faire de manière plus cohérente, tandis que les constructions qui se combinent avec beaucoup de types différents, n'auraient pas besoin d'obéir à cette condition. Nous nous attendions donc à trouver des environnements hapaxiques plus denses dans le premier cas et plus clairsemés dans le second. Pour vérifier cette hypothèse, nous confrontons fréquence-type (ou ratio *TypeToken*) et sparsité hapaxique standardisée (Z_SPARS_ENV_HAPAX, ici *z_NND1*), c'est-à-dire la distance moyenne entre les hapax et leur voisin le plus proche ; nous y ajoutons aussi le ratio *HapaxToken*, ainsi que les valeurs de sparsité sémantique brutes (OBS_SPARS_ENV_HAPAX, ici *obs_NND1*):

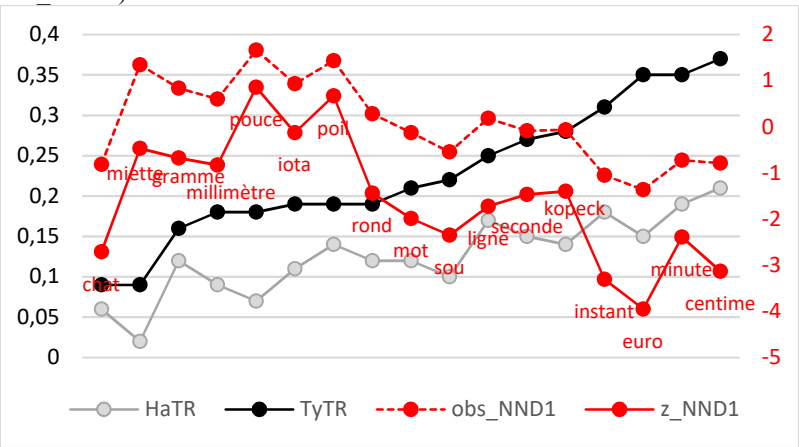
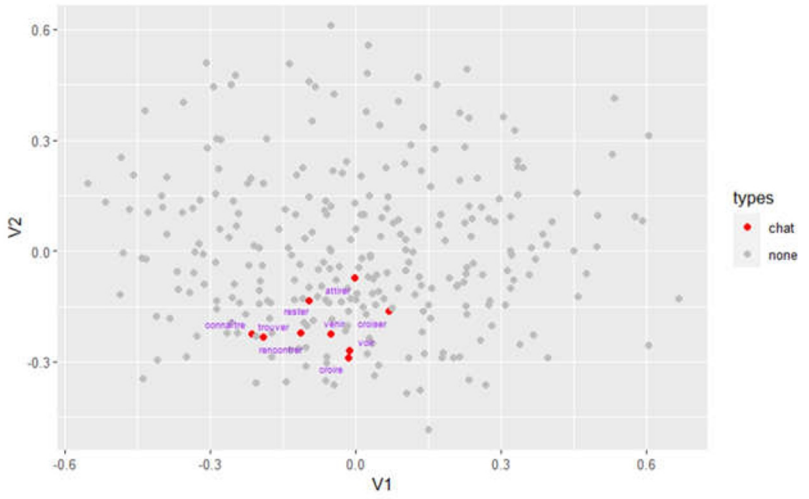


Figure 14. Productivité (*TypeToken, HapaxToken*) vs sparsité hapaxique (par rapport au plus proche voisin)

Ce graphique montre que la sparsité hapaxique présente une corrélation modérée avec le ratio *TypeToken*, mesure emblématique de l'ouverture lexicale (r = -0,59; p = 0,01) (Base R, *cor.test*). Cependant, de manière inattendue, cette corrélation est une corrélation inverse: les constructions qui ont le moins de types s'étendent de manière moins cohérente que les constructions les plus productives en termes de fréquence-type (ratio *TypeToken*). En d'autres mots, les constructions moins lexicalement diversifiées n'ont nullement besoin d'être plus cohérentes (que les constructions à fréquence-type élevé) pour s'étendre, alors que les constructions les plus diversifiées lexicalement s'avèrent justement encore plus cohérentes, malgré leur grand nombre de types (et d'hapax). Ce constat surprenant confirme l'impression signalée ci-dessus lorsque nous comparions la sparsité des environnements hapaxiques et non-hapaxiques.

On note cependant une exception: *chat*, qui, lui, est très dense, à la fois du point de vue de sa sparsité globale (OBS_SPARS_GLOB: 0,34) et lorsqu'il s'étend à de nouveaux prédicats hapaxiques (OBS_SPARS_ENV_HAPAX: 0,24). Cela ressort aussi de la visualisation moyennant le positionnement multidimensionnel (angl. *multidimensional scaling* ; Kruskal 1964), qui projette les distances entre les types dans un espace bidimensionnel, en préservant au maximum les distances réelles (R packages: MASS [Venables and Ripley 2002] et GGPLOT2 [Wickham 2016]). On voit ainsi que les prédicats de *chat* (en rouge) s'entassent dans un petit secteur du nuage des 285 prédicats attestés dans l'ensemble des 124 constructions de minimisation :



4.3 Discussion

Nos résultats montrent premièrement que la diversité lexicale (productivité réalisée) et la diversité sémantique ne sont pas nécessairement corrélées. Si on admet que la sparsité sémantique est un indicateur de la schématicité (sémantique) d'une construction (Perek 2020 : 144), on constate que productivité réalisée (fréquence-type) et schématicité ne doivent pas nécessairement être liées, ce qui rejoint Perek (2020), même pas au niveau *du slot* de la construction. Schématicité et productivité sont deux dimensions différentes, qu'il importe donc de distinguer, même si elles convergent souvent (comme avec *chat* ici).

Deuxièmement, les voisinages sémantiques des hapax (calculés sur la base de 1 à 3 voisins) ne sont pas plus denses que ceux des types plus fréquents. Globalement parlant, on ne peut donc pas dire que les hapax constituent des extensions prudentes qui s'éloignent à peine de ce qui existe déjà, comme des oisillons qui osent à peine s'éloigner de leur nid. Nos données ne permettent donc pas de prouver le rôle de la similarité sémantique (Suttle & Goldberg 2011) – et son interaction avec le nombre de types 'locaux', c'est-à-dire la couverture ou densité sémantique 'locale' (*coverage*, Suttle & Goldberg 2011 ; Perek 2016) –, qui aboutirait à une extension « item-based », de proche en proche, de la construction. Toutefois, nos résultats n'impliquent pas pour autant que la similarité sémantique n'ait plus de pertinence du tout. Seulement, la

similarité sémantique ne s'avère pas plus importante pour les poches d'innovation qu'ailleurs. En effet, il n'en reste pas moins que, globalement parlant, hapax et non hapax sont bien entourés : les densités moyennes des environnements de 1 à 3 voisins varient de 0,32 à 0,36 sur l'ensemble des minimiseurs. Ces valeurs sont bien en-deçà de la moyenne des indices de sparsité globale (0,50).

Troisièmement, et de manière plus surprenante, les constructions moins productives - présentant moins de variété lexicale réalisée - ont tendance à s'étendre de manière plus éparse que leurs homologues plus productives, ce qui contredit le modèle de Barðdal (2008). Faute d'espace, nous ne pourrons pas élaborer davantage la question ici (l'on se reportera à Lauwers & Van den Heede 2023), mais l'explication la plus plausible réside dans la nature variable des minimiseurs: les plus productifs (c'est-à-dire les plus lexicalement diversifiés), qui sont les minimiseurs qui renvoient à des valeurs monétaires (*centime, euro, kopeck*, …) et à des intervalles temporels (*minute, instant, seconde*) (cf. Figure 2), continuent à fonctionner comme des éléments sémantiquement transparents qui imposent des restrictions sémantiques à leurs prédicats, et en particulier à leurs nouvelles recrues. Ceux-ci doivent encore être sémantiquement compatibles avec le sémantisme d'origine du nom minimiseur, ce qui réduit l'empan sémantique des nouveaux types. En revanche, les minimiseurs davantage désémantisés et sémantiquement opaques (*poil, once, iota,* …) semblent être moins sensibles à l'impératif de cohérence et de similarité sémantiques depuis qu'ils ont quitté leurs niches sémantiques d'origine. A l'inverse, ces minimiseurs affranchis sur le plan sémantique s'avèrent davantage soumis à une logique de conventionnalisation collocationnelle, ce qui freine leur diversité (donc productivité) lexicale. On peut ainsi comparer *euro* et *iota* (entre parenthèses le nombre d'occurrences dans l'échantillon) :

- euro : coûter (11), mettre (11), dépenser (7), perdre (7), verser (7), prêter (4), posséder (4), débourser (4), …

- iota : changer (29), bouger (28), avancer (11), varier (9), …

Euro est compatible avec des prédicats liés à la manipulation d'argent (lui conférant le sens global de 'valeur monétaire nulle'), alors que les prédicats de *iota* ont un profil sémantiquement moins homogène (verbes de mouvement, de progression, de changement), pointant vers un sens plus abstrait de 'quantité nulle, absence donc, de changement'. Ainsi y trouve-t-on un plus grand nombre de types a priori moins attendus :

la décision de VirusTotal n'a pas affecté Cylance d'un iota (Fr10[10])

Le vent n'a pas faibli d'un iota (Fr10[10]),

alors que les rares types (a priori) moins attendus d'*euro*, rentrent tous dans le rang dès qu'on examine leur contexte d'emploi: *remplir* (*remplir les caisses d'un euro*); *servir* (*pas un euro sert à financer*), *varier* (*la dépense n'a pas varié d'un euro*).

Toutefois, il serait faux de croire que les minimiseurs transparents sont pleinement productifs dans leur niche sémantique. En effet, même les minimiseurs les plus transparents ne sont pas exempts de tendances collocationnelles, ce qui explique les ratios *TypeToken* inférieurs à 0,40. En effet, a priori, rien ne permet de prédire pourquoi *pas un instant* se combinerait surtout avec des verbes marquant l'hésitation (*douter* (27), *hésiter* (14)) face aux milliers de verbes théoriquement possibles. De plus, sur le plan sémantique, même les plus transparents se prêtent à des extensions inattendues (p.ex. *rater, envisager* + *pas un milimètre*), ce qui suggère qu'ils commencent à déjouer déjà leurs restrictions sémantiques d'origine.

Conclusions et perspectives

Arrivé à la fin de ce panorama, il est clair que le dernier mot n'a pas encore été dit. Tout d'abord, d'autres études, sur d'autres types de constructions syntaxiques (y compris en diachronie), sont nécessaires pour mieux comprendre les rapports entre les différentes mesures de la productivité et les dimensions de la problématique qu'elles véhiculent. Sur un plan plus technique, il faudrait également déterminer quelles mesures sont les moins dépendantes de la taille de l'échantillon (Pankratz et al. 2022 ; Feltgen s.p.p.). Bref, si la productivité fait partie de la connaissance (implicite) que les locuteurs ont de leur langue (Zeldes 2012 :

préface), elle reste particulièrement difficile à saisir, notamment en syntaxe. Il en est de même des rapports entre productivité et ouverture sémantique/schématicité (Perek 2020), où les hypothèses avancées se contredisent parfois.

Afin de pouvoir dénouer tous ces nœuds, il faudra aussi partir à la recherche de preuves indépendantes. En effet, n'oublions pas que l'usage attesté dans les corpus, à l'échelle de la communauté, n'est qu'une des manifestations possibles de cette boîte noire qu'est la productivité. Il reste à valider notamment les conclusions relatives à l'extensibilité des constructions et au rôle de la sémantique par le biais d'autres types d'observables du comportement langagier, qui dépassent le domaine de l'attesté, tout en ramenant la problématique à l'échelle de l'individu. A ce propos, des expériences d'acceptabilité et d'oculométrie, voire même d'EEG pourraient permettre de cerner de plus près le traitement cognitif différencié de certains contrastes observés dans les corpus (différents profils de types, constructions à productivité variable[10], etc.), tout comme les différences individuelles entre locuteurs.

Si l'approche basée sur corpus a ses limites, il en est de même de la sémantique distributionnelle. Tout d'abord, il convient de rappeler que, si elle suit d'assez près les intuitions des locuteurs en matière de similarité sémantique, la méthode ne traduit qu'une certaine vision de la sémantique, d'autant plus que la méthode 'omnibus' choisie ici, le paradigme « Bag-of-Words », est fortement attachée à la dimension référentielle, voire thématique. Il se pourrait donc qu'elle capte mal certaines similarités ou restrictions sémantiques plus abstraites. En outre, la mesure de sparsité sémantique que nous avons exploitée, basée sur des moyennes, reste assez grossière. Elle traduit aussi une vision plutôt 'continuiste' de la diversité sémantique, qui fait abstraction de la structuration interne du spectre des types, et notamment de la présence de classes sémantiques, ou, de manière plus générale, de micro-clusters sémantiques locaux, qui peuvent être plus cohérents que la moyenne. Toutefois, elle a le double avantage d'être indépendante des aléas d'un quelconque algorithme de clustering et de permettre de comparer un grand nombre de constructions sur une base égale. Enfin, notre méthode se heurte au problème de la polysémie, qui est le principal obstacle des méthodes distributionnelles (Heylen et al. 2015). Pour certains types hautement polysémiques, les coordonnées sémantiques obtenues pourraient s'avérer peu informatives et mal assorties avec l'acception activée dans la construction à l'étude. A ce propos, les approches basées sur l'apprentissage profond, et notamment les modèles dits "de transformer" comme BERT (Devlin et al. 2019) et ses avatars français Camembert (Le et al. 2020) et Flaubert (Martin et al. 2020), pourraient offrir une issue.

Références bibliographiques

Aurnague, M. & Plenat, M. (1997), Manifestations morphologiques de la relation d'attachement habituel. *Silexicales, 1*. *Mots possibles et mots existants*, Corbin, D. Fradin, B., Habert, B. Kerleroux, F. & Plénat, M. éds. Lille : Université de Lille 3 : 15-24.

Baayen, R. H. (1992), Quantitative aspects of morphological productivity. *Yearbook of Morphology 1991*, Booij, G. E. & van Marle, J. éds. Dordrecht: Kluwer Academic Publishers : 109-149.

Baayen, R.H. (1993), On frequency, transparency, and productivity. *Yearbook of Morphology 1991*, Booij, G. E. & van Marle, J. éds. Dordrecht: Kluwer Academic Publishers : 181-208.

Baayen, R. H. (2009), Corpus linguistics in morphology: morphological productivity. *Corpus Linguistics. An international handbook*, Lüdeling, A.& Kyto, M. éds, Berlin: De Gruyter : 900-919.

Baayen, R. H., et Renouf, A. (1996), Chronicling The Times: Productive Lexical Innovations in an English Newspaper. *Language*, 72 : 69-96.

Barðdal, J. (2008), *Productivity: Evidence from Case and Argument Structure in Icelandic.* Amsterdam: Benjamins.

Barðdal, J., Enghels, R., Feltgen, Q., Van Hulle, S. & Lauwers, P. (sous presse), Productivity in Diachrony. *Wiley Blackwell Companion to Diachronic Linguistics*, Ledgeway A., Breitbarth A., Kiss, K., Salmons J. & Simonenko A. éds. Hoboken, New Jersey : Wiley-Blackwell.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018), Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software, 3(30),* 774.

Bolinger, D. (1972), *Degree words*. The Hague: Mouton.

Booj, G. (2010), Construction morphology. *Language and Linguistics Compass, 3(1)* : 1-13.

Bybee, J. (1995). Regular morphology and the lexicon, *Language and Cognitive Processes*, *10(5)*: 425-455.

Bybee, J. (2006). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Bybee, J. & Eddington, D. (2006), A usage-based approach to Spanish verbs of 'becoming'. *Language, 82(2)* : 323-355.

Chomsky, N. (1965), *Aspects of the theory of syntax*. Cambridge, MA : MIT Press.

Corbin, D. (1987), *Morphologie dérivationnelle et structuration du lexique* (2 vols). Tubingen : Max Niemeyer Verlag.

Dal, G. (2003), Productivité morphologique : définitions et notions connexes. *Langue française, 140* : 3-23.

Desagulier, G. (2016), A lesson from associative learning: Asymmetry and productivity in multiple-slot constructions. *Corpus linguistics and linguistic theory, 12(2) :* 173-219.

Devlin, J., Chang M.W., Lee K. & Toutanova K. (2019), BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. Minneapolis, Minnesota : Association for Computational Linguistics : 4171-4186.

Erk, K. (2012), Vector space models of word meaning and phrase meaning: a survey. *Language and Linguistics Compass 6(10)* : 635-653.

Evert, S. (2014), Distributional Semantics in R with the wordspace Package. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, Dublin : 110-114.

Feltgen, Q. (2017), *Statistical physics of language evolution: the grammaticalization phenomenon.* Paris: Université Paris sciences et lettres dissertation.

Feltgen, Q. (s.p.p.), Testing diachronic measures of productivity using Zipf-Mandelbrot law. *Qualico* (12th International Quantitative Linguistics Conference).

Firth, J.R. (1957), *A synopsis of linguistic theory 1930-1955. Studies in Linguistic Analysis*. Oxford : Philological Society : 1-32.

Gaeta, L. & Ricca, D. (2006), Productivity in Italian word formation: a variable-corpus approach. *Linguistics, 44(1)* : 57-89.

Goldberg, A. (1995), *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago : University of Chicago Press.

Goldberg, A. (2019). Explain me this. *Creativity, competition, and the partial productivity of constructions*. Princeton: Princeton University Press.

Grabar, N., Tribout D. , Dal G. , Fradin B. , Hathout N., Lignon S., Namer F. , Plancq C. , Yvon F., Zweigenbaum P. (2006). Productivité quantitative des suffixations par *-ité* et *-Able* dans un corpus journalistique moderne. *Verbum ex machina, Actes de la 13e conférence sur le traitement automatique des langues naturelles*, Mertens P., Fairon, C., Dister, A. & Watrin P. éds. Louvain-la Neuve : Presses universitaires de Louvain : 167-177.

Gries, S. (2010). Useful statistics for corpus linguistics, *A mosaic of corpus linguistics: selected approaches*, Sánchez A. & Almela M. éds. Frankfurt am Main: Peter Lang : 269-291.

Gries, S. (2015), More (old and new) misunderstandings of collostructional analysis: On Schmid and Küchenhoff (2013). *Cognitive Linguistics, 26 (3)* : 505-536.

Heylen, K., Wielfaert, T., Speelman, D. & Geeraerts, D. (2015), Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua, 157* : 153-172.

Hoeksema, J. (2002), Minimaliseerders in het Standaardnederlands. *Tabu, 32* : 105-174.

Jackendoff, R., & Audring, J. (2019), *The texture of the lexicon. Relational Morphology and the Parallel Architecture.* Oxford : Oxford University Press.

Joubarne, C. & Inkpen, D.. 2011, Comparison of Semantic Similarity for Different Languages Using the Google n-gram Corpus and Second-Order Co-occurrence Measures. *Advances in Artificial Intelligence. Canadian AI 2011. Lecture Notes in Computer Science, 6657,* Butz, C. & Lingras, P. éds. Berlin/Heidelberg : Springer : 216-221.

Jurafsky D. & Martin J.H. (2024), Chapter 6 : Vector semantics and embeddings. *Speech and Language Processing* (3e éd.). Publication en ligne : https://web.stanford.edu/~jurafsky/slp3/6.pdf.

Kassambara, A. & Mundt, F. (2020), *Factoextra: Extract and visualize the results of Multivariate Data Analyses*. R package version 1.0.7.

Kassambara, A. (2017). *Practical guide to principal component methods in R: PCA, M(CA), FAMD, MFA, HCPC, factoextra*. STHDA.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29(1)* : 1-27.

Langacker, R. (1987), *Foundations of Cognitive Grammar. Volume I: Theoretical Prerequisites*. Stanford: Stanford University Press.

Lauwers, P. (2014), From lexicalization to constructional generalizations. On complex prepositions in French. In: *Romance Construction Grammar*, Gonzálvez-García, F. & Boas, H. éds. Amsterdam etc.: Benjamins : 79-111.

Lauwers, P., & Van den Heede, M. (en préparation), Productivity and semantic diversity within a family of minimizing constructions in French.

Le H., Vial L., Frej J., Segonne, V., Coavoux M., Lecouteux, B., Allauzen A., Crabbe B., Besacier L. & Schwab, D. (2020). FlauBERT: Unsupervised language model pre-training for French ». Marseille : LREC.

Lê, S., Josse, J., & Husson, F. (2008), FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software, 25(1)* : 1-18.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017), Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language,* 92 : *57*-78.

Martin L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary L., de la Clergerie E., Seddah D. & Sagot B. (2020), CamemBERT: a tasty French language model. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics : 7203-7219.

Nyrop, K. (1908), *Grammaire historique de la langue française, tome 3*. Copenhague : Gyldendalske Boghandel Nordisk Forlag ; Leipzig/New York/Paris : Harrassowitz, Stechert & Picard.

Perek F. (2020), Productivity and Schematicity in constructional change. *Nodes and Networks in Diachronic Construction Grammar*, Sommerer, L. & Smirnova, E. éds, Amsterdam/Philadelphia : Benjamins : 141-166.

Perek, F. (2016), Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics,* 54 : 149-188.

Pinker, S. (1999), *Words and Rules: The Ingredients of Language*. London: Phoenix.

Plevoets, K. (2015), *svs: Tools for Semantic Vector Spaces*. Gent: Ghent University.

Pankratz, E., von der Malsburg, T. & Vasishth, S. (2022), Shannon entropy is a more comprehensive and principled morphological productivity measure than the standard alternatives. *PsyArXiv, 8*, juin 2022 (en ligne).

Suttle, L., & Goldberg, A. (2011), The partial productivity of constructions as induction. *Linguistics, 49(6),* 1237-1269.

Trousdale, G. (2008), A constructional approach to lexicalization processes in the history of English: evidence from possessive constructions. *Word Structure,* 1 : 156-177.

Van den Heede, M. (en préparation), *The minimizing construction in Dutch and French*, Thèse de doctorat. Université de Gand.

Van den Heede, M., & Lauwers, P. (2023), Syntactic productivity under the microscope: the lexical and semantic openness of Dutch minimizing constructions. *Folia Linguistica, 57(3)* : 723-761.

Van Hulle, S., Enghels, R. & Lauwers, P. (soumis pour publication), The many guises of productivity: a case-study of Spanish inchoative constructions.

Van Wettere, N. (2018), *Copularité et Productivité : une analyse contrastive des verbes attributifs issus de verbes de mouvement en français et en néerlandais*. Thèse de doctorat. Université de Gand.

Van Wettere, N. (2021), Productivity of French and Dutch (semi-)copular constructions and the adverse impact of high token frequency. *International Journal of Corpus linguistics, 26* : 396-428.

Van Wettere, N. (2022), The hapax/type ratio. An indicator of minimally required sample size in productivity studies? *International Journal of Corpus Linguistics, 27(2)* : 166-190.

Venables, W. N. & Ripley, B.D. (2002), *Modern applied statistics with S. Fourth edition*. New York: Springer

Wickham, H. (2016), *ggplot2: Elegant graphics for data analysis*. New York: Springer.

Zeldes, A. (2012), *Productivity in Argument Selection: From Morphology to Syntax*. Berlin: De Gruyter.

Corpus utilisés

COW corpus:

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. *Proceedings of challenges in the management of large corpora* (CMLC-3), IDS publication server : 28-34.

Schäfer, R. & Bildhauer, F. (2012), Building large corpora from the web using a new efficient tool chain. *Proceedings of the eighth international conference on language resources and evaluation* (LREC'12) : 486-493.

TenTen corpus:

Jakubíček, M., Kilgarriff A., Kovář V., Rychlý, P. & Suchomel, V. (2013). The TenTen corpus family. *7th International Corpus Linguistics Conference CL*. Lancaster : 125-127.

New, B. *Lexique* (http://www.lexique.org/)

---

[1] Je tiens à remercier ici Ludovic De Cuypere pour avoir exploré avec moi le prétraitement des corpus COW en vue de la création de matrices distancielles. Un grand merci aussi à Kris Heylen, qui m'a initié à la sémantique distributionnelle, ainsi qu'à Koen Plevoets pour m'avoir orienté dans l'univers des *packages* R.

[2] Consulté le 18 mars 2024.

[3] Comme le nombre d'occurrences est stable dans notre étude de cas, fréquence-type et ratio *TypeToken* coïncident.

[4] Voici la formule (Gries 2010 : 272), avec $0.\log_2 0 = 0$: $H = \sum_{i=1}^{n}(p(x) \cdot \log_2 p(x))$

[5] En effet, plus on voit d'occurrences (en agrandissant le corpus), plus cela devient difficile de rencontrer de nouveaux types (et d'hapax). La construction 's'épuise' pour ainsi dire. Il convient donc d'évaluer la diversité lexicale indépendamment de la *fréquence-token*, qui doit être neutralisée. Ce problème méthodologique est particulièrement ardu en diachronie, étant donné les fluctuations de fréquence d'une construction au fil du temps (voir notre article collectif, Barðdal et al., à par.).

[6] On pourrait objecter que, puisqu'on fonctionne avec des échantillons stables (à 100 occurrences), qu'un petit nombre de types est tout simplement la conséquence mathématique du fait que les types les plus fréquents ont absorbé la plupart des occurrences de l'échantillon. Certes, le top 3 donne déjà le ton et contribue donc à la corrélation négative, mais il n'en reste pas moins que le reste du spectre fréquentiel peut encore varier beaucoup. A priori, un sommet très élevé de la pyramide réduit l'espace-type restant mais ne *détermine* pas le nombre de types 'derrière' le top 3, et encore moins

le nombre d'hapax. Ce n'est qu'a posteriori qu'une telle corrélation entre la hauteur du top 3 et le nombre de types 'restants' peut être établie avec certitude.

[7] Notons que pour les prédicats complexes, nous avons retenu l'élément sémantique central (p.ex. *avoir en caisse -> caisse*). Pour le présentatif *il y a*, très fréquent dans notre jeu de données (p.ex. *il n'y a pas un chat à la fac en ce moment*), nous avons choisi *trouver* comme substitut. Il s'ensuit une liste de 287 prédicats, dont deux seulement ne figuraient pas dans la matrice distancielle de 50 000 lemmes.

[8] Voir Lauwers & Van den Heede (en prép.) pour des détails plus techniques et d'autres applications de la méthode.

[9] *Pas un clou* ne compte qu'un seul type ; il n'a donc pas d'indice de sparsité.

[10] En ce moment, de telles expériences sont en cours dans le cadre du projet *Language productivity@work* (https://www.languageproductivity.ugent.be/) à l'Université de Gand.