



**HAL**  
open science

# Nouns of humans, transparency and opacity: A corpus-based study of compound nouns in English

Elise Mignot, Daniel Henkel

## ► To cite this version:

Elise Mignot, Daniel Henkel. Nouns of humans, transparency and opacity: A corpus-based study of compound nouns in English. *Nommer l'humain Descriptions, catégorisations, enjeux*, L'Harmattan, 2021, Langages de discours en débats. hal-03783961

**HAL Id: hal-03783961**

**<https://hal.science/hal-03783961>**

Submitted on 26 Sep 2022

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

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

# Noms d'humains, transparence et opacité : étude sur corpus du cas des composés en anglais

Élise MIGNOT

*Sorbonne Université*

*EA 7332 CeLiSo (Centre de Linguistique en Sorbonne)*

Daniel HENKEL

*Université Paris 8 Vincennes-St. Denis*

*TransCrit EA1569*

## Résumé

L'étude porte sur les noms d'animés humains en anglais, et, plus précisément, sur les noms *composés* qui dénotent des humains, tels que *milkman*, *yes-man*, *plastic lady*. Dans un corpus d'un million de mots, étiqueté par catégories grammaticales (POS) et lemmes grâce au logiciel *TreeTagger*, un premier recensement par expressions régulières (regex) visant tous les syntagmes nominaux a donné lieu ensuite à un relevé manuel des différents schémas correspondant à la composition nominale. Les humains paraissent moins susceptibles d'être dénotés par des composés que les autres entités, ce que nous attribuons à la relative transparence des noms composés, et à la façon dont nous conceptualisons les humains, une personne étant vue comme un tout qui transcende la somme de ses parties. Les noms d'animés humains tendent à être opaques car une dénomination trop descriptive réduit une personne à un seul aspect, ce qui revient à trahir sa complexité.

*Mots clefs : noms composés, humains, transparence, opacité, corpus*

## Abstract

This study focuses on nouns denoting human beings in English, and more specifically on compound nouns such as *milkman*, *yes-man* or *plastic-lady*. In a 1-million-word corpus tagged for POS and Lemma with *TreeTagger*, regular expressions were used to produce an exhaustive inventory of all Noun Phrases, followed by a manual classification of compound nouns. Compound nouns were found to be used less frequently to designate human beings, in comparison with other semantic categories, which can be explained by the greater transparency of compound nouns, and the way in which we conceptualize human beings, i.e. as a whole transcending the sum of their parts. Nouns for human beings tend to be opaque insofar as an overly specific denomination reduces a person's intrinsic complexity to a single aspect.

*Mots clefs : compound nouns, human, transparency, opacity, corpus*

## 1. Introduction

Nous nous penchons sur les noms d'animés humains en anglais contemporain, et plus particulièrement sur les noms *composés* dénotant des animés humains, comme par exemple *milkman* (livreur de lait), *yes man* (un homme qui dit « oui » à tout et est donc perçu comme lâche), *girlfriend* (petite amie).

Par « noms d'animés humains » nous entendons les noms d'individus, et non pas les groupes d'humains tels que *gang, government, family*, etc. En effet, ceux-ci sont déjà, au moins en partie, du côté du non-humain, car il s'agit de construits conceptuels. Même si les membres du groupe sont des humains, le groupe lui-même peut être vu comme une entité autre, or nous cherchons véritablement à cerner les spécificités de la dénotation de l'humain, au sens de « personne ».

Par « nom composé », nous entendons : nom constitué de deux bases, dont (le plus souvent), l'une est une tête (Bauer 2009 : 343 et 348, Bauer et Huddleston 2002 : 1630-1631, Katamba 1994 : 264, Plag 2003 : 5). Nous avons pris en compte tous les types de composés, c'est-à-dire tous les types morphologiques, accentuels, et sémantiques.

Parmi les types morphologiques on rencontre les schémas suivants. (La liste est adaptée de Bauer 1983 : 202 sq et 2009 : 342-356).

- Les composés nom + nom, comme *milkman*
- Les composés verbe + nom, comme *playpit*
- Les composés adjectif + nom, comme *blackbird*
- Les composés préposition (ou adverbe) + nom, comme *in-crowd*
- Les composés adverbe + nom, schéma peu productif car il ne fonctionne qu'avec des adverbes de temps ou de lieu (*now generation*)
- Les « dephrasal compounds », correspondant à des figements et lexicalisations de constructions syntaxiques, comme *lady-in waiting, son-in-law*.

Nous incluons dans notre corpus les composés illustrant les deux schémas accentuels, accent initial (*milkman*) ou accent final (*silk scarf, cotton dress*). Nous suivons Bauer (1983 : 104) pour qui l'accent n'est pas un critère pour décider si une séquence peut être un composé ou pas, particulièrement dans le cas des noms composés (par opposition au adjectifs composés, par exemple).

Nous incluons également tous les types sémantiques (Bauer 1983 : 30-31, Bauer 2009 : 353, Bauer, Lieber et Plag 2013: 463-483, Dressler 2006 : 23-44), c'est-à-dire, notamment, les composés endocentriques (dont la tête est le deuxième élément – c'est le cas de *milkman*, un type de *man*), les exocentriques (dont la tête est extérieure à la structure car ils dénotent une entité autre que celle dénotée par chacun des deux noms – c'est le cas pour *birdbrain*, qui ne dénote pas un type de cerveau mais un type de personne) (Bauer 2008), et ceux qui n'ont pas de tête évidente (les composés coordinatifs tels que *murder-suicide* ou *singer-songwriter*, où les deux noms semblent être sur le même plan).

Nous retenons également tous les types sémantiques au sens où nous incluons tous les composés quelle que soit la relation sémantique qui unit le deux termes. Ces relations sémantiques peuvent être très variées (Bauer 2009 : 353, Downing 1977, Hatcher 1960, Levi 1978, Warren 1978).

Dans la littérature sur les noms composés, il ne semble pas y avoir de traitement différencié sur le critère humain - non-humain. Ainsi la différence n'est pas mentionnée dans les ouvrages consacrés à la composition de Lieber et Štekauer (2009) ou Scalise et Vogel (2010) par exemple.

Nous nous penchons sur les *composés* (dans la perspective de la référence aux humains) car ils sont analysables et, par rapport à d'autres procédés de formation des mots, (relativement) transparents. Ceci a une conséquence sur l'application des dénominations à certaines entités plutôt qu'à d'autres.

Les composés sont analysables dans la mesure où on reconnaît différentes parties de la dénomination : un composé est constitué de deux bases reconnaissables, qui ont, dans la plupart des cas, une existence autonome dans le lexique. Il s'agit là d'une caractéristique formelle.

Ils sont par ailleurs (relativement) transparents. Il s'agit cette fois d'une caractéristique sémantique : le sens du tout est (au moins en partie) déductible du sens des parties. Ceci est à mettre en relation avec les notions d' « arbitraire absolu » et « arbitraire relatif » de Saussure :

Le principe fondamental de l'arbitraire du signe n'empêche pas de distinguer dans chaque langue ce qui est radicalement arbitraire, c'est-à-dire immotivé, **de ce qui ne l'est que relativement**. Une partie seulement des signes est absolument arbitraire ; chez d'autres intervient un phénomène qui permet de reconnaître des degrés dans l'arbitraire sans le supprimer : **le signe peut être relativement motivé**. (Saussure 1966 : 180-181 [Ns. soul.])

Le « relativement motivé » correspond au cas où on comprend la raison d'être de la dénomination, car on reconnaît dans un mot construit au moins une base qui appartient au lexique. Les noms simples sont du côté de l'arbitraire absolu, et les noms composés du côté de l'arbitraire relatif, donc d'une certaine motivation.

S'il est vrai qu'un composé tel que *blackbird* est dans une certaine mesure opaque, à l'intérieur du domaine de la formation des mots, les composés sont plus transparents que d'autres formations. En effet nous reconnaissons des bases. Dans les noms suffixés, tels que *baker*, nous reconnaissons certes *-er*, mais ce n'est pas une base, et son sens n'est pas très précis. En effet, *-er* dénote le plus souvent un agent, mais ce n'est pas toujours le cas. Le nom *breaker*, par exemple, peut dénoter une vague qui se brise (dans le contexte du surf), et le *-er* n'a donc rien d'agentif. Par ailleurs, ce suffixe peut dénoter des animés (*baker*) ou des inanimés (*dishwasher*). Une dernière différence entre les bases et les suffixes est que les bases sont autonomes ; elles appartiennent à la syntaxe. La syntaxe, contrairement au lexique, est transparente, dans la mesure où le sens du tout est déductible du sens des parties (Malmberg, 1977 : 383-411). Il ressort donc que, même si le lexique, de façon générale, est caractérisé par l'opacité, nous pouvons poser des degrés d'opacité. Les noms simples sont bien sûr les plus opaques, et les composés, étant constitués de plusieurs bases, sont les moins opaques.

Or, nous entendons montrer ici que les humains s'accommodent mal de dénominations transparentes. Nous observons qu'ils servent peu à dire les humains. Il semble exister une asymétrie dans le lexique, selon qu'on nomme un animé humain ou une autre entité.

Précisons enfin que cette étude succède à une autre (Mignot et Marty 2018), sur la même question, conduite à partir du *Concise Oxford Dictionary* (Stevenson & Waite : 2011). Le but de la présente étude est donc de confronter ces premières conclusions à une étude portant sur l'usage.

Ce premier inventaire compte 7196 noms (de tous types, simples ou construits), dont 1028, c'est-à-dire 14%, dénotent des humains. A titre de comparaison, si l'on retient seulement les noms *composés* du corpus, 7%, c'est à dire la moitié de la proportion, dénotent des humains. Les humains sont donc moins susceptibles que d'autres entités d'être dénotés par des noms composés.

Une autre façon de présenter les données est de se pencher sur les seuls noms d'humains. Les noms composés représentent 17,6% des noms d'humains. A titre de comparaison, 32,3% des noms d'inanimés sont des composés. On peut encore une fois conclure que les humains sont moins susceptibles d'être dénotés par les composés que d'autres entités.

Cette première étude, conduite sur les mots tirés du dictionnaire, a en outre permis de montrer que les noms composés qui, en dépit de la tendance, dénotent des humains, présentent les particularités sémantiques suivantes (qui ne sont pas mutuellement exclusives).

- Ils ont tendance à être dévalorisants. C'est le cas par exemple de *pack rat*, *faintheart*, *fair-weather friend*, *plastic lady* (alors que *plastic bag*, qui dénote un inanimé, n'est pas dévalorisant).
- Ils impliquent une relation interpersonnelle. La personne dénotée par *paper boy*, par exemple, livre le journal chez quelqu'un).

Les noms composés d'entités autres que les animés humains ne présentent pas ces caractéristiques, et il semble, encore une fois, que nous traitons les humains différemment des autres entités.

Cette première étude, menée à partir d'un échantillon de mots du dictionnaire, demandait à être complétée par une étude sur corpus, afin de vérifier si les résultats concordent.

## 2. Méthodes

Les résultats dont nous faisons état à présent proviennent d'un corpus d'un peu plus d'un million de mots contenant 10 œuvres de 10 auteurs différents, d'environ 100 000 mots chacun, sélectionnées de manière aléatoire parmi une cinquantaine d'œuvres du domaine public<sup>1</sup> en provenance du Projet Gutenberg et datant de 1850 à 1925 environ.

**Tableau 1.** *Caractéristiques du corpus utilisé*

<u>Auteur</u>	<u>Titre</u>	<u>Pays/Date</u>	<u>Nombre de mots</u>
Connor, R.	<i>The Major</i>	(Canada, 1919)	120937 mots
Forster, E.M.	<i>Howards End</i>	(U.K., 1910)	112375 mots
Galsworthy, J.	<i>Beyond</i>	(U.K., 1917)	119642 mots
Glaspell, S.	<i>The Visioning</i>	(U.S.A., 1911)	118206 mots
Howells, W.D.	<i>The Landlord at Lions Head</i>	(U.S.A., 1897)	121412 mots
Lawrence, D.H.	<i>Aaron's Rod</i>	(U.K., 1922)	116811 mots
Parker, G.	<i>The Right of Way</i>	(Canada, 1901)	123391 mots
Phillips, D.G.	<i>The Second Generation</i>	(U.S.A., 1906)	116149 mots
Reade, C.	<i>White Lies</i>	(U.K., 1857)	121719 mots
Wright, H.B.	<i>The Eyes of the World</i>	(U.S.A., 1914)	116130 mots

*année médiane : 1910 total : 1 186 772 mots*

Tous les textes ont été nettoyés (suppression du métatexte), et étiquetés par catégories grammaticales et lemmes dans *TreeTagger*<sup>2</sup>, afin de procéder à l'extraction automatique des SN au moyen de la formule regex suivante :

- $(?b\S+/(CD|QD|DT|PP\$)\S+){0,2}((?b\S+/RB[RS]?S+)*?b\S+/J[JV][RSGN]?S+((\S+/CC\S+)(/,/,))?)?(?b\S+/N[NP]S?S+)*?b\S+/N[NP]S?S+$

à savoir :

1 Il s'agit d'un sous-corpus tiré de celui décrit dans Henkel (2014). Quoique le choix d'étudier des textes du domaine public s'explique d'abord par des raisons pratiques, il présente également l'avantage de rendre nos résultats vérifiables et reproductibles pour d'autres chercheurs. D'autre part, pour que le corpus ne soit pas trop éloigné de l'anglais contemporain du point de vue diachronique, la date de 1850 a été fixée comme limite chronologique antérieure.

2 L'étiquetage par défaut a subi deux modifications : d'une part, les quantificateurs MANY, MUCH, MORE, ENOUGH et quelques autres, étiquetés « JJ » (« adjectif ») par *TreeTagger*, ont été ré-étiquetés en « QD » (c.-à-d. « quantifieur-déterminant ») ; d'autre part, étant donné que seules certaines formes participiales en fonction épithète reçoivent l'étiquette « JJ », alors qu'aucun consensus n'existe quant au statut grammatical des participes en fonction épithète, tous les autres participes (« VVG », « VVN ») se trouvant entre un déterminant et un nom (c.-à-d. en fonction épithète, à l'intérieur d'un SN) ont été ré-étiquetés en « JVG », « JVN », de sorte que tous les participes en position épithète puissent être assimilés aux adjectifs (« J- ») lors de notre recensement. Lorsqu'un nom était précédé d'un participe, sans déterminant matériel, p.ex. *O broken promises*, le participe n'a pas été ré-étiqueté, (pour ne pas inclure parmi les épithètes des résultats comme *he had broken promises all his life*) mais le nom a néanmoins été recensé, en tant que SN simple, et corrigé, le cas échéant, lors de l'examen manuel des échantillons.

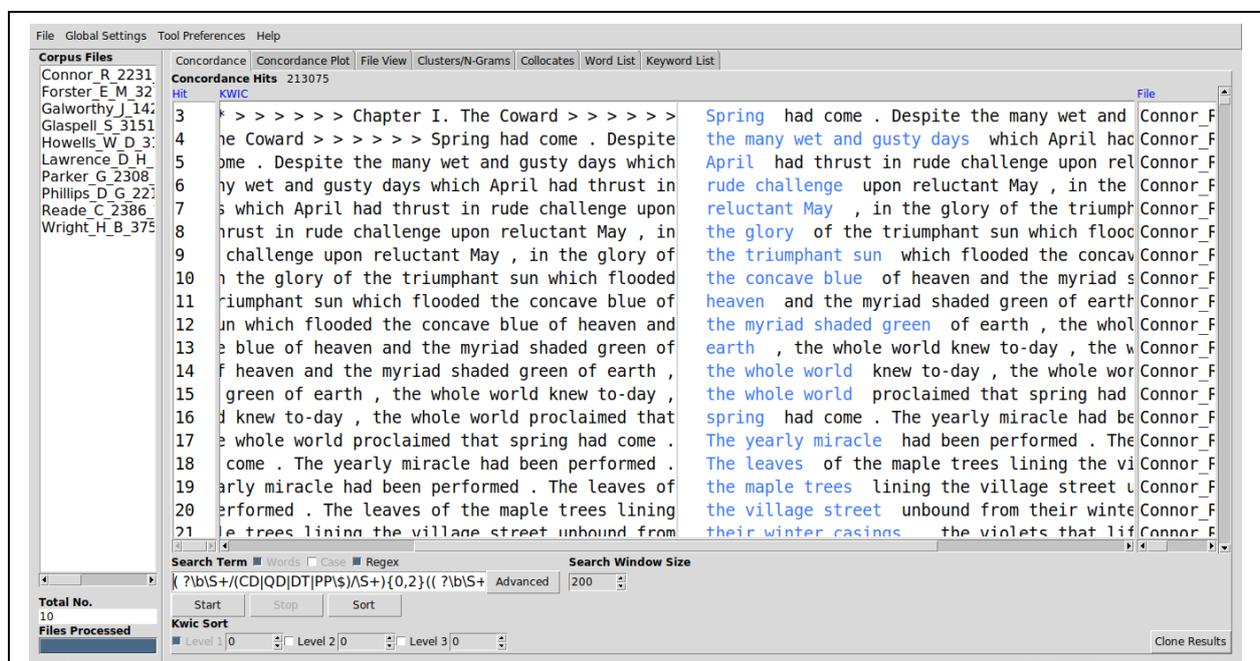
- un Nom commun (NN, NNS) ou propre (NP)

ayant éventuellement, à sa gauche, des :

- déterminants (DT, PPS)
- quantifieurs (CD, QD)
- adjectifs (JJ, JJR, JJS), participes (JVG, JVN) ou noms (NN, NNS, NP) en fonction épithète

Comme le montre la capture d'écran du concordancier *AntConc* ci-dessous (Fig. 1), cette méthode de recensement a ainsi permis d'inventorier les SN du corpus de manière quasi exhaustive<sup>3</sup>, quelle que soit leur structure interne, y compris :

- des SN simples tels que *Spring* (1.3), *April* (1.5), *heaven* (1.11), *earth* (1.13)
- des SN complexes contenant plusieurs modes de détermination ou quantification, ainsi que plusieurs épithètes, tels que *the many wet and gusty days* (1.4)
- des SN contenant des noms en fonction épithète, tels que *the maple trees* (1.19), *the village street* (1.20)



**Fig. 1** Résultats de la recherche par formule regex dans *AntConc* (capture d'écran)

Au total, 222 673<sup>4</sup> SN ont été recensés, soit environ 20 000 par texte (chaque texte ayant fait également l'objet d'un recensement individuel : max. Parker  $n=25\,467$ , min. Glaspell  $n=19\,894$ ).

À ce stade de notre enquête, cependant, est survenue une difficulté due au fait qu'il n'existe en anglais aucune convention graphique permettant de repérer les noms composés de manière automatique et exhaustive. On trouve, en effet, des noms composés :

3 Nous disons « quasi » car la possibilité d'une erreur d'étiquetage ne peut jamais être totalement exclue. Lors de l'examen manuel de nos échantillons, le taux d'erreur observé était d'environ 2-3 %.

4 La différence entre ce résultat et celui qui s'affiche dans la capture d'écran d'*AntConc* tient au fait que notre inventaire a porté directement sur les textes, convertis en tableaux .csv, sans passer par le concordancier, lequel omet parfois quelques résultats selon les options choisies (p.ex. celles qui touchent à la prise en compte des espaces ou aux caractères non alphanumériques). La fonction regex d'*AntConc* donne néanmoins des résultats suffisamment proches pour illustrer notre méthode de recensement.

- s'écrivant en 1 seul mot, tels que *chairman* ou *landlady*,
- s'écrivant avec un trait d'union, tels que *farm-hand* ou *sleep-walker*,
- ou bien sans aucun indice graphique, comme *business man* (qui s'écrit aussi en 1 seul mot) ou *shop girls*.

Par conséquent, compte tenu de l'impossibilité<sup>5</sup> d'obtenir un recensement exhaustif des SN contenant des noms composés, nous avons été obligés de procéder à un examen manuel portant sur des échantillons de 2000 SN prélevés de manière aléatoire dans chacun des textes du corpus (soit environ 10%). Ont été retenus comme noms composés :

- toutes les séquences N+N (voire N+N+N, etc.)
- toutes les unités graphiques, avec ou sans trait d'union, formées à partir d'une base nominale (p.ex. *gentleman*, *playfellow*, *sister-in-law*)
- les noms de discours (c.-à-d. ayant reçu des marques morphosyntaxiques nominales comme le pluriel) composés de plusieurs éléments (p.ex. *the offscourings*, *those go easys*).

Comme dans l'étude qui a servi de point de départ à celle-ci, nous n'avons relevé que les noms d'humains au sens strict, et non pas les noms de groupes d'humains (tels que *club*, *gang*, *gouvernement*, *family*).

### 3. Résultats

Seront exposées ci-après cinq séries de résultats :

- d'abord les noms les plus fréquents en anglais globalement
- la proportion des noms d'êtres humains globalement, en termes de types lexicaux et d'occurrences
- l'inventaire des noms composés recensés, par type lexical
- la proportion des noms d'êtres humains parmi les noms composés
- les Ratios de Tokens par Type observés dans nos échantillons

#### 3.1. Les noms les plus fréquents

Dans le Tableau 3 ci-dessous figurent les 20 lexèmes nominaux les plus fréquents dans le discours narratif globalement, tel que représenté dans notre corpus :

**Tableau 2a.** 20 Lexèmes nominaux les plus fréquents en anglais (discours narratif). Nombre total d'occurrences, fréquence moyenne par 1000 mots.

<u>LEXÈME</u>	<u>n=</u>	<u>Fréq./1k</u>
1. man	3858	3.25
2. thing	2448	2.06
3. eye	2058	1.73
4. time	1917	1.62

<sup>5</sup> Les composés avec un trait d'union sont les plus pratiques à inventorier, en raison du signe typographique. Les suites N1 N2 sont aussi relativement faciles à cibler grâce à l'étiquetage grammatical, mais avec un taux non négligeable de faux-positifs. Quant aux composés qui s'écrivent en un seul mot, on peut, bien sûr, anticiper les cas les plus fréquents et les mieux connus, en ciblant le nom qui sert de base (p.ex. *man* dans *chairman*, *showman* etc.), mais non sans introduire dans les résultats un biais dû un *a priori* théorique. Seul un examen manuel permet, d'une part, d'éliminer des faux-positifs tels que *Roman*, *shaman* ou *cayman* (que l'on trouve en cherchant la base *man*) et de repérer en même temps des composés aussi imprévisibles que *car demillers*, chapelgoer ou *go easys*.

5. face	1841	1.55
6. hand	1809	1.52
7. day	1801	1.52
8. life	1748	1.47
9. way	1574	1.33
<b>10. woman</b>	1476	1.24
<b>11. girl</b>	1463	1.23
12. something	1174	0.99
13. word	1157	0.97
<b>14. mother</b>	1144	0.96
15. world	1111	0.94
<b>16. people</b>	1111	0.94
17. house	1063	0.9
18. nothing	1026	0.86
19. heart	1018	0.86
<b>20. father</b>	980	0.83

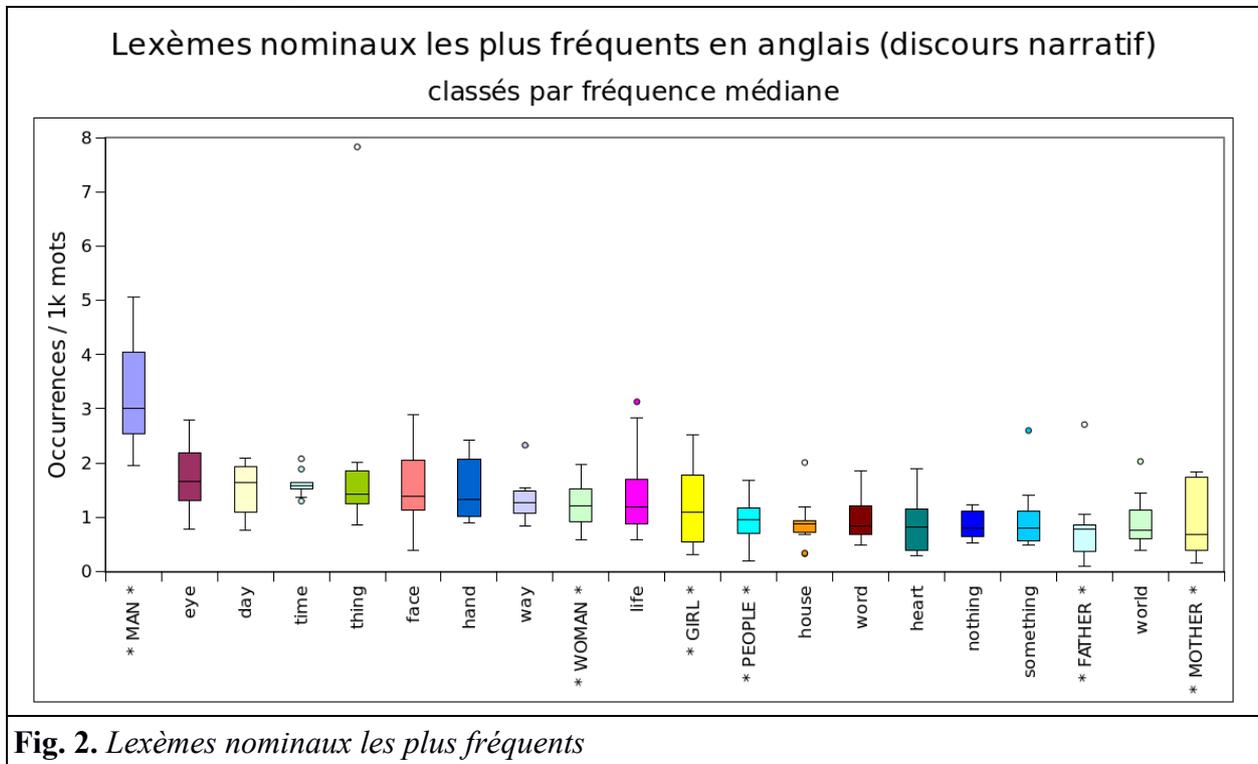
Parmi ceux-ci, on relève six noms désignant des êtres humains : *man*, *woman*, *girl*, *mother*, *people*, *father*. Cependant, les fréquences moyennes calculées pour l'ensemble du corpus cachent d'importantes disparités d'un auteur à un autre, comme le montre le Tableau 2b, dans lequel ont été ajoutées les fréquences maximale, minimale et médiane observées entre les différents auteurs.

**Tableau 2b** 20 Lexèmes nominaux les plus fréquents en anglais (discours narratif). Nombre total d'occurrences, fréquence moyenne par 1000 mots, fréquence maximale, médiane et minimale par 1000 mots.

<u>LEXÈME</u>	<u>n=</u>	<u>Fréq./1k</u>	<u>Fréq.Max./1k</u>	<u>Fréq.Méd./1k</u>	<u>Fréq.Min./1k</u>
<b>1. man</b>	3858	3.25	5.06	3	1.95
2. thing	2448	2.06	<b>7.83</b>	<b>1.42</b>	0.86
3. eye	2058	1.73	2.8	1.66	0.78
4. time	1917	1.62	2.08	1.58	1.3
5. face	1841	1.55	2.89	<b>1.4</b>	<b>0.4</b>
6. hand	1809	1.52	2.43	1.34	0.9
7. day	1801	1.52	2.1	1.64	0.76
8. life	1748	1.47	<b>3.13</b>	<b>1.19</b>	0.59
9. way	1574	1.33	2.33	1.27	0.84
<b>10. woman</b>	1476	1.24	1.98	1.21	0.59
<b>11. girl</b>	1463	1.23	<b>2.53</b>	<b>1.09</b>	<b>0.31</b>
12. something	1174	0.99	<b>2.6</b>	<b>0.8</b>	0.48

13. word	1157	0.97	1.86	0.85	0.48
<b>14. mother</b>	1144	0.96	<b>1.84</b>	<b>0.69</b>	<b>0.15</b>
15. world	1111	0.94	<b>2.03</b>	<b>0.76</b>	0.4
<b>16. people</b>	1111	0.94	1.68	<b>0.96</b>	<b>0.2</b>
17. house	1063	0.9	2.01	<b>0.88</b>	<b>0.33</b>
18. nothing	1026	0.86	1.23	0.81	0.53
19. heart	1018	0.86	1.89	<b>0.83</b>	<b>0.29</b>
<b>20. father</b>	980	0.83	<b>2.71</b>	<b>0.79</b>	<b>0.1</b>

Pour cette raison, nous avons préféré évaluer la fréquence des lexèmes nominaux, non pas en termes de fréquence moyenne, mais en tenant compte de la variation à l'intérieur d'une *fourchette* de fréquences observées d'un texte à l'autre, ce qu'on peut représenter sous forme de boîtes de Tukey (ou « boîtes à moustaches ») :



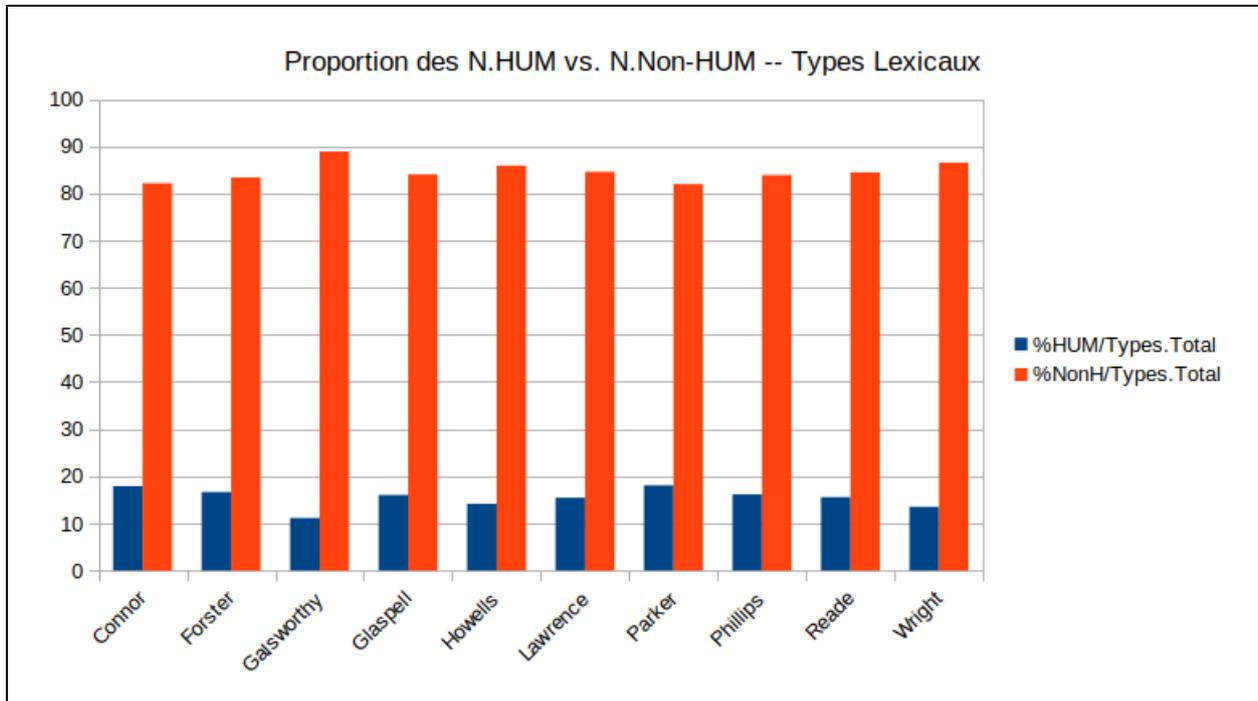
Pour chaque lexème nominal, on voit ainsi, entre les bornes supérieures et inférieures, la totalité de la variation observée<sup>6</sup>, à l'intérieur de laquelle les 50 % du milieu, c'est-à-dire ceux qu'on peut considérer comme étant les plus représentatifs de la tendance centrale, se trouvent dans la partie colorée. Plus la boîte est petite, moins il y a de variation en termes de fréquence entre les auteurs (inversement, une ampleur plus étendue indique, non pas une fréquence plus grande, mais plus de variation). Cette représentation graphique nous montre ainsi que ces noms très fréquents forment un continuum très progressif, où les fréquences s'inscrivent plus ou moins dans la même fourchette, – à

<sup>6</sup> Sauf si une valeur s'écarte de plus d'une fois et demie par rapport à l'écart interquartile, auquel cas celle-ci est considérée exceptionnelle et se trouve isolée, en dehors de la fourchette.

une exception près : « *man* », dont la médiane de 3 occ./1000 mots et même l'ensemble des fréquences observées se situent très nettement au-dessus de celle de tous les autres noms.

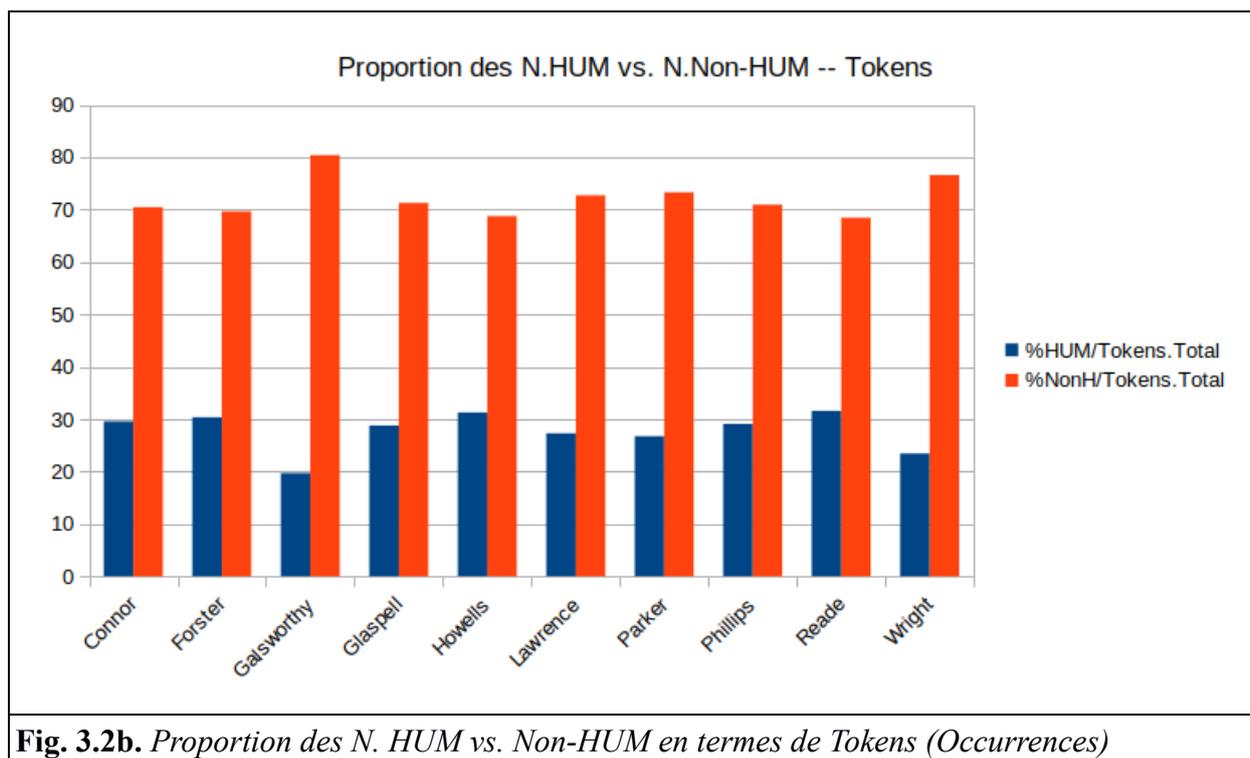
### 3.2. Proportion des noms d'êtres humains

En termes de Types Lexicaux, il ressort de nos observations que les noms dénotant des êtres humains représentent entre 11 et 18 % des types lexicaux, tandis que 80-90 % du vocabulaire nominal provient des autres classes sémantiques :



**Fig. 3.2a.** *Proportion des N. HUM vs. Non-HUM en termes de Types Lexicaux*

En revanche, en termes d'occurrences (ou de tokens) les noms d'êtres humains occupent une proportion du discours nettement plus élevée, entre 20 et 30% selon les auteurs :



**Fig. 3.2b.** Proportion des N. HUM vs. Non-HUM en termes de Tokens (Occurrences)

En clair, les noms d'êtres humains sont davantage *récurrents*, du moins dans le discours narratif. Ils ont tendance à être répétés, réutilisés, d'où une fréquence supérieure et une diversité lexicale inférieure aux autres types sémantiques. Ces proportions sont, rappelons-le, celles de tous les noms quel que soit leur type morphologique – simple ou composé – qui nous serviront ainsi de base de comparaison pour évaluer les proportions des noms désignant des êtres humains ou non-humains parmi les noms composés dans la section 3.4 ci-après.

### 3.3. Inventaire des noms composés par type lexical

Le Tableau 3 ci-dessous contient l'inventaire complet des noms composés recensés dans les échantillons aléatoires prélevés dans chaque œuvre du corpus :

**Tableau 3.** Noms composés désignant des êtres humains recensés dans 10 échantillons aléatoires de 2000 SN. Nombre de Tokens, nombre de Types Lexicaux (bases nominales)

	<u>Tokens</u>	<u>Types*</u>	
<b>Connor</b>	15	7	<i>the Yiddish colour bearer, Mr. Chair-r-rman, Mr. Chairman, A little Jew chap, an ordained clergyman, The car dwellers, a gentleman, the gentleman, business men, the big business men, a Trinity man, no mind-reader, the Scotchman, our genial YMCA Secretary, a young University student</i>
<b>Forster</b>	11	10	<i>the bride-elect, a clergyman, gentlefolk, Gentlemen, its inmates, the lay-reader, the offscourings, the soul offspring, a parlourmaid, So-and-so, the stepmother</i>

<b>Galsworthy</b>	8	5	<i>the ballet-master, an almost accomplished horsewoman, the page-boy, a page-boy, this madman, nice gentleman, business men, a chapelgoer</i>
<b>Glaspell</b>	9	7	<i>a daughter-in-law, a gentleman, chorus girls, shop girls, our regiment men, a staff officer, a playfellow, her sister-in-law, the trouble-maker</i>
<b>Howells</b>	15	11	<i>advance-agent, Boston architects, The coachman, the coachman, the coachman, the farm-hand, foot-passengers, The gentleman, every Boston girl, New York girls, his gran'father, their kinswomen, a trance medium, the night porter, a Harvard student</i>
<b>Lawrence</b>	13	8	<i>fellow-passengers, a footman, the modern, social freebooter, a Frenchman, gentlemen, a gentleman, mostly middle-class well-to-do Italians, the landlady, the landlady, the newcomer, his own ivory queen, a sleep-walker, some van-men</i>
<b>Parker</b>	23	10	<i>a child-woman, the stern Churchman, an Englishman, a gentleman, the gentleman, all gentlemen, gentleman, a jail- bird-robber-murderer, the medicine-man, an exclusive fellow-passenger, quack-doctor, cheerful raftsmen, every river-driver, coarse river-drivers, the river-men, the shepherd, your shepherd, the soldier-citizens, this tailor-man, the timber-merchant, our beautiful and admired townswoman, a waterman, a waterman</i>
<b>Phillips</b>	17	10	<i>the all-conqueror, a breadwinner, future brother-in-law, the day-laborer, go easys, the Frenchman, girl friends, the poor old gentleman, gentleman, young gentleman, her grandmothers, a wonderful, fine old nobleman, manufacturing partner, two-thousand-dollar Romney, a showman, a master workman, workman</i>
<b>Reade</b>	7	6	<i>the commander-in-chief, no fellow-smokers, this old gentleman, his heir-at-law, an artillery man, sister-in-law, a peasant woman</i>
<b>Wright</b>	11	7	<i>the Chinaman, the Chinaman, the Chinaman, his girl comrade, the shadowy horseman, housekeeper, long-ago lovers, our girl musician, the most popular portrait painter, culture-club women, the last workman</i>

Individuellement, chaque échantillon contenait entre 7 et 23 tokens, et entre 5 et 11 types lexicaux<sup>7</sup> différents. Il va de soi que chacun de ces échantillons est bien trop petit pour formuler la moindre conclusion, mais en opérant un classement global, en fonction de la base lexicale, comme ci-dessous (cf. Tableau 4), on ne peut qu'être frappé par la place qu'occupent les composés de *man*, tels que *clergyman*, *workman*, *business man* (en 1 ou 2 mots) etc., lesquels ne représentent pas moins de 40 % de tous les noms composés d'êtres humains :

**Tableau 4** *Types lexicaux observés parmi les noms composés HUM dans 10 échantillons aléatoires de 2000 SN.*

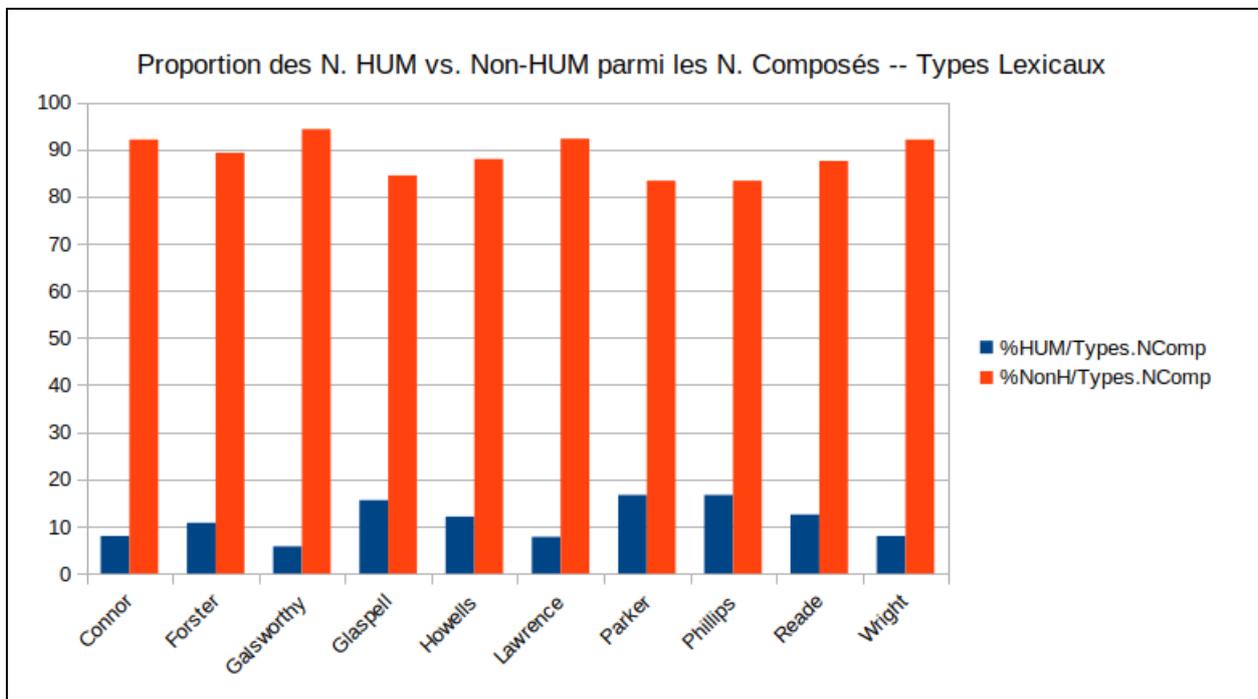
<sup>7</sup> Il convient de préciser que ces chiffres sont donnés en tenant compte du nom qui a servi de base à la composition, c.-à-d. *fellow* dans *playfellow* et *maid* dans *parlourmaid*, afin de calculer les ratios token/type dans les mêmes conditions que pour les noms simples, lesquels ont tous été comptabilisés de la même manière avec ou sans épithète.

Type Lexical	n=	% Tokens N.Comp.HUM
-man ( <i>river-man, showman, workman, nobleman, etc.</i> )	52	40.31 %
-woman ( <i>peasant woman, kinswomen, townswoman, etc.</i> )	6	4.65 %
-boy, -girl, -lady, -father, -mother ( <i>chorus girls, landlady, page-boy, etc.</i> )	11	8.53 %
-in-law ( <i>daughter-in-law, etc.</i> )	5	3.88 %
-er ( <i>river-driver, housekeeper, sleep-walker, etc.</i> )	24	18.6 %
rôle ( <i>parlourmaid, timber-merchant, University student, etc.</i> )	11	8.53 %
autres	20	15.5 %
<b>Total Tokens : 129</b>		

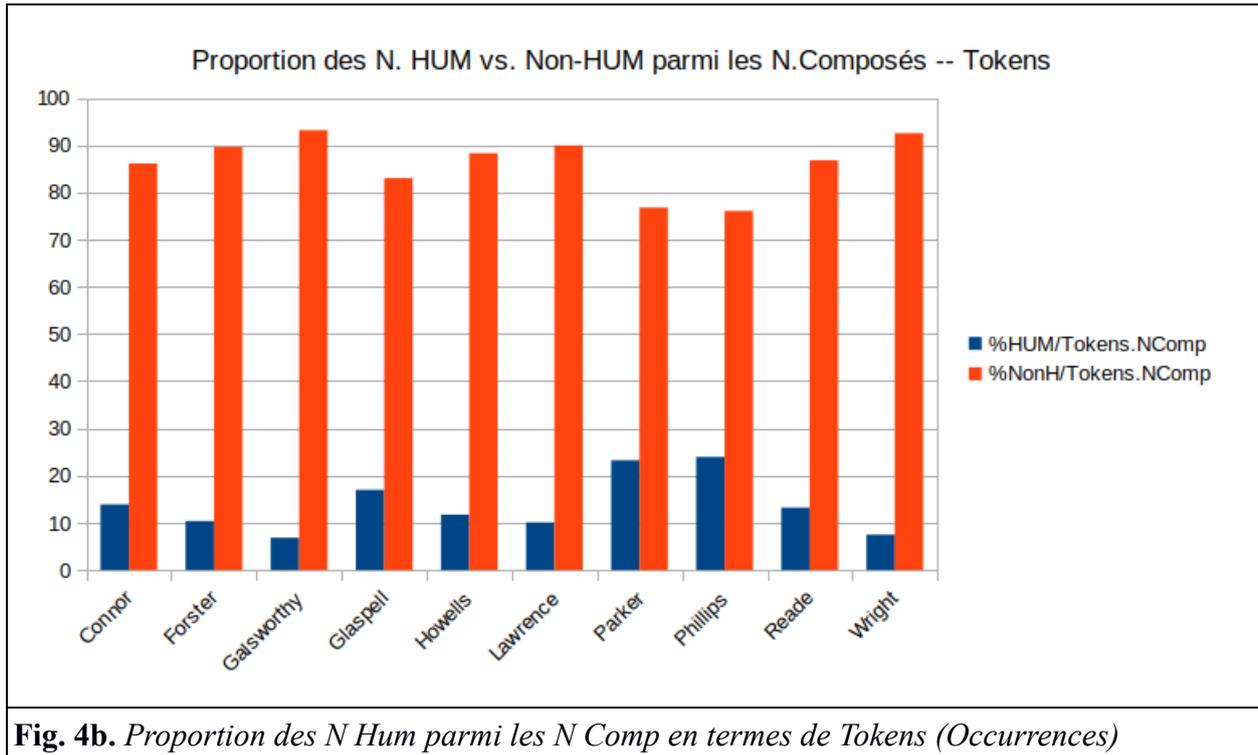
Après *man*, ce sont les dérivés en *-er* comme *housekeeper* et *sleepwalker* qui constituent la deuxième classe morpho-sémantique la plus nombreuse, auxquels on peut associer les noms de rôle comme *timber-merchant* et *parlourmaid* qui arrivent en troisième position (*ex aequo* avec *-boy, -girl, -lady*). Autrement dit, les noms composés d'êtres humains servent le plus souvent – pas uniquement, mais le plus souvent – à caractériser les êtres humains en fonction d'une activité, notamment socio-professionnelle.

### 3.4. Proportion des noms d'êtres humains parmi les noms composés

Sont représentées ci-dessous (Fig. 4a-b) les proportions des noms désignant des êtres humains, cette fois-ci parmi les noms identifiés morphologiquement comme des noms composés dans nos échantillons, d'abord en termes de Types Lexicaux et ensuite en termes d'occurrences :



**Fig. 4a.** Proportion des N Hum parmi les N Comp en termes de Types Lexicaux

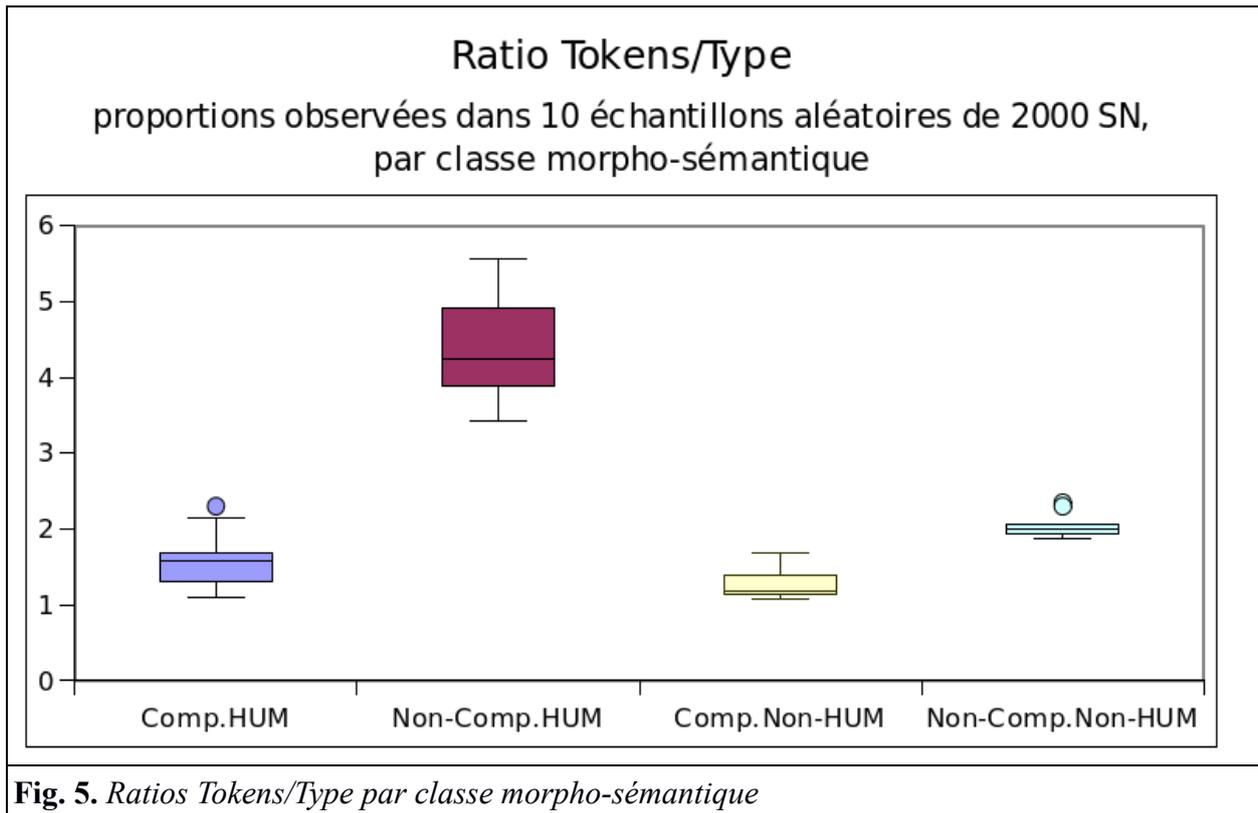


**Fig. 4b.** Proportion des *N Hum* parmi les *N Comp* en termes de *Tokens* (Occurrences)

Ces résultats sont à rapprocher de ceux donnés en 3.2 pour l'ensemble des noms, tous types morphologiques confondus. En termes de types lexicaux, les noms d'êtres humains sont moins bien représentés, avec une médiane de 10,5 % (au lieu de 15,6 % pour l'ensemble des types morphologiques), tandis qu'en termes d'occurrences la médiane est de 11,7 % (contre 29 % précédemment).

### 3.5. Ratios Tokens/Type

Une dernière analyse permettra d'apprécier d'une autre manière à quel point les noms composés d'êtres humains sont atypiques par rapport à l'ensemble de leur classe sémantique. On voit ci-dessous (Fig. 5) les ratios de token par type, c'est-à-dire le nombre de fois, en moyenne, qu'un type lexical donné a été retrouvé dans nos échantillons :



Dans l'ensemble, ce rapport est d'environ 1 à 2 tokens par type – sauf les noms d'êtres humains, autres que composés, qui sont répétés en moyenne entre 4 et 5 fois par échantillon. Les noms composés, en revanche, s'emploient de manière ponctuelle, avec 1 à 2 tokens par type, comme les noms d'objets ou d'entités non-humains.

#### 4. Discussion

Les données tirées de l'usage confirment donc celles de l'étude sur dictionnaire. La catégorie ontologique (humain ou non humain) du référent semble jouer un rôle dans le type de nom sélectionné. Nous interpréterons maintenant cette moindre proportion de noms composés dans la référence aux humains. Nous posons qu'elle tient à la façon dont nous conceptualisons l'humain.

Les humains sont perçus comme plus complexes que les autres entités : ils sont mobiles, ils participent à plusieurs procès, parfois plusieurs en même temps, leur apparence change avec le temps. Pourtant, un être humain doit être vu comme un tout. Les humains sont plus que la somme de leurs attributs, de leurs actes, ou de leurs parties. Ce sont des personnes, qui ont une identité. Ces points sont développés en philosophie (Chauvier (2003 : 10) entre autres) et a de nombreuses implications, par exemple légales, au travers du concept de « responsabilité » (Chauvier 2003 : 10).

La notion de « personne » est certes complexe : on peut se demander si une personne reste la même si elle perd la mémoire, ou si Dr Jekyll et Mr Hyde sont une seule et même personne (Chauvier 2003 : 52). Mais il s'agit-là de cas particuliers, et le fait même qu'ils soient saillants indique qu'ils ne correspondent pas à la conceptualisation par défaut de ce qu'est une personne.

Dans la tradition philosophique, il existe deux façons principales de définir une personne : l'approche réductionniste et l'approche non-réductionniste. Nous adoptons la seconde. L'approche réductionniste, développée par Parfit (1984 : 346), après Hume (2000 [1748]), propose que l'unité d'une vie n'implique rien de plus que les diverses relations entre les expériences qui ont eu lieu

pendant cette vie. En d'autres termes, l'existence d'une personne ne consiste qu'en une série interconnectée d'événements mentaux et physiques (Parfit 1984 : 211).

Nous n'adoptons pas cette position car, comme le note Campbell (1994 : 178), elle correspond à un point de vue de troisième personne : afin de voir une personne comme une série d'états mentaux, il nous faudrait être en position d'observateur. Or nous ne nous percevons pas nous-mêmes comme le ferait un observateur extérieur. Nous nous pensons à la première personne. Nous suivons donc Chauvier (2003 : 117) qui estime que nous ne pouvons pas nous concevoir en termes réductionnistes, et que nous ne sommes pas une association transitoire d'états mentaux et cérébraux. Cet auteur estime que le concept même de « personne » inclut de façon inhérente un point de vue de première personne (Chauvier 2003 : 117). En d'autres termes, les personnes ont un moi. Nous adoptons donc l'approche non-réductionniste, c'est-à-dire holistique, qui est aussi plus proche de la façon dont nous nous percevons : nous sommes plus que la somme de nos parties, qu'il s'agisse de parties du corps, d'états psychologiques, etc.

Nous posons en outre que ceci a une conséquence sur la façon dont nous nommons les humains. Nous observons que les dénominations d'humains ont tendance à être plus opaques que les dénominations d'autres entités, et proposons que cette exigence d'opacité tient au fait qu'il est difficile de réduire les humains à une de leur partie. Les dénominations d'humains doivent être opaques, car, si elles sont transparentes, elles sont descriptives, au moins dans une certaine mesure, et réduisent de ce fait la personne à un aspect. Les composés, comme on l'a vu, sont des dénominations plus transparentes que d'autres. Ils sont donc moins aptes à dénoter les humains.

## 5. Conclusion

Il ressort de ce travail que les données de l'étude en contexte corroborent celles de l'étude effectuée sur dictionnaire. Les humains sont moins susceptibles d'être dénotés par des composés, qui sont des formations relativement transparentes.

Nous supposons que la relative rareté des composés en référence aux animés humains est due à la façon dont on conceptualise les humains, c'est-à-dire comme plus que la somme de leurs parties. Réduire quelqu'un à une caractéristique revient à ignorer la complexité essentielle de l'humain.

Rappelons qu'à l'occasion de la première étude, sur dictionnaire, qui a servi de point de départ à celle-ci, nous avons observé que lorsque les humains sont dénotés par des composés, les dénominations tendent à être dévalorisantes et/ou à inclure une notion de relation interpersonnelle. Dans la présente étude, conduite à partir de l'usage, cette hypothèse reste en suspens, et devra donc être vérifiée ou précisée par des analyses ultérieures.

## Bibliographie

- Bauer, L. (1983). *English word formation*. Cambridge : Cambridge University Press.
- Bauer, L. (2008). *Les composés exocentriques de l'anglais*. In D. Amiot (Dir.), *La composition dans une perspective typologique* (pp. 35-47). Arras : Artois Presse Université.
- Bauer, L. (2009). *A typology of compounds*. In R. Lieber & P. Štekauer (Dirs.), *The Oxford handbook of compounding* (pp. 343- 356). Oxford : Oxford University Press.
- Bauer, L., & Huddleston R. (2002). *Lexical word formation*. In R. Huddleston & J.K. Pullum (Dirs.), *The Cambridge grammar of the English language* (pp. 1621-1722). Cambridge : Cambridge University Press.
- Bauer, L., Lieber R., & Plag I. (2013). *The Oxford reference guide to English morphology*. Oxford : Oxford University Press.
- Campbell, J. (1994). *Past, space and self*. Cambridge, Massachusetts : MIT Press.

- Chauvier, S. (2003). *Qu'est-ce qu'une personne?* Paris : Vrin.
- Downing, P. (1977). *On the creation and use of English compound nouns*. *Language*, 53, 810-842.
- Dressler, W. U. (2006). *Compound types*. In G. Libben & G. Jarema (Dirs.), *The representation and processing of compound words* (pp. 23-44). Oxford : Oxford University Press.
- Hatcher, A. G. (1960). *An introduction to the analysis of English noun compounds*. *Word*, 16, 356-373.
- Henkel, D. (2014). *L'adjectif en anglais et en français*. Paris : Thèse de doctorat, Université Paris-Sorbonne
- Hume, D. (2000) [1748]. *An inquiry concerning human understanding*. Oxford : Clarendon Press.
- Katamba, F. (1994). *English words*. London : Routledge.
- Levi, J. (1978). *The syntax and semantics of complex nominals*. New York: Academic Press.
- Lieber, R., & Štekauer, P. (Dirs.). (2009). *The Oxford handbook of compounding*. Oxford : Oxford University Press.
- Malmberg, B. (1977). *Signes et symboles. Les bases du langage humain*. Paris : Picard.
- Mignot E. & C. Marty (à paraître en 2018). « Denominations of humans : the case of compounds in English ». *Numéro spécial de International Journal of Language and Culture* (John Benjamins), L. Gardelle et S. Sorlin (Dirs.), *From culture to language and back: the animacy hierarchy in language and discourse* (pp. 162-181).
- Parfit, D. (1984). *Reasons and persons*. Oxford : Oxford University Press.
- Plag, I. (2003). *Word formation in English*. Cambridge : Cambridge University Press.
- Saussure, Ferdinand de. (1916 / 1967). *Cours de linguistique générale*. Paris : Payot.
- Scalise, S., & Vogel, I. (Dirs.). (2010). *Cross-disciplinary issues in compounding*. Amsterdam : John Benjamins.
- Stevenson, A., & Waite, M. (Dirs.). (2011). *Concise Oxford English dictionary. Twelfth edition*. Oxford : Oxford University Press.
- Warren, B. (1978). *Semantic patterns of noun-noun compounds*. Göteborg : Acta Universitatis Gothoburgensis.