# Le système des quantifieurs

## *Modèles théoriques et résultats empiriques*

**Michel DELARCHE (EILA/LANSAD)**

**Présentation au séminaire CLILLAC**
**(2 avril 2012)**

**Die Mathematiker sind eine Art Franzosen: Redet man zu ihnen, so übersetzen sie es in ihre Sprache, und dann ist es alsbald ganz etwas anders. (Goethe)**

***Les mathématiciens sont comme les Français: quoi qu'on leur dise, il le traduisent en leur propre langage, et cela devient aussitôt tout à fait autre chose.***

**Michel DELARCHE**
**Univ Paris Diderot, Sorbonne Paris Cité, CLILLAC-ARP**
**UFR EILA, case 7002, 75205 Paris cedex 13**
**delarche@eila.univ-paris-diderot.fr**

# Le système des quantifieurs
# Modèles théoriques et résultats empiriques

## La modélisation des quantifieurs

- Modèles sémantiques formels
- Universaux sémantiques
- Anneaux de quantifieurs: un modèle sémantique axiomatique

## Consolidation empirique à travers les co-occurrences

- Travailler aux limites des corpus existants
- Exploiter les moteurs de recherche

## La détection semi-automatique en analyse de corpus

- Mise en oeuvre de l'outil NooJ
- Gestion des ambigüités morpho-syntaxiques et sémantiques

## Contribution potentielle à l'étude de styles et de genres

- L'essai de vulgarisation scientifique
- Les romans de Jane Austen
- Les romans de Maupassant

université PARIS DIDEROT PARIS 7

# Première partie:
## *Modélisation formelle des quantifieurs*

## Sémantique formelle

## Théorie des universaux sémantiques

## Représentation systémique contrastive de la sémantique des quantifieurs

## Consolidation empirique

**Michel DELARCHE**
**Univ Paris Diderot, Sorbonne Paris Cité, CLILLAC-ARP**
**UFR EILA, case 7002, 75205 Paris cedex 13**
**delarche@eila.univ-paris-diderot.fr**

# La modélisation des quantifieurs
# Représentations formelles (1/9)

Logique + Philosophie Analytique + Linguistique + Informatique
→ General Quantifier Theory (GQT)

**La GQT veut exprimer rigoureusement la sémantique des quantifieurs naturels à travers des grammaires formelles, afin de décrire certaines propriétés des quantifieurs en vue de les catégoriser théoriquement**

**Quelques références:**


Barendregt HP *The Lambda Calculus*, North Holland, 1984
Meyer B *Introduction to the Theory of Programming Languages*, Prentice-Hall, 1990
Articles fondateurs:
Montague, R *The Proper Treatment of Quantification in Ordinary English*
J. Hintikka, J. Moravcsik, P. Suppes (eds.): Approaches to Natural Language.
Dordrecht 1973, 221–242.
Barwise, J, and Cooper, R. *Generalized quantifiers and natural language*
Linguistics and Philosophy 4(2): 159–219, 1981
Synthèses récentes:
Peters, S. and Westerstål, D. *Quantifiers in Language and Logic,* Oxford, 2007
Cann, R., Kempson, R. and Gregoromichelski, E.
*Semantics - An introduction to Meaning in Language,* Cambridge, 2009


université PARIS DIDEROT PARIS 7

# La modélisation des quantifieurs
# Représentations formelles (2/9)

**Logique propositionnelle: dans cette approche on combine les propositions (énoncés élémentaires ayant une valeur de vérité) par différents opérateurs logiques ($\neg$, $\wedge$, $\vee$, $\rightarrow$) ce qui permet de spécifier les règles de déduction à un niveau de granularité assez grossier, reliant des propositions dérivées (inférences) ou équivalentes (synonymes) en construisant des « tables de vérité » qui modélisent le rôle des connecteurs logiques (and, or, except/but, (n)either...(n)or, if...then...else...)**

**Calcul des prédicats: on travaille à un niveau de granularité plus fin (prédicats reliant des variables). Par exemple, une expression du genre "sujet-verbe-complément" devient une fonction v(s,c) appelée « prédicat à deux places »; un composé "adjectif-nom"ou "nom-copule-attribut" sont représentés par des prédicats à une seule place a (n) etc.**

**Le calcul de prédicats inclut les quantifieurs existentiel $\exists$ et universel $\forall$ pour modéliser la sémantique par des modèles ensemblistes (les variables parcourent différents domaines)**

université PARIS DIDEROT PARIS 7

# La modélisation des quantifieurs
# Représentations formelles (3/9)

**Théorie des types: on développe des modèles plus complexes à partir de la généralisation de la notion de fonction par le lambda-calcul qui représente abstraitement la liaison entre variables et fonctions à différents niveaux de récursivité (une « fonctionnelle » est une fonction généralisée dont les arguments sont eux-mêmes des fonctions)**

**Par exemple, un verbe ditransitif VDT est du type:**

$e \rightarrow (e \rightarrow (e \rightarrow t)$ **soit la formule:** $\lambda x\, \lambda y\, \lambda z$ **VDT (x, y, z)**

**Avec 'give' comme instanciation la formule se lit: « x gives y z »**

**NB: on peut parenthéser pour modifier l'ordre d'évaluation:**
**ainsi,** $\lambda x\, (\lambda y\, (\lambda z$ **Vdt (x, y, z))) peut s'interprèter: « z is given to y by x »**

**Ces modèles opèrent à l'interface entre syntaxe et sémantique et ils permettent (entre autre) de formaliser des équivalences par application de schémas transformationnels**

**On peut ainsi construire des modèles sémantiques totalement formalisés en définissant un "domaine de discours" où la dénotation consiste à identifier des sous-domaines.**

**Définir le sens des expressions se ramène à déterminer des fonctions indicatrices sur les domaines de valeurs possibles**

université PARIS DIDEROT PARIS 7

# La modélisation des quantifieurs
# Représentations formelles (4/9)

**Ingrédients de la modélisation formelle des quantifieurs:**
**Prédicat Quantifié = Syntagme Nominal Quantifié + Verbe**
***Quantified Nominal Phrase (QNP) + Verb***
**→ Domaine de référence du quantifieur (domaine du nom)**
**→ Portée (scope) du quantifieur (domaine de la prédication)**
**On peut écrire des expressions formelles n'utilisant que les opérateurs de la logique du premier ordre (avec ∃ et ∀) pour les quantifieurs les plus simples (A, ALL) mais un ordre plus élevé (cadre de la théorie des types) est requis pour d'autres quantifieurs**
**Dans le cas général, on a le type: $(e \rightarrow t) \rightarrow ((e \rightarrow t) \rightarrow t)$**
**En effet, un quantifieur relie un nom commun ( type: $e \rightarrow t$) à un verbe prédicatif (également de type: $e \rightarrow t$)**

**Un quantifieur généralisé est un opérateur abstrait liant deux ensembles: le restricteur (NP) et la prédication (VP)**

université PARIS DIDEROT PARIS 7

# La modélisation des quantifieurs
# Représentations formelles (5/9)

**En développant ce genre de modélisation on peut définir des propriétés intéressantes comme la conservativité:**

**Q (Y) est conservatif ssi: $X \in Q(Y) <=> (X \cap Y) \in Q(Y)$**

**(Ou Q dénote le quantifieur, X le domaine et Y le prédicat)**

**Ceci signifie que les entités qui ne font pas partie du domaine de référence n'ont pas d'impact sur la « portée nucléaire » ("nuclear scope") définie par la prédication.**

**Par exemple, si je dis: « tous les chats miaulent » le fait de savoir si un autre animal peut miauler n'a aucun impact sur la valeur de vérité de cette proposition.**

**Cela peut sembler une évidence, et pourtant cette propriété n'est pas toujours vérifiée. Par exemple, si on considère la proposition:**

**« Beaucoup de Scandinaves ont reçu le prix Nobel »**

***(exemple donné dans l'ouvrage de Cann & al.)***

université PARIS DIDEROT PARIS 7

# La modélisation des quantifieurs
# Représentations formelles (6/9)

**Dans une interprétation fréquentiste, cette phrase admet 3 sens:**
**1°) « si X est un citoyen scandinave, il y a une probabilité élevée que X ait reçu le prix Nobel », ce qui est pragmatiquement absurde**
**2°) « si X a eu un prix Nobel, il y a une probabilité élevée pour que X soit scandinave », ce qui est tout aussi contrefactuel**
**3°) « dans les pays scandinaves, le ratio prix Nobel / habitant est plus élevé que pragmatiquement attendu (par exemple plus élevé que la moyenne par pays/région du monde) » qui est le seul sens raisonnable**
**Le fond du problème est que, dans cette occurrence particulière, 'many' n'est pas conservatif: ce qu'il signifie dépend d'autre chose que du seul domaine de référence (l'ensemble des Scandinaves)**

**D'où l'impossibilité d'une sémantique formelle "pure" des langues naturelles (c-à-d se situant totalement en amont de l'interprétation pragmatique) :**
'*many* may be treated as **underspecified** for its full interpretation. To interpret an utterance of the word, therefore, the hearer has to make certain choices based on context (including world knowledge), i.e. pragmatic choices must be allowed to precede semantic interpretation if we do not want to take the easy way out and say that *many* is multiply **ambiguous**' (Cann & al. 2009 – p114.)

université PARIS DIDEROT PARIS 7

# La modélisation des quantifieurs
# Représentations formelles (7/9)

**Conclusion: la modélisation par la spécification de modèles à base de domaine et de portée ne permet pas de représenter la sémantique des quantifieurs de manière complète même si on peut formaliser certaines propriétés logiques intéressantes (comme la monotonie croissante ou décroissante par rapport au restricteur ou à la portée: par exemple, NO ou FEW sont monotones décroissants vis-à-vis de la portée (dans le jargon moins formalisé de la grammaire énonciative on parle de « polarité négative » pour FEW)**

**Cependant, une limitation forte de cette approche est qu'elle n'aborde les quantifieurs que de manière individuelle en ne disant rien de leurs inter-relations: des notions "évidentes" telles que l'ordre des quantifieurs ou la distance qui sépare intuitivement deux quantifieurs ne sont pas représentées, d'où l'intérêt d'introduire une autre approche**

# La modélisation des quantifieurs
# Représentations formelles (8/9)

**Modèles algébriques: les types abstraits de données (ADT)**
**→ substrat formel des langages informatiques « à objets »**

**Types**

**Stack [E]** ***(type abstrait paramétré par un autre type E)***

**Fonctions**

**Empty: Stack → BOOLEAN**
**New: → Stack**
**Push: E x Stack → Stack**
**Pop: Stack → Stack**
**Top: Stack → E**

**Préconditions**

**Pop (s) : NOT empty (s) ; Top (s) : NOT empty (s)**

**Axiomes**

**∀ s ∈ Stack, ∀ e ∈ E,**
**Empty (New)**
**NOT Empty (push (e, s)**
**Top (Push (e,s)) = e**
**Pop (Push (e,s)) = s**

***(je reprends ici un exemple de l'ouvrage de B. Meyer)***

# La modélisation des quantifieurs
# Représentations formelles (9/9)

**Les ADT permettent de caractériser un type d'objet par les opérations que l'on peut lui appliquer (c-à-d une spécification d'interface)**

**Je me suis inspiré de ce modèle pour développer un modèle algébrique proposant une vision « systémique » des quantifieurs scalaires afin d'inscrire dans un cadre mathématique des propriétés considérées traditionnellement comme relevant de la pragmatique (comme l'ordre « naturel » entre les quantifieurs)**

**L'idée est d'enrichir la formalisation de la sémantique des quantifieurs à travers des opérateurs binaires qui les mettent en relation, de façon à faire ressortir les spécificités de chacun vis-à-vis de tous les autres.**

**Encore faut-il assurer une certaine traçabilité de ce modèle par rapport à l'analyse empirique des langues naturelles.**
**Pour bâtir et consolider le modèle on a donc 3 points à vérifier:**

- **l'universalité des quantifieurs retenus,**
- **le respect des axiomes du modèle algébrique,**
- **l'observabilité empirique des relations prévues par le modèle**

université PARIS DIDEROT PARIS 7

# La modélisation des quantifieurs
# Universaux sémantiques

**La vieille quête de la proto-langue universelle (notre langue commune d'avant la tour de Babel...) a fait place depuis les années 70 à un programme de recherche des « universaux sémantiques », c-à-d de sèmes primitifs qui aient des représentations dans toutes les langues terrestres (y compris des langues sans forme écrite).**

**Le travail pionnier d'A. Wierzbicka a été complété et développé par ses collègues – en particulier australiens – et a débouché sur des listes de quelques dizaines de notions primitives, dont les quantifieurs:**

**ALL, MANY/MUCH et SOME (et possiblement FEW, selon Goddard)**

**Certaines autres primitives universelles sont susceptibles d'être interprétées et/ou combinées pour identifier d'autres quantifieurs:**

**NOT → NO** **A + PART → SOME**

**NEAR → ALMOST** **VERY + SMALL + PART → VERY LITTLE**

Quelques références:

Wierzbicka, A. *Semantic Primitives Across Languages: A Critical Review*
p 445-450 Semantic and Lexical Universals : Theory and empirical findings
(C. Goddard and A. Wierzbicka Eds) John Benjamin 1994

Goddard, C. *The search for the shared semantic core of all languages* p 5-40.
Meaning and Universal Grammar - Theory and Empirical Findings/. Vol I.
John Benjamin 2002

université PARIS DIDEROT PARIS 7

# La modélisation systémique des quantifieurs
# Anneaux minimaux de quantifieurs

**Résumé de l'épisode précédent (séminaire CLILLAC d'avril 2011):**

- vision systémique contrastive de la sémantique des quantifieurs scalaires
- modèle algébrique d'anneau commutatif AQ (Min, Gap)
- extension par des opérateurs unaires (approximation, intensification)
- quelques exemples dans 6 langues européennes (DE, EN, ES, FR, IT, RU)

**Mon programme de travail 2011-2012 visait la consolidation du modèle:**

- consolidation des ensembles de base à partir des universaux
- exploitation des corpus aisément accessibles (BNC, COCA, CDE)
- recherche complémentaire de co-occurrences via Google

**Exemple: tables d'opérateurs Min et Gap pour {NO, SOME, MUCH, ALL}**

| Min | Q-n | Q-c | Q-g | Q-t |
|---|---|---|---|---|
| Q-n | Q-n | Q-n | Q-n | Q-n |
| Q-c | Q-n | Q-c | Q-c | Q-c |
| Q-g | Q-n | Q-c | Q-g | Q-g |
| Q-t | Q-n | Q-c | Q-g | Q-t |

| Gap | Q-n | Q-c | Q-g | Q-t |
|---|---|---|---|---|
| Q-n | Q-n | Q-c | Q-g | Q-t |
| Q-c | Q-c | Q-n | Q-g | Q-g |
| Q-g | Q-g | Q-c | Q-n | Q-g |
| Q-t | Q-t | Q-c | Q-g | Q-n |

université PARIS DIDEROT PARIS 7

# Consolidation empirique
# Aux limites des corpus existants

**Certaines propriétés structurales exprimées par ce modèle algébrique sont non-triviales et difficiles à vérifier empiriquement:**

**Par exemple, Gap (Qt, Qg) = Qg signifie que: « des co-occurrences de MUCH/MANY et ALL existent pour lesquelles la différence entre les deux est égal à MUCH/MANY »**

**A première vue, on pourrait se laisser aller à penser que cet écart est plutôt majoré par SOME soit l'équation: Gap (Qt, Qg) = Qc**

**Les corpus en ligne existants ne permettent pas de trancher, car les co-occurrences de Qt et Qg sont peu nombreuses:**

| | COCA | BNC | CDE |
|---|---|---|---|
| much/many or all | 37 | 9 | |
| Mucho(s) or todo(s) | | | 9 |

**Pas de co-occurrences contrastant 'many' avec lui-même sous la forme "many X... but many X..." sur le même domaine de référence: "many intellectuals... but many ordinary people..." dans COCA**

université PARIS DIDEROT PARIS 7

# Consolidation empirique
# Exploiter les moteurs de recherche

**Pour les langues qui ne sont pas équipées de corpus en ligne facilement exploitables ou pour détecter des co-occurrences complexes et/ou rares, la seule solution est Mister Google, avec tous les problèmes que cela pose:**

- **– instabilité de l'espace de recherche,**
- **– nombreuses duplications d'instances,**
- **– nombreuses erreurs de scripteurs inattentifs ou non-natifs...**

*(par exemple, sur la Toile, environ 10% des occurrences de 'muchos' et 'muchas' contiennent des erreurs d'accord en genre et/ou en nombre !)*

**En contrepartie on détecte environ 1000 fois plus de co-occurrences de quantifieurs que dans les corpus contrôlés:**

*Par exemple, le 13 août 2011, Google livrait 10 800 co-occurrences contrastives de Qg et Qt en français*: « l'UNESCO, **beaucoup, sinon tous** ici, la connaissent bien » (Source : cidecquebec2011.org/docs/GSaoumaForero.pdf)

**On peut ainsi trouver des exemples prouvant que Gap (Qg, Qt) = Qg:**

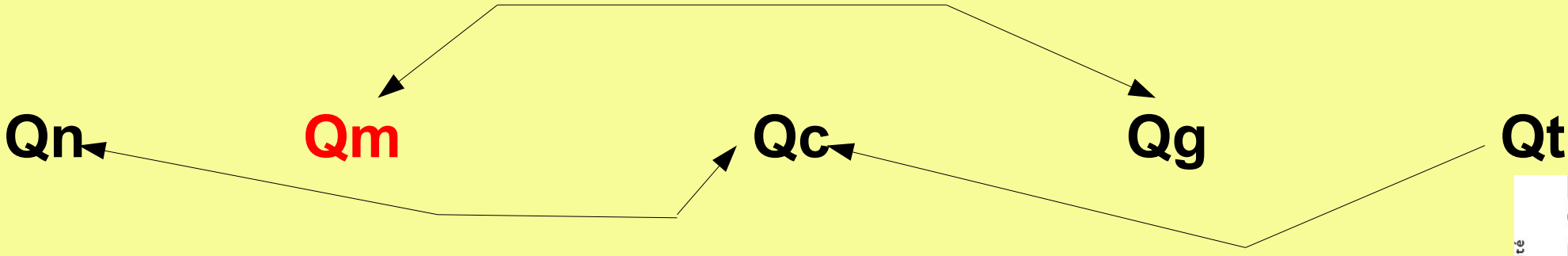
***Many agreed, but many disagreed as well***
***Many agreed, but many others had a different point of view***
***Many agreed but many opposed my views***
***Etc...***

université PARIS DIDEROT PARIS 7

# La modélisation systémique des quantifieurs
# Extension et consolidation du système (1/3)

**Pour que les quantifieurs scalaires « fassent système » la cohérence (consistency) et la complétude (completeness) de l'ensemble doivent pouvoir être vérifiées: tous les opérateurs usuels doivent être internes.**

**Dans une approche formelle, la négation d'un quantifieur renvoie à tout l'ensemble complémentaire comme champ des valeurs possibles:**
**Ainsi, NOT (Qg) = (Qn ∨ Qc∨ Qt) dans le cadre du modèle à 4 valeurs**
**Mais en réalité, quand on dit "pas beaucoup" on ne veut dire ni "tout" ni "rien" ni "une certaine quantité": le modèle à 4 valeurs est incomplet vis-à-vis de l'opérateur NOT, et maintenir l'internalité de la négation nous amène donc à étendre le modèle avec une 5ème valeur:**




**Qm (= 'peu de') intègre donc le système comme représentation canonique du "pas beaucoup" (combinaison des universaux NOT et MANY/MUCH)**

université PARIS DIDEROT PARIS 7

# La modélisation systémique des quantifieurs
# Extension et consolidation du système (2/3)

**Généralisation des extensions par combinaison d'universaux:**

**NOT (MUCH/MANY) = LITTLE/FEW (codage: Qm)**

**A + SMALL + PART = A LITTLE (codage: Qp)**

**NEAR + NO = ALMOST/NEARLY NO (codage: Qan) etc.**

**=> vérification empirique par les co-occurrences contrastives:**

| Co-occurrences contrastives | COCA (11/8/11) | BNC (11/8/11) | CDLE (11/8/11) |
|---|---|---|---|
| Q-an / Q-n<br>No or almost/nearly/practically no<br>Ningun o casi ningun (+inflections)<br>Nada o casi nada (de) | 0 | 0 | <br><br>1<br>8 |
| Q-an / Q-m<br>Few/little or almost/nearly/practically no<br>Poco o casi nada (de)<br>Poco o casi ningun (+ inflections) | 1 | 0 | <br><br>1<br>1 |
| Q-m / Q-n<br>Little or no<br>Few or no<br>Poco o ningun (+inflections)<br>Poco o nada (de) | <br>2583<br>152 | <br>830<br>32 | <br><br><br>4<br>1 |

**On peut observer au passage d'autres phénomènes intéressants**

# La modélisation systémique des quantifieurs
# Extension et consolidation du système (3/3)

**On peut ensuite rechercher des co-occurrences contrastives via l'Internet pour pallier les limitations des corpus:**

| Contraste | Langue | Citations | total |
|---|---|---|---|
| Q-n / Q-an | EN | Here's what is robbing our government: **no or almost no** taxes for huge multi-billion-dollar corporations. (source: huffingtonpost.com – 2011-Mar-3) | 438,000 |
| | | | |
| Q-n / Q-an | FR | Si ce courant passe dans un fil de cuivre, il n'y a **aucun, ou presque aucun** effet. (source : documents.irevues.inist.fr - *article : Bistouri électrique et coagulation par plasma d'argon (APC) D. COUMAROS, P. SCHLÜTER)* | 12,940 |
| Q-m / Q-an | FR | Dans l'armée, **peu ou presque pas de** citoyens. (source : mediterranee-antique/info/fontane/rome) | 44,900 |
| Q-n / Q-m | FR | Grève des postes: **peu ou pas de** courrier ce matin encore (source : sudpresse.be - 2011-Feb-14) | 6,300,000 |

**On peut donc étendre le modèle contrastif inter-langue à des quantifieurs dérivés (qui ne sont pas des universaux primitifs)**

# La modélisation systémique des quantifieurs
# Anneau maximum à 8 éléments

| Gap | Q-n | Q-an | Q-im | Q-m | Q-p | Q-c | Q-g | Q-t |
|---|---|---|---|---|---|---|---|---|
| Q-n | *Q-n* | Q-an | Q-im | *Q-m* | *Q-p* | *Q-c* | *Q-g* | *Q-t* |
| Q-an | Q-an | Q-n | Q-im | Q-m | Q-p | Q-c | Q-g | Q-t |
| Q-im | Q-im | Q-im | Q-n | Q-m | Q-p | Q-c | Q-g | Q-g |
| Q-m | *Q-m* | Q-m | *Q-m* | *Q-n* | *Q-p* | *Q-c* | *Q-g* | *Q-g* |
| Q-p | *Q-p* | Q-p | *Q-p* | *Q-p* | *Q-n* | *Q-c* | *Q-g* | *Q-g* |
| Q-c | *Q-c* | Q-c | *Q-c* | *Q-c* | *Q-c* | *Q-n* | *Q-g* | *Q-g* |
| Q-g | *Q-g* | Q-g | *Q-g* | *Q-c* | *Q-g* | *Q-g* | *Q-n* | *Q-g* |
| Q-t | *Q-t* | Q-t | Q-g | *Q-g* | *Q-g* | *Q-g* | *Q-g* | *Q-n* |

**Je n'ai conservé que les valeurs compatibles avec la structure algébrique d'anneau et révélées empiriquement comme distinctes (par exemple Qim = 'très peu' ≠ Qm = 'peu', mais Qig = 'vraiment beaucoup' = Qg='beaucoup') Qat = "presque tout" imposerait Gap (Qat, Qt) = Qc, et l'associativité de Gap ne serait pas conservée (NB: de plus, Wiersbicka pense que dans certaines langues, Qat n'a pas d'existence distincte de Qt)**

**Pour plus de détails, lire le papier "RingsOfQuantifiers" sur ma page perso: http://www.eila.univ-paris-diderot.fr/user/michel_delarche**

université PARIS DIDEROT PARIS 7

# Deuxième partie:

## *Analyse statistique de corpus*

**Mise en oeuvre de l'outil NooJ**

**Intérêt et limites de l'analyse automatique**

**Contribution à l'analyse d'un genre littéraire:**
**L'essai de vulgarisation scientifique**

**Contribution à l'analyse de style et de genre:**
**Jane Austen et Maupassant**

université PARIS DIDEROT PARIS 7

# Détection semi-automatique des occurrences
# Mise en oeuvre de l'outil NooJ (1/4)

**NooJ est un environnement d'analyse linguistique:**

- **Performant**
- **Multi-langue**
- **Incluant des dictionnaires et grammaires prédéfinies**
- **Programmable par ajout de grammaires morpho-syntaxiques**
- **Pour la recherche de concordance et l'analyse sémantique**

**Il a été développé par Max Silberzstein (voir http://www.nooj4nlp.net) qui m'a obligeamment et diligemment aidé à résoudre mes problèmes techniques**

**J'ai développé de petites grammaires pour repérer et annoter (avec le codage présenté précédemment) les occurrences de quantifieurs dans des textes en anglais, espagnol et français:**

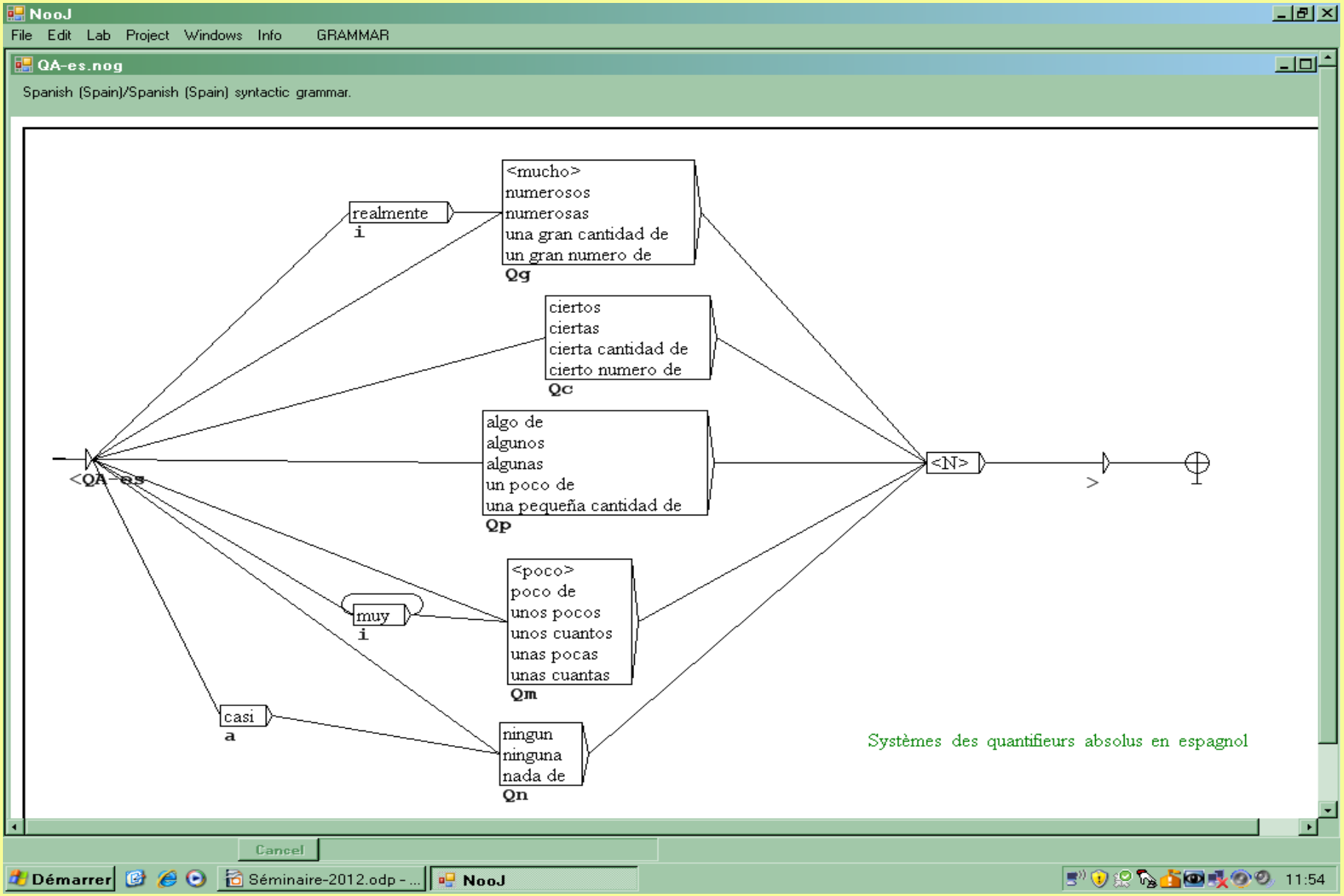
**Texte-source → analyse grammaticale automatique → liste d'occurrences annotées**
**liste d'occurrences → inspection et nettoyage manuel → occurrences pertinentes**
**Occurrences pertinentes → comptage → statistiques d'utilisation des quantifieurs**

**Pour chaque langue, j'ai construit deux graphes d'analyse:**
**- un graphe pour extraire les quantifieurs utilisés sans référence explicite à une totalité de référence (quantifieurs absolus)**
**- un graphe pour les quantifieurs faisant référence à une totalité sous-jacente (quantifieurs relatifs)**

université PARIS DIDEROT PARIS 7

# Détection semi-automatique des occurrences
# Mise en oeuvre de l'outil NooJ (2/4)

## Exemple de graphe d'analyse (quantifieurs absolus en espagnol)



# Détection semi-automatique des occurrences
# Mise en oeuvre de l'outil NooJ (3/4)

**Occurrences annotées de quantifieurs relatifs (chap. 1 de Pride & Prejudice):**

take possession before Michaelmas, and **some of his/<RQc>** servants are to be in
her the preference." "They have **none of them/<RQn>** much to recommend them,"

**Occurrences annotées de quantifieurs absolus (chap. 1 de Pride & Prejudice):**

want of a wife. However **little known/<AQm>** the feelings or views of
as the rightful property of **some one/<AQc>** or other of their daughters
about it." Mr. Bennet **made no answer/<AQn>** . "Do not you want to
tell me, and I have **no objection/<AQn>** to hearing it." This was
as he comes." "I see **no occasion/<AQn>** for that. You and the
a woman has not often **much beauty/<AQg>** to think of." "But, my
you; and I will send **a few lines/<AQp>** by you to assure him
a good word for my **little Lizzy/<AQm>** ." "I desire you will do
no such thing. Lizzy is **not a bit/<AQn>** better than the others; and
in vexing me. You have **no compassion/<AQn>** on my poor nerves."
it, and live to see **many young men/<AQg>** of four thousand a year
the neighbourhood." "It will be **no use/<AQn>** to us if twenty such
a woman of mean understanding, **little information/<AQm>**, and uncertain temper.

**Si l'on ne veut pas trop complexifier la grammaire d'analyse, on doit accepter d'avoir à finaliser manuellement le filtrage des occurrences**

université PARIS DIDEROT PARIS 7

# Détection semi-automatique des occurrences
# Mise en oeuvre de l'outil NooJ (4/4)

## Quelques limites de NooJ:

**1°) Les textes scientifiques comporte souvent des valeurs numériques notées avec des puissances de 10 ou des symboles spéciaux, mais NooJ ne sait pas traiter les exposants, par exemple *(in* J. Monod *Le hasard et la nécessité* - Ed. Seuil - 1970)**
***« On peut estimer ce nombre à 2500 ± 500, pour la bactérie Escherichia Coli (5 x $10^{-13}$ g en poids et 2 µ en longueur environ) »***

**Solution: remplacer les symboles et formes non reconnus par des caractères reconnus: $10^{-13}$ → 10^-13 )**

2°) **La représentation des verbes français dans le dictionnaire interne de NooJ n'inclut pas la distinction entre le transitif et l' intransitif (qui serait fort utile pour distinguer les compléments indirects des partitifs de quantification):**

**Prendre du bon temps → AQ Qc**
**Se souvenir du bon vieux temps → AQ Qc**

université PARIS DIDEROT PARIS 7

# Détection semi-automatique des occurrences
# Ambigüités morpho-syntaxiques et sémantiques

**Principaux problèmes rencontrés:**

**1°) Le pluriel indéfini français peut avoir le sens de AQc (=un certain nombre de, une certaine quantité de) et des écrivains (pas G de M) utilisent 'de' plutôt que 'des' pour introduire ce genre de nuance:**
**« Ils passent de bons moments ensemble, de plus pénibles aussi. »**
**(P. Garnier *La théorie du panda* – Ed. Zulma – 2008)**
**« Il y avait déjà passé des heures bien dures. »**
**(G. de Maupassant *Notre coeur* )**
**(Dans mes statistiques, tous les pluriels indéfinis ont été exclus)**

**2°) (' 'Some', 'certain', 'quelque', 'cierto'...) + Nom Singulier peuvent être on non des quantifieurs Qc (eg 'some time' vs 'some student') selon que le nom peut fonctionner en "mass count" ou pas**

**3°) 'little' peut-être adjectif ou quantifieur, parfois indémêlablement: 'a little hesitation' = 'un peu d'hésitation' ou 'une petite hésitation' ?**

**4°) Les formes incomplètes restent non-détectées: « Mais voilà, il faudrait de l'argent, beaucoup... » (GdM *Pierre et Jean*)**
**(AQc est détecté, mais AQg reste non-détecté à cause de l'ellipse)**

université PARIS DIDEROT PARIS 7

# Exploration de corpus littéraires: genre et style
# La vulgarisation scientifique (1/4)

**Pourquoi étudier la vulgarisation?**

- un genre littéraire varié mais globalement assez bien délimité
- un « genre-interface » hybride qui entremêle des éléments du style « savant » (sans tout l'appareillage académique) et du style « journalistique »
- une multiplicité de disciplines « à succès » (mathématiques, cosmologie, mécanique quantique, biologie, climatologie, histoire, archéologie, sociologie...)
- un terrain de choix pour l'analyse des différentes stratégies d'expression de l'information quantitative

**Premier axe d'exploration empirique:**

- poids relatifs des quantifieurs scalaires et des données numériques
- interprétations possibles quant aux variations du discours de la vulgarisation (parties « interprétatives » vs parties « techniques »)

**Eléments de mon futur programme de travail:**

- délimitation de sous-genres (taxinomie de la vulgarisation)
- comparaisons inter-langues (français, anglais, espagnol)
- poids relatifs du style de l'auteur et des spécificités du genre

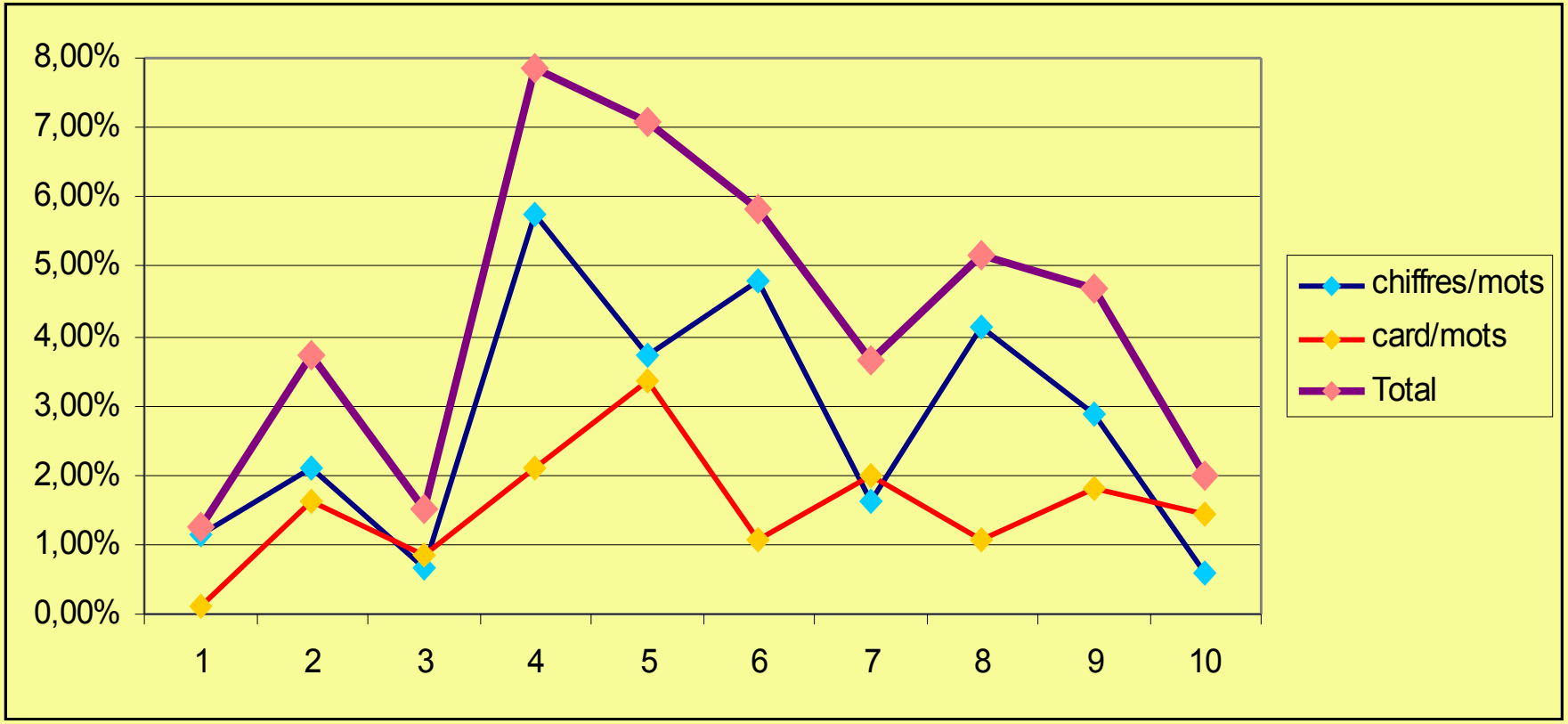
université PARIS DIDEROT PARIS 7

# Exploration de corpus littéraires: genre et style
# La vulgarisation scientifique (2/4)

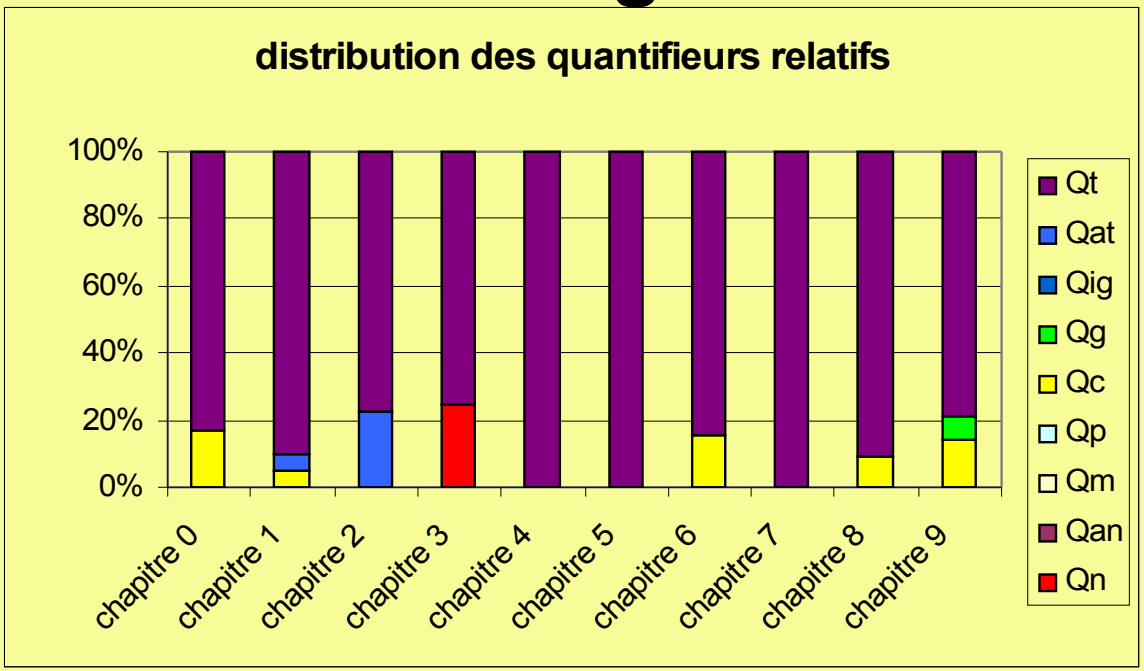
**Ouvrage exploité: J. Monod *Le Hasard et la nécessité*, Le Seuil, 1970**

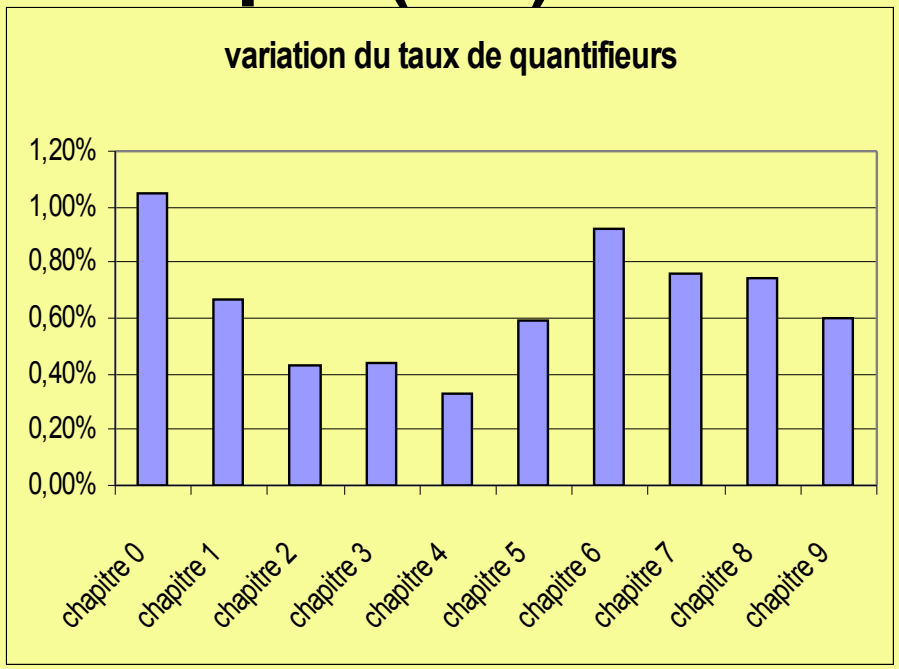
- **écrit par un scientifique reconnu (Prix Nobel de biologie 1965)**
- **décrivant les acquis de la biologie moléculaire des années 50-60**
- **s'inscrivant dans la tradition humaniste des sciences**
- **portant un discours philosophique « anti-animiste » radical**

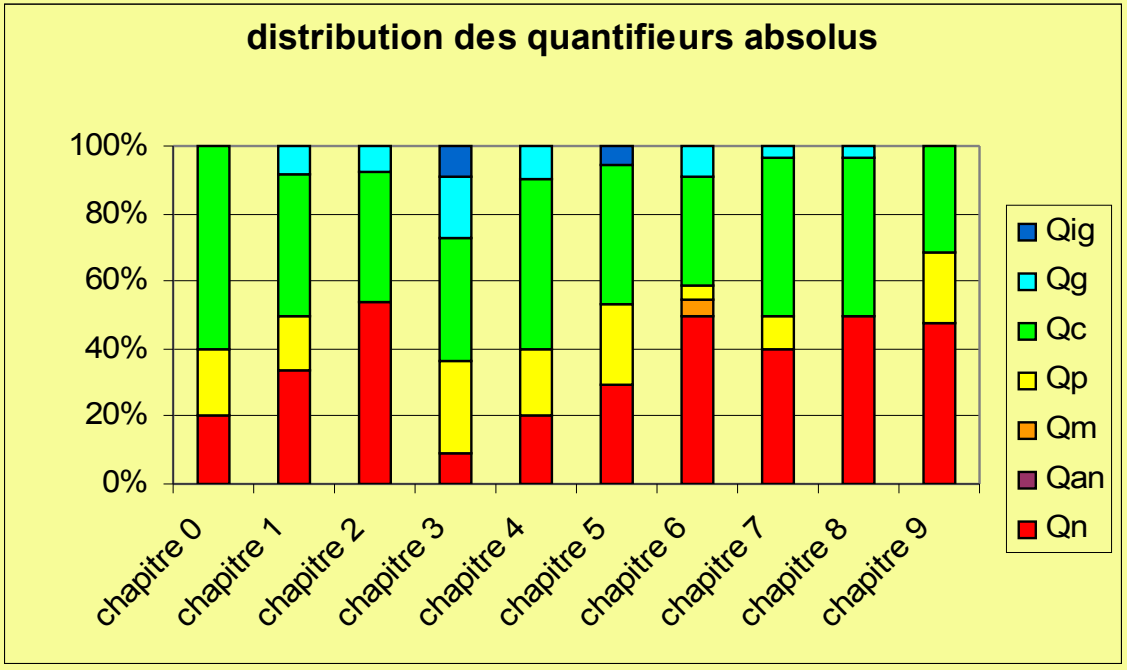
**Variation des taux de numéraux (lettres et chiffres) au fil des chapitres:**



# Exploration de corpus littéraires: genre et style
# La vulgarisation scientifique (3/4)



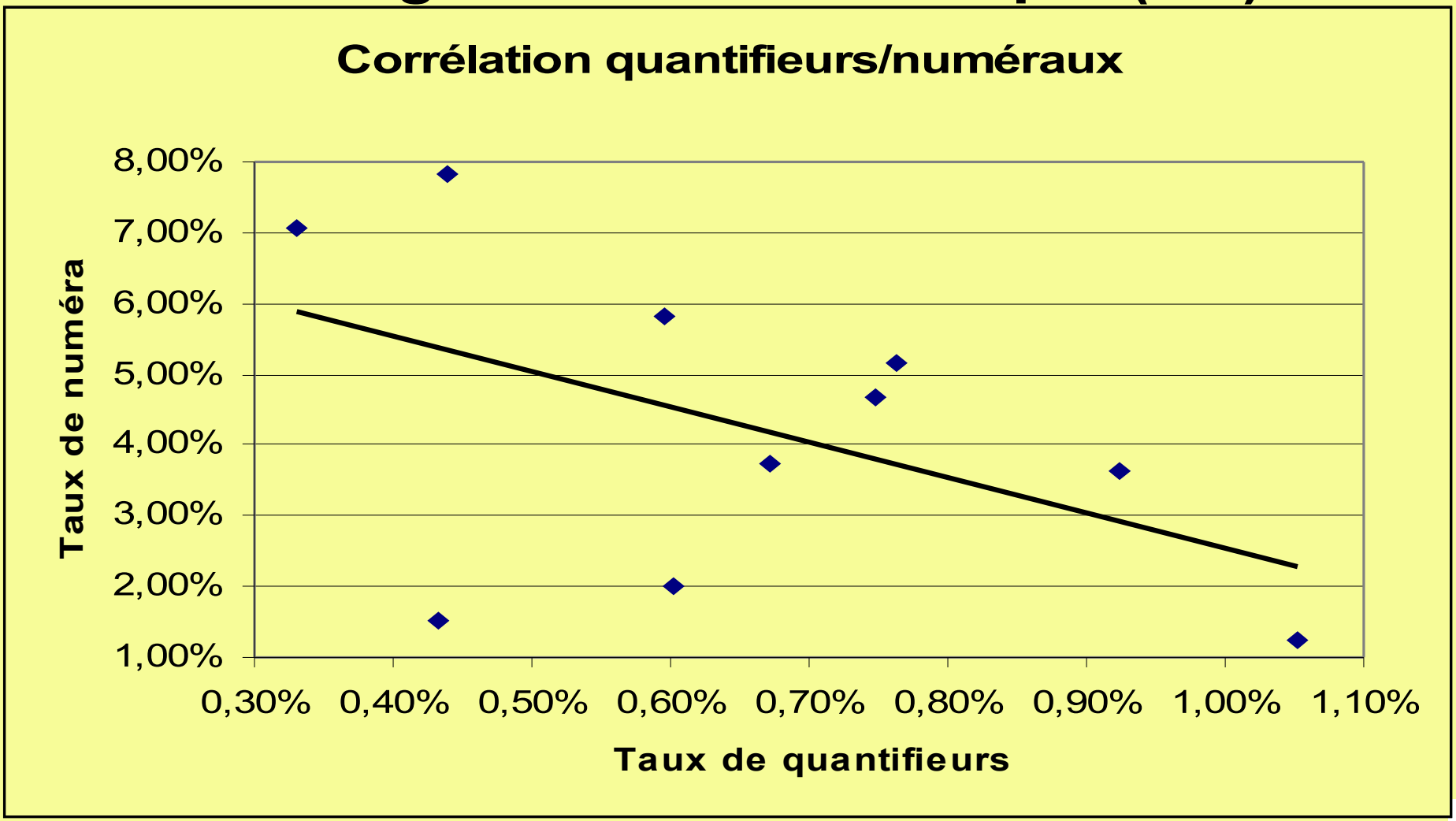




**Les principales caractéristiques de la distribution sont:**
**1°) la domination du "tout" et du "rien" (discours des généralisations conclusives)**
**2°) le poids relativement important de Qc (marqueur de non-généralité ou d'incertitude)**

université PARIS DIDEROT PARIS 7

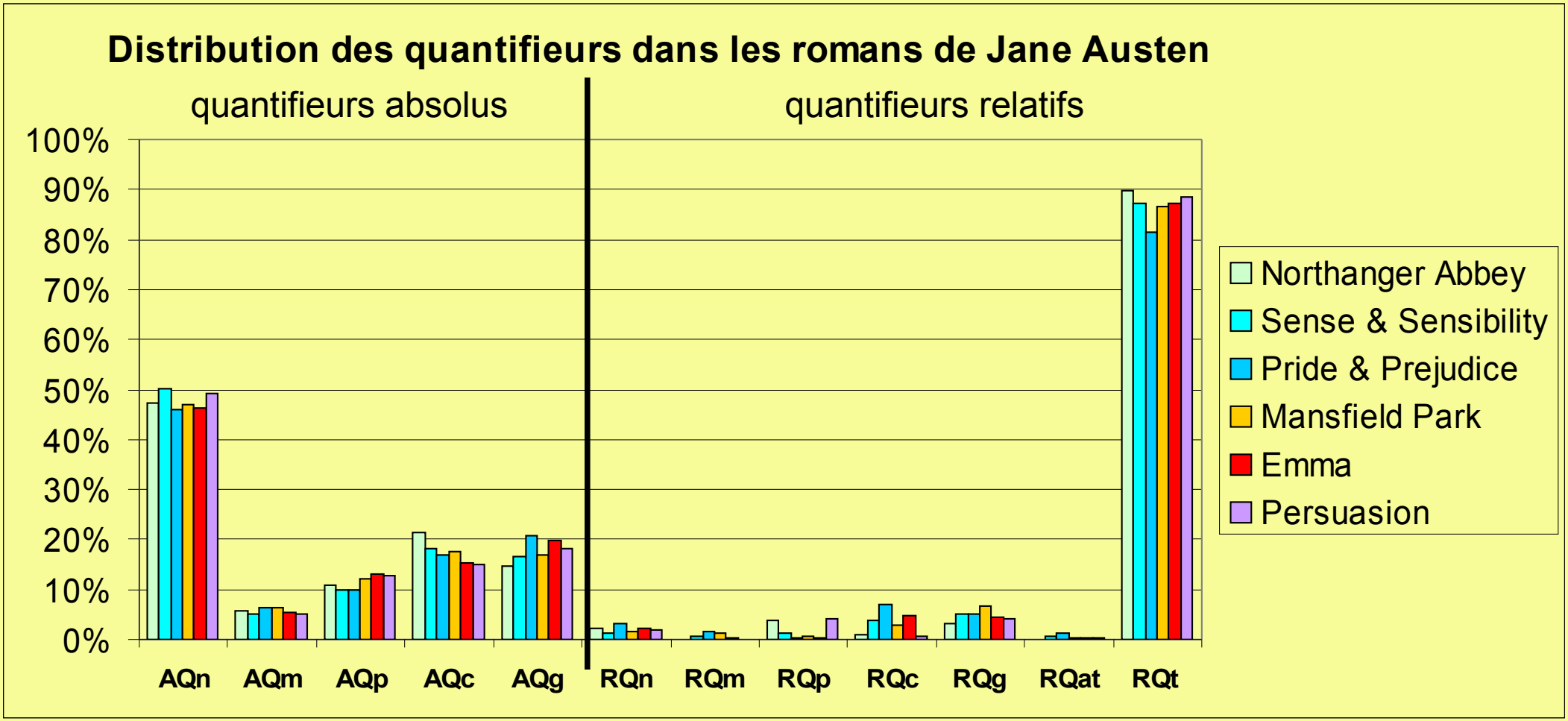
# Exploration de corpus littéraires: genre et style
# La vulgarisation scientifique (4/4)




**On observe une corrélation négative entre les deux formes de la « quantativité », ce qui reflète une certaine dichotomie entre les chapitres à vocation d'exposition des résultats scientifiques et ceux qui sont davantage dédiés aux généralisations philosophiques avec deux chapitres discursifs hors corrélation (chapitres 2 et 9)**

université PARIS DIDEROT PARIS 7

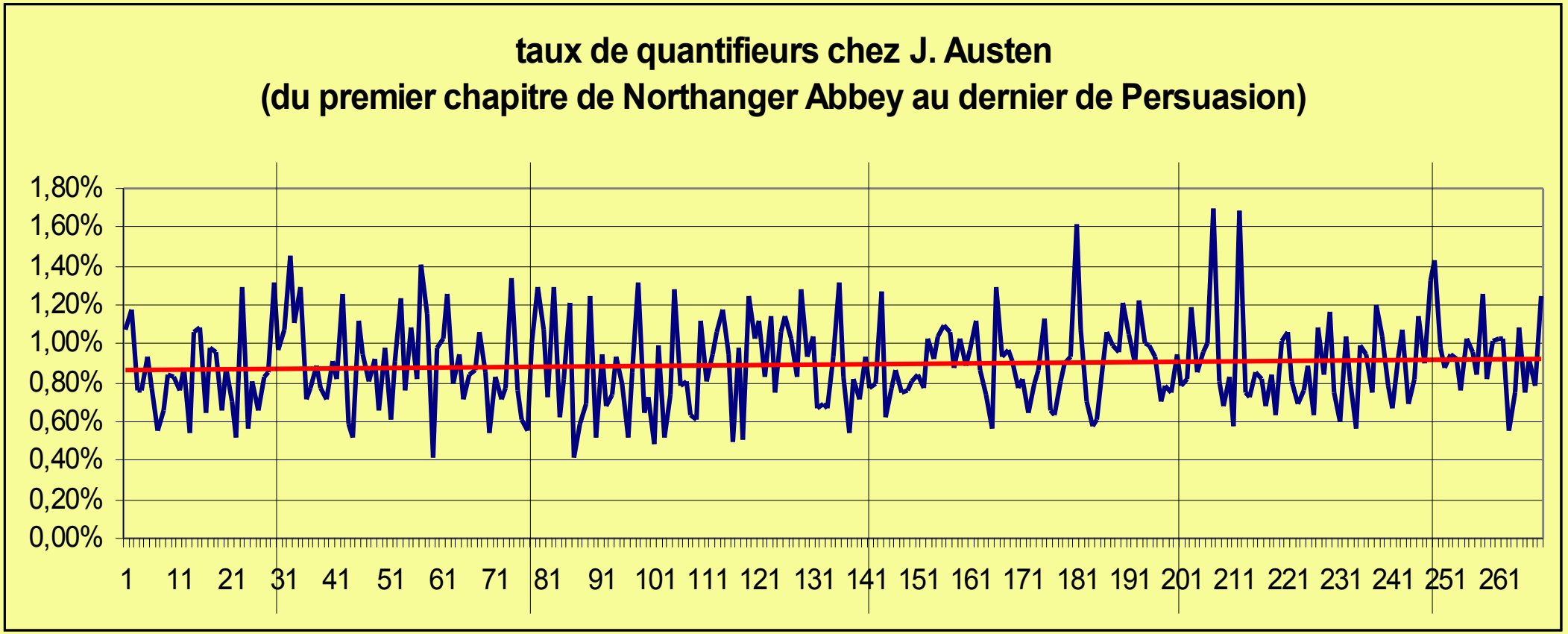
# Exploration de corpus littéraires: genre et style
# Les six romans de Jane Austen (1/6)




**Distribution très homogène, dominée par le "tout" et le "rien"**
**("presque rien": totalement absent, "presque tout": résiduel)**

université PARIS DIDEROT PARIS 7

# Exploration de corpus littéraires: genre et style
# Les six romans de Jane Austen (2/6)

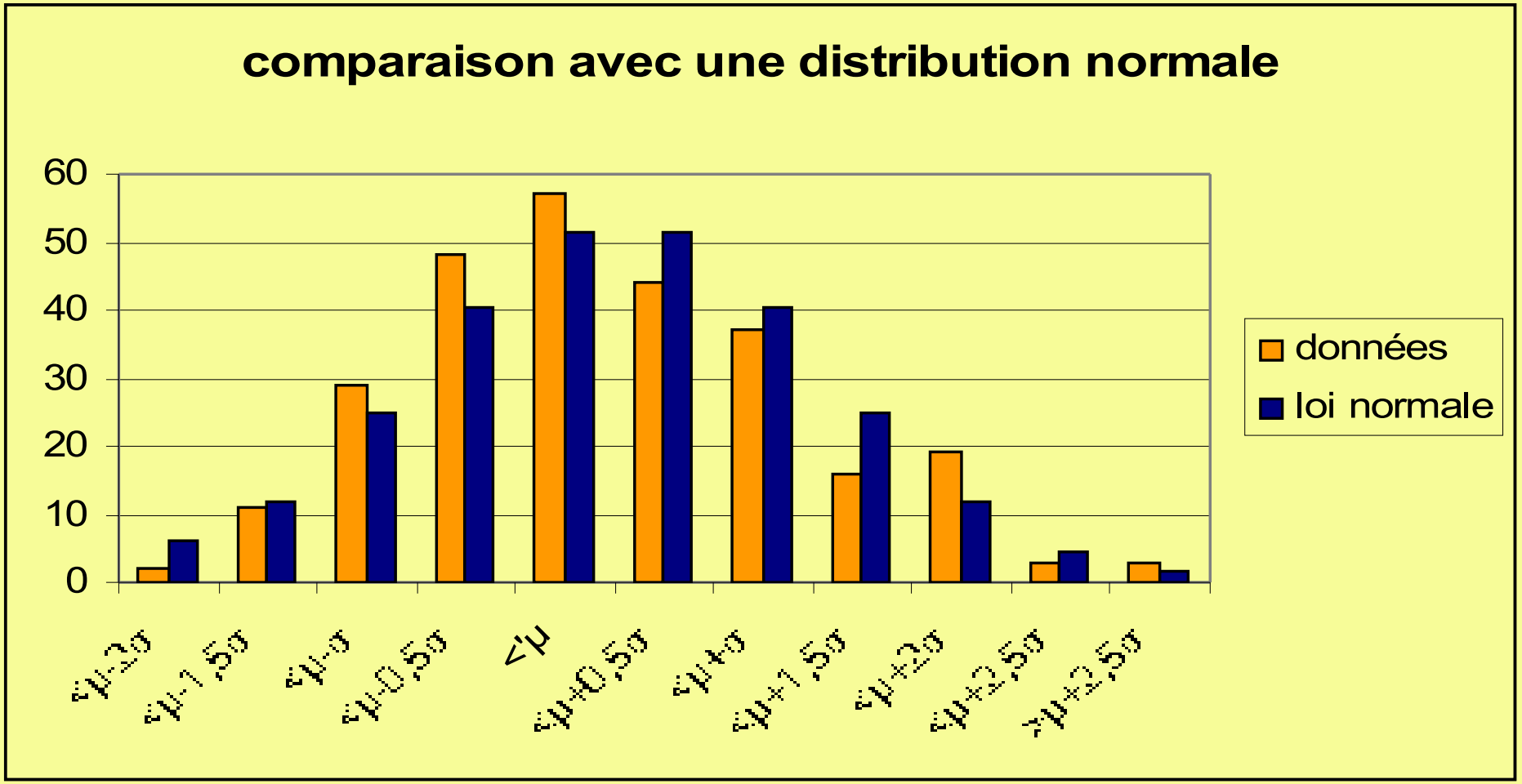



**Sur 15 ans, on n'observe aucune tendance globale d'évolution (grande stabilité de la fréquence d'usage des quantifieurs et de leur distribution)**

**D'où deux interrogations:**
**Le système des quantifieurs est-il un marqueur stylistique fiable ?**
**L'analyse du signal obtenu fait-elle sens narratologiquement ?**

université PARIS DIDEROT PARIS 7

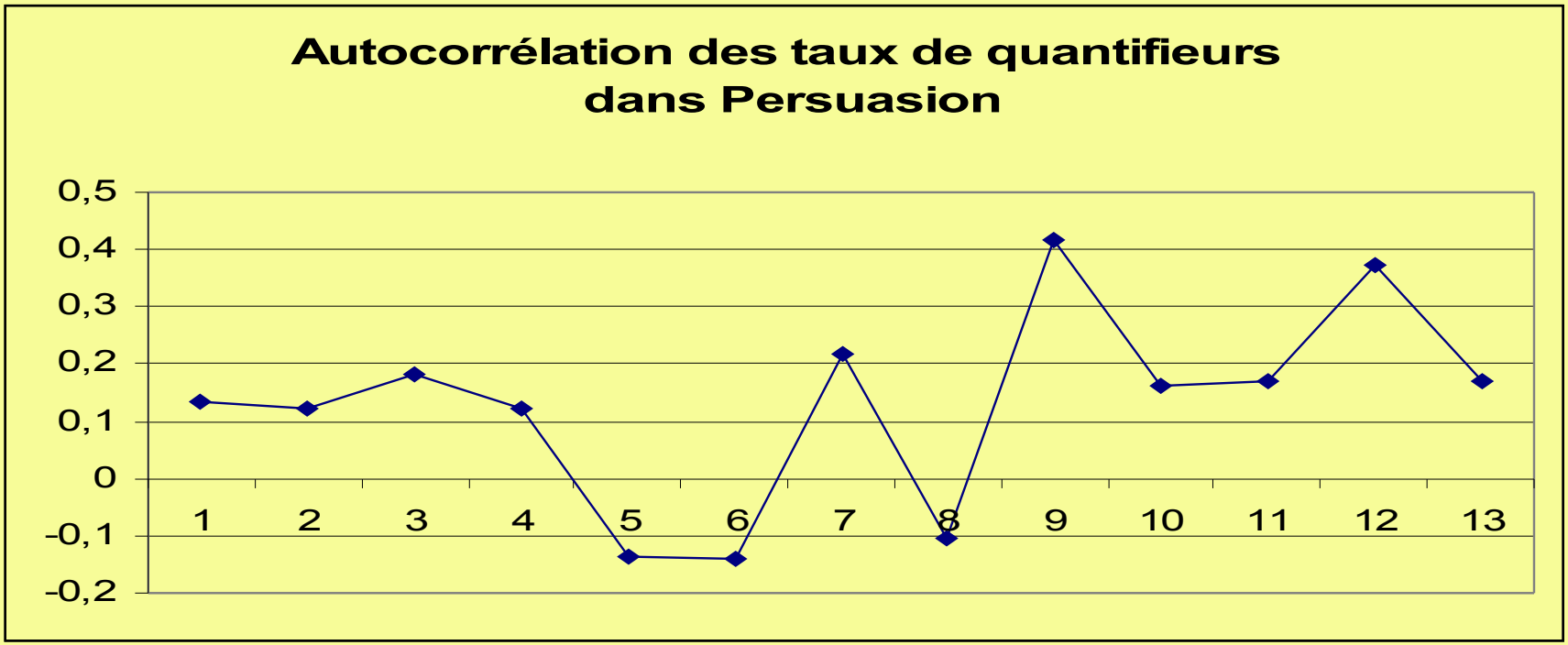
# Exploration de corpus littéraires: genre et style
# Les six romans de Jane Austen (3/6)




**m= 0,892%, eam = 0,181% et σ= 0,227% (CV= 0,25 et EAM/σ=0,79)**
**coefficient d'asymétrie = 0,6 (zéro pour une loi normale)**
**kurtosis (métrique du moment d'ordre 4) = 0,51**
**3 chapitres (≈1,1%) au-delà de μ+3σ (distribution à queue épaisse)**

université PARIS DIDEROT PARIS 7

# Exploration de corpus littéraires: genre et style
# Les six romans de Jane Austen (4/6)

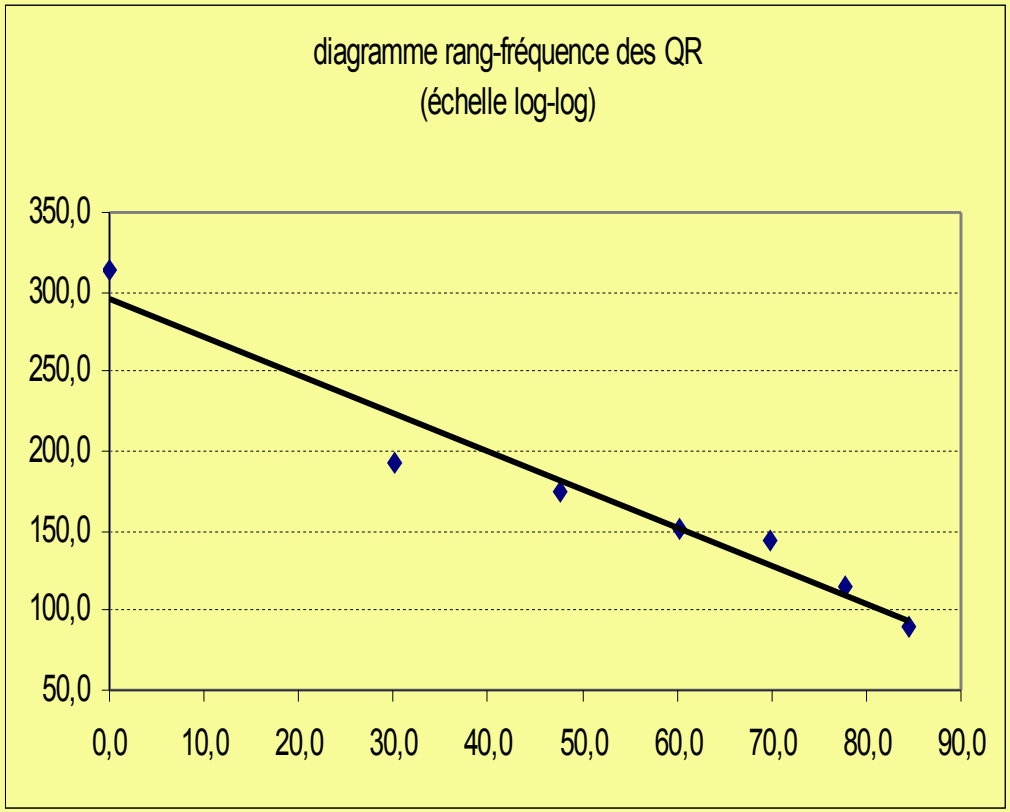


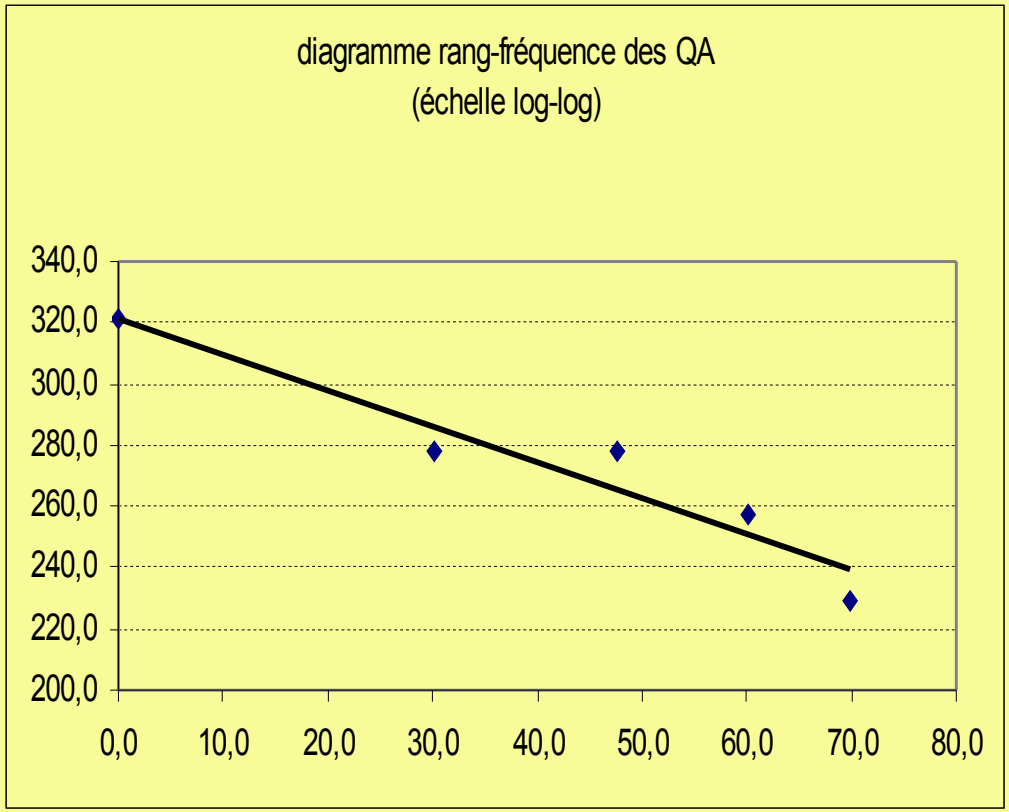

**Pour détecter dans une oeuvre des macro-régularités narratives qui seraient masquées par des fluctuations aléatoires, on peut calculer des estimateurs d'autocorrélation entre les chapitres, par exemple. Ici, les pics observés à x = 9 (resp. x =12) indiquent que les taux globaux de quantifieurs observés aux chapitres10 et suivants (resp. 13 et suivants) sont corrélés à ceux du début.**
**Ces coefficients d'auto-corrélation restent faibles mais peuvent servir d'indice qualitatif pour localiser échos et symétries narratives.**

université PARIS DIDEROT PARIS 7

# Exploration de corpus littéraires: genre et style
# Les six romans de Jane Austen (5/6)

## Chez Jane Austen, les distributions de chacun des deux groupes de quantifieurs suivent la loi de Mandelbrot-Zipf:




**Selon Mandelbrot, la distribution exponentielle décroissante observée par Zipf provient simplement de ce que les textes contiennent de l'information (qu'on appelle aussi « néguentropie »). Le fait que le micro-système des quantifieurs vérifie cette loi montre que le (très) petit sous-ensemble en question contribue, à son échelle, à cette production d'information**

université PARIS DIDEROT PARIS 7

# Exploration de corpus littéraires: genre et style
# Les six romans de Jane Austen (6/6)

## Quelques pistes d'interprétation des résultats:

- le style aphoristique de Jane Austen s'exprime volontiers par généralisations (souvent ironiques) d'où la forte proportion des occurrences de AQn et RQt
- ces généralisations ne s'accompagnent d'aucune "prudence épistémologique" qui appellerait des restrictions ou exceptions prenant la forme du "presque tout" du "presque aucun(e)" ou d'un empirisme casuiste passant par "un certain nombre de"
- ceci se combine dans certains passages à des effets rhétoriques de répétition qui contribuent à renforcer la fréquence de Qn:

"she had no conversation, no stile, no taste, no beauty" (P&P)

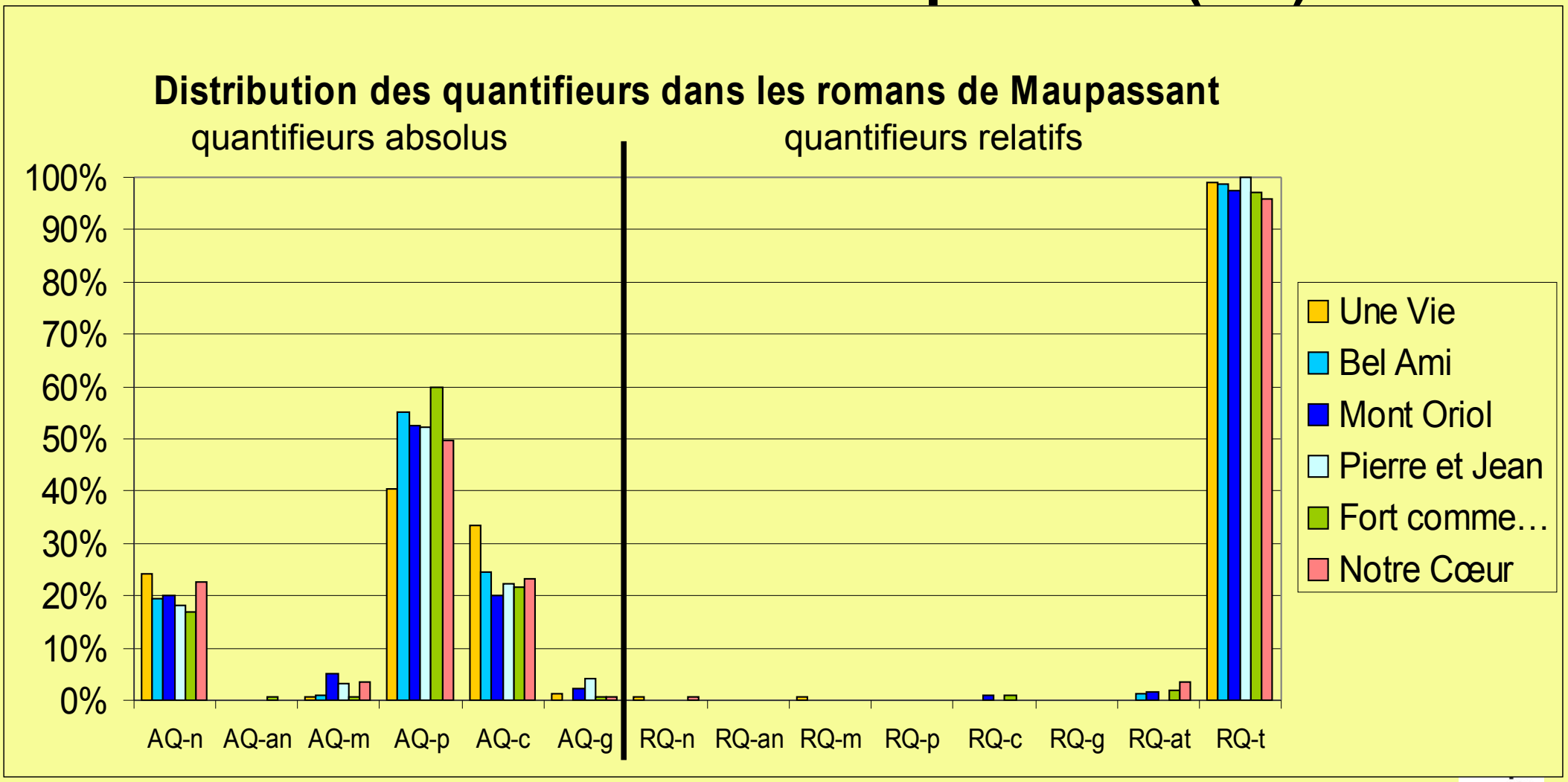
## Points à creuser:

Y a-t-il une composante de genre statistiquement détectable chez des écrivains contemporains de Jane Austen, ce qui relativiserait le degré de spécificité de son style ? (roman gothique, roman féminin du XVIIIème siècle anglais...)

La statistique des quantifieurs peut-elle permettre de distinguer les pastiches contemporains des originaux ?

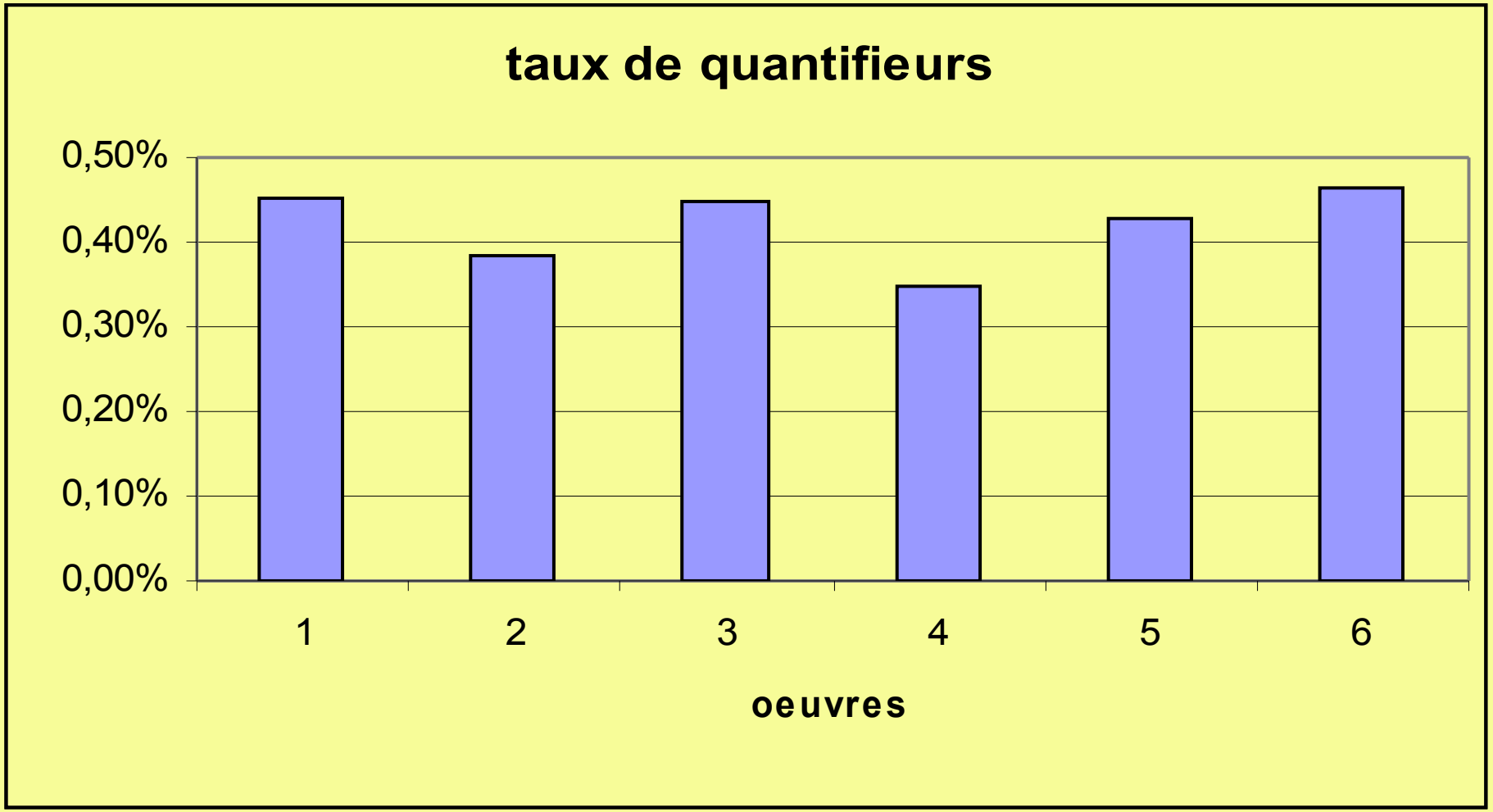
université PARIS DIDEROT PARIS 7

# Exploration de corpus littéraires: genre et style
# Les six romans de Maupassant (1/4)




**Distribution un peu moins homogène que chez Austen, et dominée par le "tout" et le "quelques". Le "beaucoup" est presque absent (AQg n'est quasiment représenté que par des périphrases: "une masse de", "une foule de", "un tas de" etc.)**

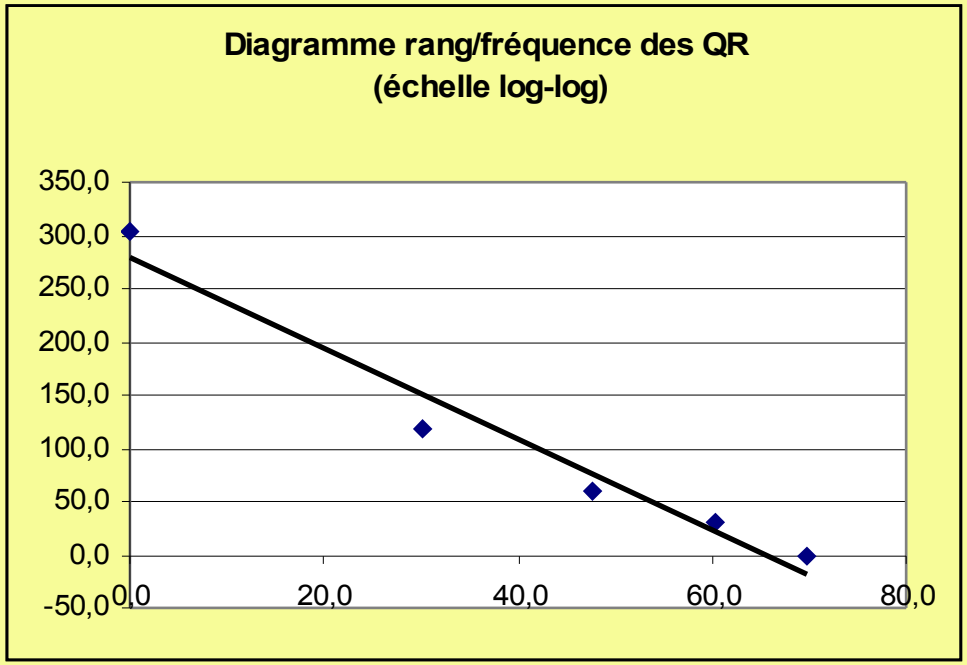
# Exploration de corpus littéraires: genre et style
# Les six romans de Maupassant (2/4)

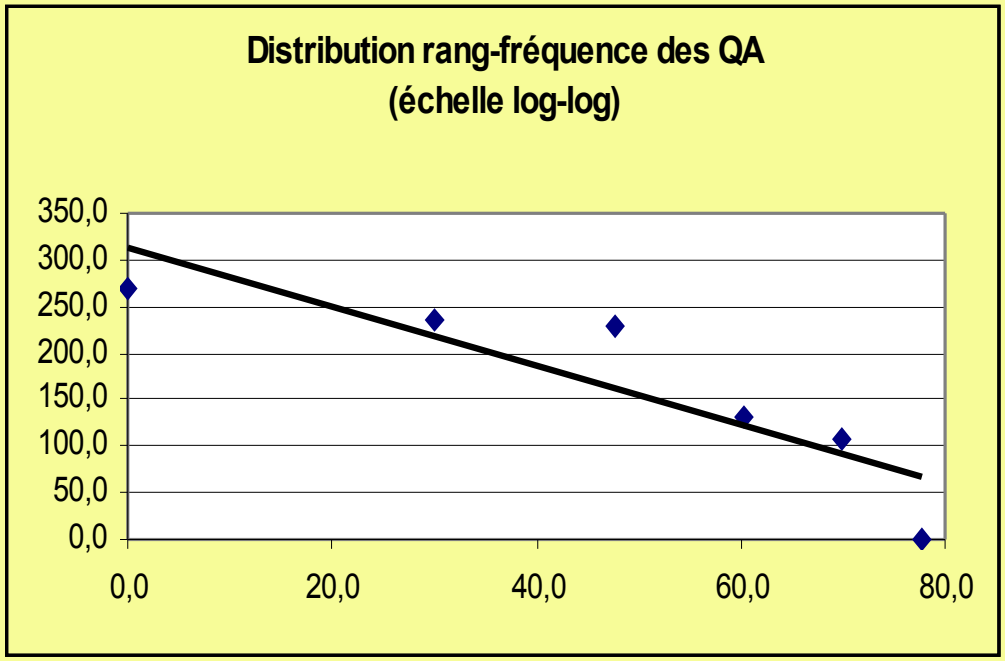


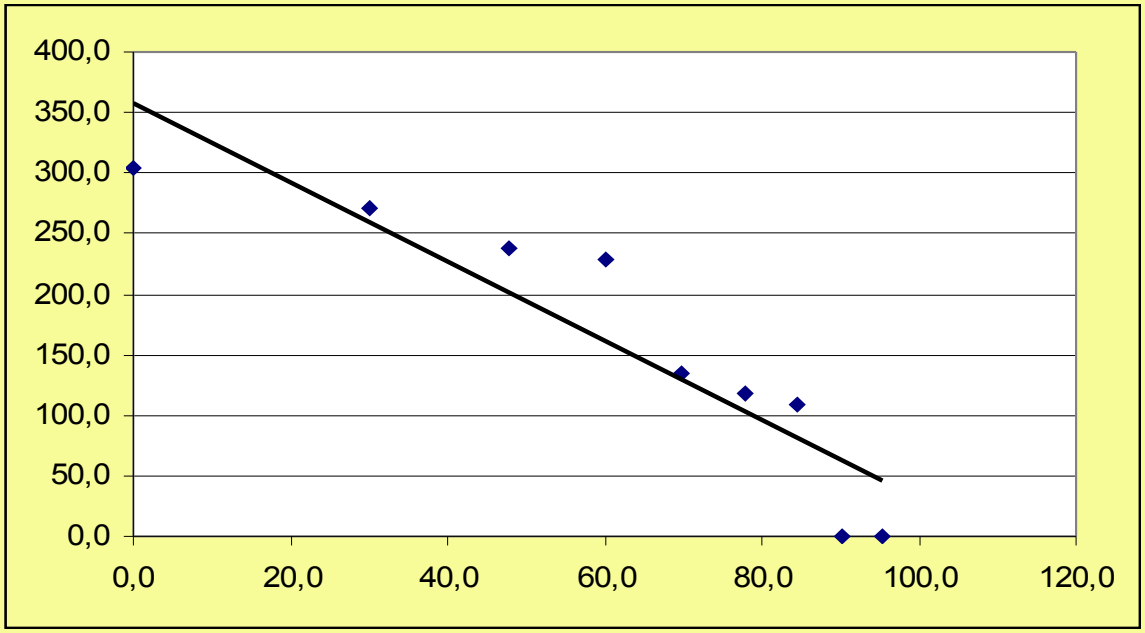

**Guy de Maupassant utilise deux fois moins de quantifieurs que Jane Austen, avec également une très grande stabilité globale (mais sur une période de seulement 6 ans)**

université PARIS DIDEROT PARIS 7

# Exploration de corpus littéraires: genre et style
# Les six romans de Maupassant (3/4)




**Les quantifieurs de Maupassant ne suivent pas très bien la loi de Mandelbrot-Zipf: le sous-système des quantifieurs fonctionne de manière moins homogène que chez Jane Austen (superposition de divers jeux de langage chez GdeM ?)**

université PARIS DIDEROT PARIS 7

# Exploration de corpus littéraires: genre et style
# Les six romans de Maupassant (4/4)

## Quelques pistes d'interprétation:

**- le style naturaliste de Maupassant s'exprime volontiers par des descriptions d'éléments épars dans le paysage, d'où la forte proportion des AQp et AQc (et l'absence de contrepartie en RQ marque le caractère non clos des domaines de référence)**

**- il s'y ajoute un repérage imprécis des durées et des espaces dans une tonalité de brièveté ("quelques instants", "quelques secondes", "quelques minutes", 'quelques heures", quelques jours", "quelques semaines", quelques mois", "quelques pas")**

**- des répétitions contribuent à renforcer la fréquence de Qn: "on ne put la soupçonner d'aucune liaison, d'aucune amourette, d'aucune intrigue" (*Notre Coeur*)**

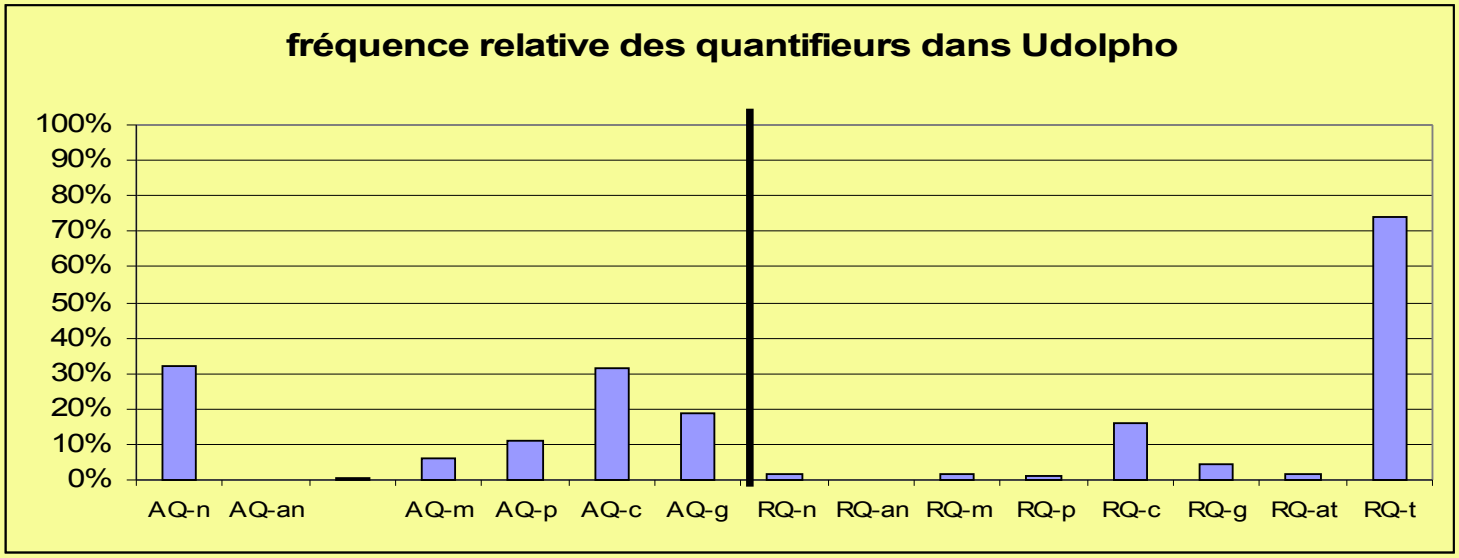
## Points à creuser:

**Y a-t-il une composante de genre statistiquement détectable chez les romanciers contemporains de Maupassant, ou bien le maniement des quantifieurs est-il propre à chaque auteur?**

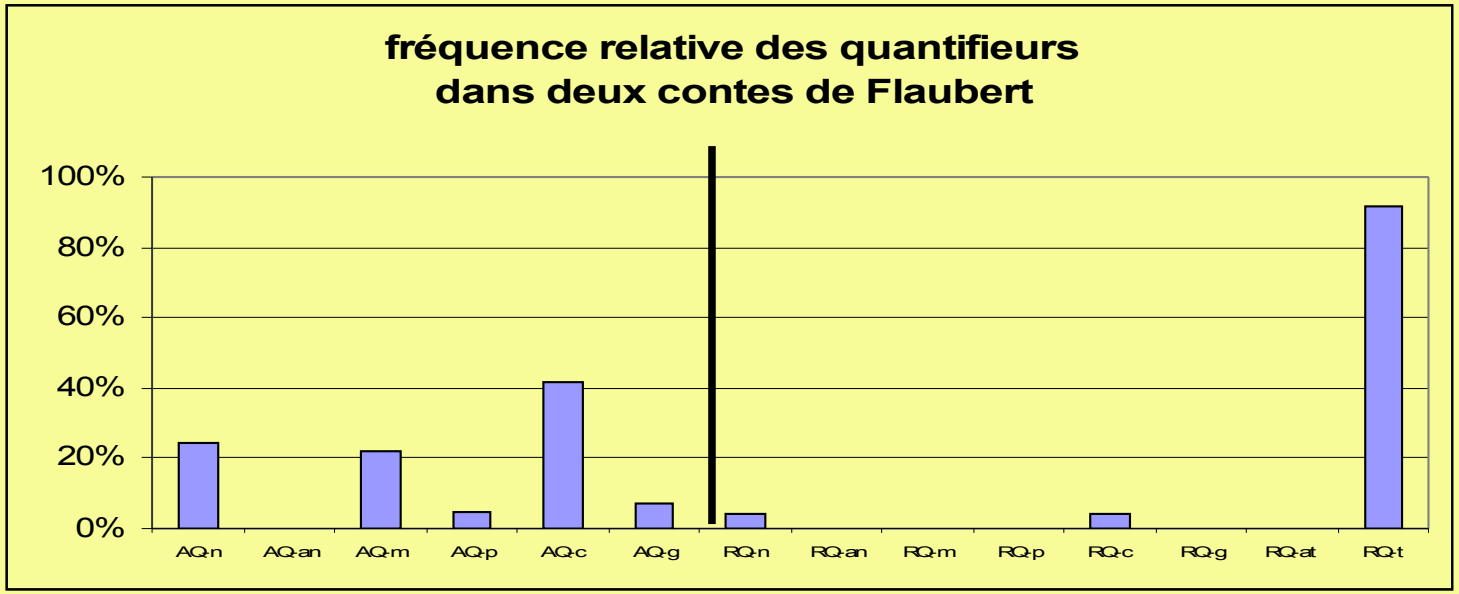
université PARIS DIDEROT PARIS 7

# Exploration de corpus littéraires: genre et style
# Premières analyses comparatives (à consolider)

## J. Austen et A. Radcliffe montrent des profils de distribution très différents:




## Et il en va de même pour Flaubert et Maupassant:




université PARIS DIDEROT PARIS 7

# Conclusion:
# encore très très beaucoup de travail
# en perspective pour le désordinateur



**Yadékestions ?**

université PARIS DIDEROT PARIS 7