Les chaînes de référence en français : analyse d'un corpus diachronique de textes narratifs (début 12e – fin 15e s.)

CÉLINE GUILLOT-BARBANCE, ENSL-IHRIM

Contexte : Projet ANR DEMOCRAT

Objectifs

- produire un modèle intégré et discursif de la référence qui rende compte de la variation (diachronique, générique, linguistique)
- produire un corpus annoté (12e – 21s.) qui serve de référence et de corpus d'apprentissage (disponible sur Ortolang fin mars)
- produire un outil d'annotation et de manipulation des données annotées
- produire un système de détection automatique des coréférences (TAL)

Equipes

- Paris, Lattice
- Strasbourg, Lilpa
- Lyon, Ihrim et Icar

Remerciements

- Matthieu Decorde, Serge Heiden, Alexei Lavrentiev, Bénédicte Pincemin (Ihrim)
- Matthieu Quignard (Icar)
- Sophie Bordes (stagiaire Lyon2)

Définitions : chaîne de référence (CR)

« suite des expressions d'un texte entre lesquelles l'interprétation construit une relation d'identité référentielle » (Corblin 1985 : 174)

Quant li rois Clodoveu et le roine Baltelt eurent tant esté emsamle qu‘il eurent .ii. enfans, si vint en volenté au roi qu'il alast en pelerinaige en le sainte tere de Jherusalem. Il cela tant chele volenté a le roine que ele s'aperchut qu'il estoit en son cuer tormentés d'aucune secree pensee qu'ele ne savoit mie (*Vie de sainte Bathilde* 1, p. 5)

Définitions : mention, CR, paire, singleton

Mention : occurrence d'une expression référentielle

CR : suite de plus de 2 mentions coréférentes

Paire : 2 mentions coréférentes

Singleton : 1 mention isolée

-> Pas de limite finale à la CR

-> Intégration de toutes les mentions coréférentes indépendamment des limites structurelles (chapitres/paragraphes, discours direct, etc.) et des catégories grammaticales (Np, etc.)

Arrière-plan théorique

- **Étude des chaînes de référence du point de vue d'une « linguistique de l'écrit » diachronique** (Combettes 2012)
  - Chaînes de référence et genres textuels en diachronie
  - Chaînes de référence et modes de structuration des textes en diachronie
  - Chaînes de référence et progression thématique en diachronie

- **Grille d'analyse pour l'étude des chaînes de référence**
  - Mesures de paramètres multiples (*Langages* 195, 2014 ; *Langue française* 195, 2017)
    - Densité référentielle, nombre de chaînes, longueur des chaînes, composition des chaînes, etc.
  - Combinaison des approches paradigmatisante et syntagmatique (Schnedecker 2017)

Méthodologie et outils

Traitement des données (environ 50 000 mots au total)

- Annotation manuelle de 5 extraits de 10 000 mots tirés de 5 textes (stagiaire)
  - Création des mentions
  - Attribution d'un référent
- Catégorisation automatique des mentions et révision manuelle des catégories
- Création automatique des chaînes de référence

Analyse des données à l'aide de scripts

Outils de traitement et d'analyse

- Plateforme d'analyse textométrique TXM (http://textometrie.ens-lyon.fr)
- Scripts groovy

Tableau 1 : Présentation du corpus

Texte	Auteur	Titre	Date	Forme	Genre	Taille des extraits
roland	anonyme	Chanson de Roland	ca 1100	vers	épique	13 059
eneas1	anonyme	Eneas	ca 1155	vers	roman	11 647
sbath1	anonyme	Vie de sainte Bathilde (Version I)	2e m. 13e s.	prose	hagiographie	11 175
jehpar	anonyme	Roman de Jean de Paris	1494	prose	roman	11 915
commyn1	Philippe de Commynes	Mémoires	ca. 1490-1505	prose	mémoires	11 268

Structure des textes du corpus

	Structure de niveau 1	Structure de niveau 2
roland	112 laisses	-
eneas1	(initiales rubriquées)	-
sbath1	(initiales rubriquées)	-
jehpar	25 titres	-
commyn1	livre 1	2 titres

Évaluation de la qualité de l'annotation

Quignard *et al*., à par.

Deux mesures de l'accord inter-annotateurs

- Délimitation des mentions
- Attribution d'un référent à une mention

Résultats

- Les scores sont globalement bons (gamma = 0.77)
- L'existence et la délimitation des mentions peuvent varier
- L'attribution du référent à la mention est plus stable
- L'accord inter-annotateurs est moins bon sur les singletons et les paires que sur les chaînes (CR)

-> L'attribution des mentions aux CR est assez fiable

Analyse paradigmatique

DONNÉS SYNTHÉTIQUES SUR LES MENTIONS, LES RÉFÉRENTS ET LES CHAÎNES

Densité référentielle par texte

	Mentions	Mentions / mots
roland	5081	39%
eneas1	4223	36%
sbath1	4202	38%
jehpar	4380	37%
commyn1	3659	32%

Référents et CR par texte

	CR	Référents des CR + paires + singletons	%age des référents dans CR
roland	127	1873	7%
eneas1	107	1635	6%
sbath1	130	1703	8%
jehpar	90	1634	5%
commyn1	148	1710	9%

Mentions et CR par texte

	Mentions	CR	Mentions dans CR	%age des mentions dans CR
roland	5081	127	3284	65%
eneas1	4223	107	2672	63%
sbath1	4202	130	2579	61%
jehpar	4380	90	2801	64%
commyn1	3659	148	2042	56%

Référents animés humains et CR

	Nombre de CR	Nombre d'animés humains
roland	127	66 (52%)
eneas1	107	51 (48%)
sbath1	130	71 (55%)
jehpar	90	53 (59%)
commyn1	148	101 (68%)

Taille des CR par texte (a)

Texte	Nombre de mentions dans la CR							
	< 10	entre 10 et 19	entre 20 et 29	entre 30 et 39	entre 40 et 49	entre 50 et 59	entre 60 et 89	> 90
Roland	76	24	7	8	3	2	1	6 1851 m.
Eneas	65	15	7	8	2	3	4	3 1173 m.
Sbtah1	91	23	6	3	-	2	1	4 1453 m.
jehpar	52	10	5	5	1	5	6	6 1387 m.
Commyn1	111	20	8	-	3	-	1	5 752 m.

Taille des CR par texte (b)

	Nombre de mentions dans la CR			
	< 10	entre 10 et 19	entre 20 et 89	> 90
Roland	76 (60%)	24 (19%)	21 (16%)	6 (5%)
Eneas	65 (61%)	15 (14%)	24 (22%)	3 (3%)
Sbath1	91 (70%)	23 (18%)	12 (9%)	4 (3%)
Jehpar	52 (58%)	10 (11%)	22 (24%)	6 (7%)
Commyn1	111 (75%)	20 (14%)	12 (8%)	5 (3%)

Bilan : grandes tendances

- Pas de corrélation entre la densité référentielle et le nombre de CR
- Corrélation inverse entre le nombre de CR courtes et le nombre de singletons/paires (biais de l'annotation ?)
- Écart très important entre le nombre de CR courtes et toutes les autres
- Les textes qui ont beaucoup de CR sont ceux qui ont beaucoup de CR courtes (commyn1, sbath1 et roland)
- Les données ne reflètent pas une évolution diachronique ni une opposition vers/prose
- Importance du genre ?

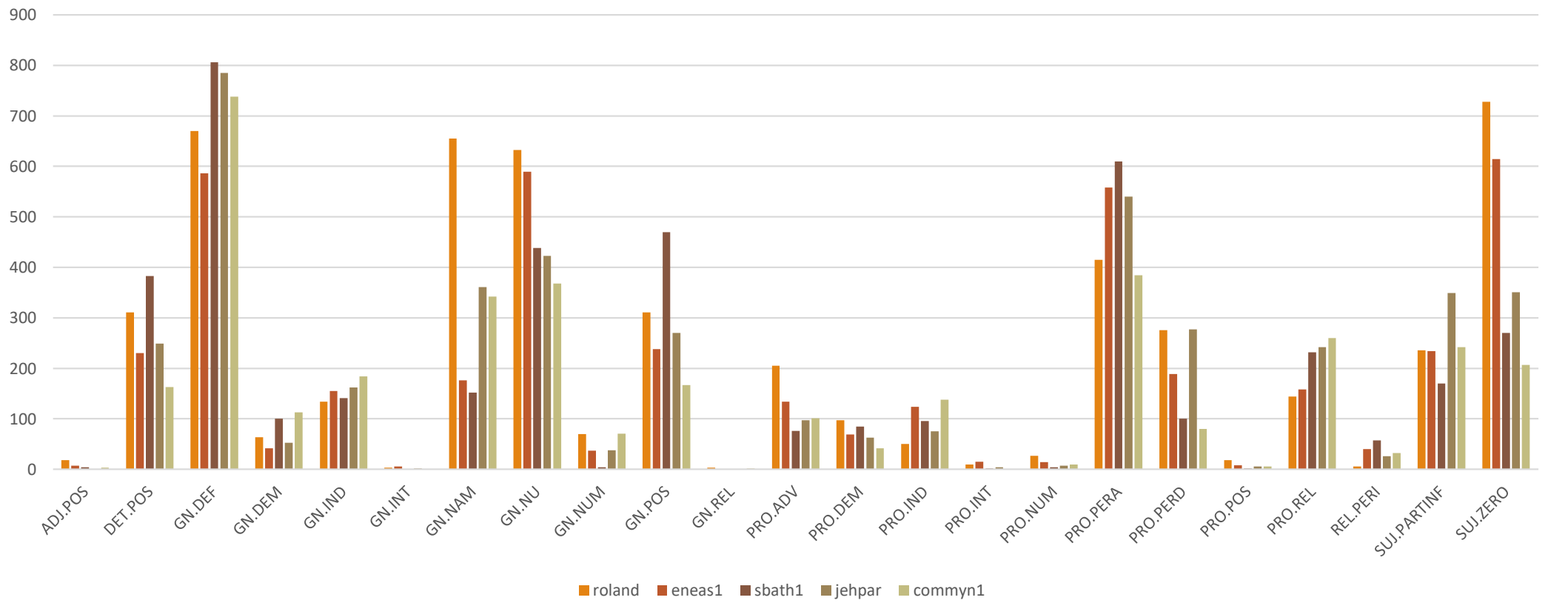
Bilan : profil de chaque texte

- **roland** : forte densité référentielle, beaucoup de CR qui totalisent une part importante des mentions, toutes les tailles de CR sont bien représentées
- **eneas1** : densité référentielle relativement faible et assez peu de CR, seulement 3 CR très longues, un nombre moyen de CR courtes et moyennes, peu de référents humains dans les CR
- **sbath1** : densité référentielle relativement importante et beaucoup de CR, surtout courtes et 4 CR très longues, nombre moyen de CR intermédiaires
- **jehpar** : peu de CR mais qui comptabilisent beaucoup de mentions, 6 référents dont les CR sont assez longues, très peu de CR courtes et un nombre moyen de CR intermédiaires, beaucoup de référents humains dans les CR
- **commyn1** : densité référentielle faible, beaucoup de CR mais globalement courtes (beaucoup de CR courtes, les 5 CR longues ne sont pas très longues, peu de CR intermédiaires), beaucoup de référents humains dans les CR

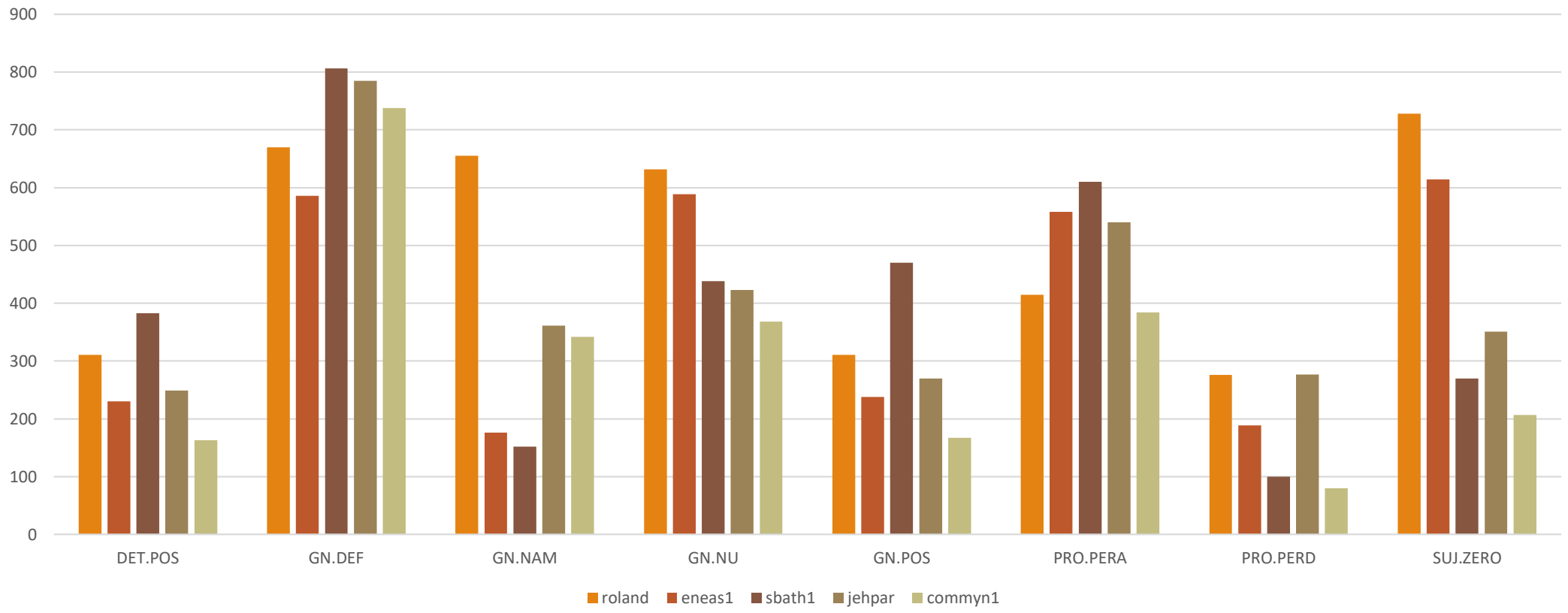
Analyse paradigmatique

COMPOSITION DES MENTIONS ET DES CHAÎNES

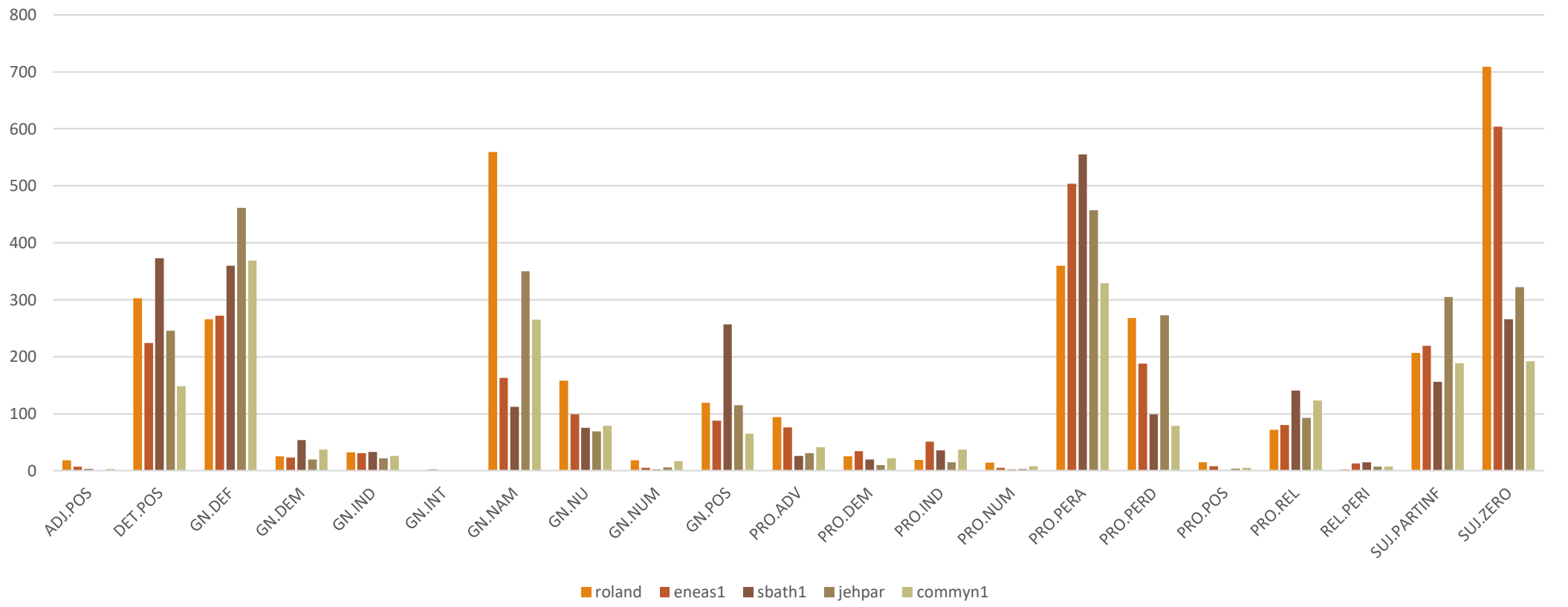
Catégories grammaticales des mentions



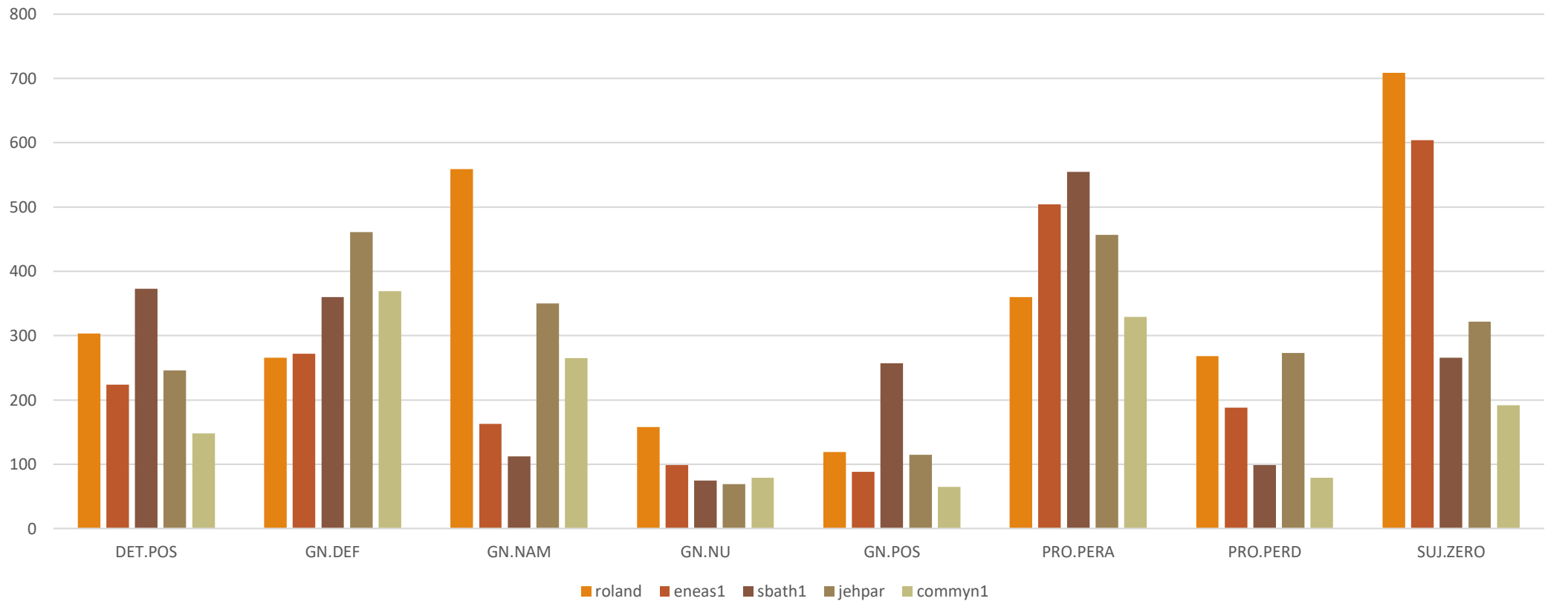
Principales catégories grammaticales des mentions



Catégories grammaticales des CR



Principales catégories grammaticales des CR



Bilan : grandes tendances

- Mentions
  - Le sujet zéro domine dans les deux textes les plus anciens (en vers)
  - Le GN défini domine dans les trois textes les plus récents (en prose)
- CR
  - sujet zéro et PP dominent dans les 3 textes anciens (roland sujet zéro, eneas1 sujet zéro et PP, sbath1 PP)
  - les expressions nominales sont plus fréquentes dans les textes du 15^{e} s. (GN défini catégorie la plus fréquente dans commyn1, fréquences du PP et du GN défini proches dans jehpar)
- Le Np est très inégalement représenté : fréquent dans Roland, relativement dans jehpar et commyn1, rare dans eneas1 et sbath1

Bilan : profil de chaque texte

- **roland** : beaucoup de catégories fréquentes, importance du sujet zéro, y compris en dehors des CR, relative rareté du PP, fréquence du Np
- **eneas1** : fréquence du sujet zéro et du PP, surtout dans les CR, peu d'expressions nominales
- **sbath1** : fréquence du PP dans les CR et du GN défini dans les mentions, rareté du Np, relative rareté du sujet zéro
- **jehpar** : fréquence du GN défini dans les mentions, fréquences assez proches du PP, GN défini, Np et sujet zéro dans les CR
- **commyn1** : fréquence des expressions nominales (GN défini un peu plus fréquent que le PP dans les CR, Np bien représenté), et recul du sujet zéro pas compensé par le PP

Analyse syntagmatique

Méthode d'exploration syntagmatique

- Création de sous-corpus de petite taille
  - Analyse partielle de chaque texte (7000 sur 10 000 mots)
  - 14 sous-corpus par texte correspondant à 14 portions de 500 mots
- Exploration de chaque portion de texte grâce à des outils de synthèse
  - Données plus fines sur les CR au fil du texte
- Exploration de chaque portion de texte grâce à l'outil « Progression »
  - Visualisation des CR à travers la linéarité du texte

Analyse syntagmatique

DONNÉES SYNTHÉTIQUES SUR LES CHAÎNES

Risques de concurrence et d'ambiguïté référentielle : nombre de CR

	Nombre de CR par portion de texte														Moyenne
roland	22	24	25	25	18	16	18	25	25	18	19	21	23	17	21
eneas1	15	19	23	19	16	13	18	24	17	16	13	14	11	14	17
sbath1	10	17	10	16	12	16	12	13	11	9	9	20	15	24	14
jehpar	17	16	16	22	14	18	14	18	19	17	20	17	19	15	17
commyn1	19	15	25	22	24	23	30	21	25	18	21	17	14	21	21

Risques de concurrence et d'ambiguïté référentielle : référents humains

	Nombre de référents animés animés humains dans les CR par portion de texte														Moyenne
roland	14	16	11	15	12	13	13	15	13	12	13	13	12	7	13
eneas1	12	12	14	9	8	4	4	12	8	9	10	9	8	10	9
sbath1	8	17	6	9	7	8	8	8	7	6	7	11	9	15	9
jehpar	12	10	10	12	8	15	10	14	14	12	12	9	11	10	11
commyn1	11	10	12	10	14	12	17	11	14	11	15	11	11	12	12

Proéminence des référents : CR longues

	Nombre de CR de plus de 10 mentions par portion de texte													
roland	3	3	2	4	5	3	4	4	3	3	4	3	3	3
eneas1	4	3	5	2	5	1	2	2	4	3	6	4	2	5
sbath1	2	1	4	3	3	3	6	4	3	4	3	3	4	2
jehpar	3	2	5	4	4	2	2	4	4	5	2	4	3	1
commyn1	3	3	3	2	1	3	2	2	1	3	0	3	2	1

Proéminence des référents : CR les plus longues

	Nombre de mentions de la CR la plus longue par portion de texte													
roland	36	25	61	46	30	63	52	30	54	43	34	39	34	41
eneas1	41	41	27	22	20	21	13	22	51	45	26	20	42	27
sbath1	49	49	44	38	58	42	38	45	39	40	50	24	38	17
jehpar	24	46	22	29	26	34	51	46	24	22	36	33	64	44
commyn1	24	37	26	27	15	11	12	24	24	18	9	12	18	22

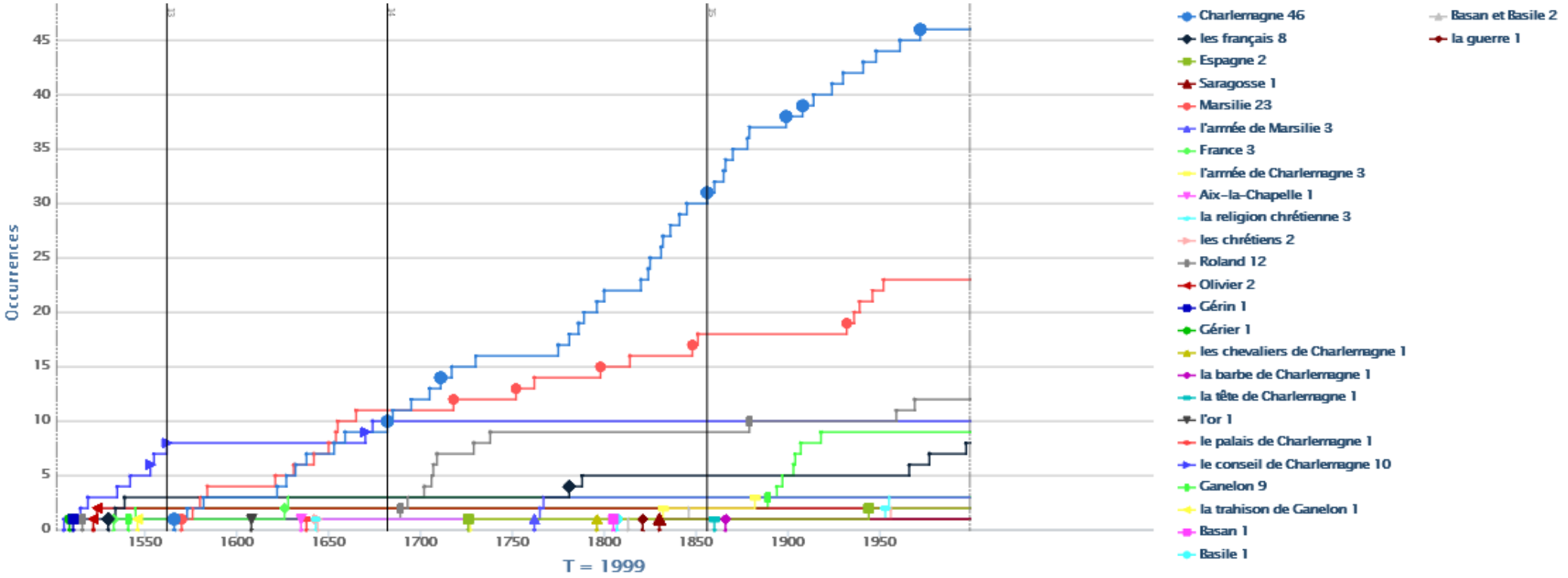
Bilan : profil des textes

- **roland** : beaucoup de référents humains instanciés, nombre de CR longues assez stable et en général au moins une CR très longue, 3 ou 4 référents proéminents et souvent 1 très proéminent
- **eneas1** : peu de référents humains instanciés, CR longues variables en nombre et en longueur, plusieurs référents relativement proéminents et parfois un référent très proéminent
- **sbath1** : peu de référents humains instanciés, CR longues très longues et en nombre variable, à peu près toujours un référent très proéminent et parfois plusieurs qui le sont relativement
- **jehpar** : beaucoup de référents humains instanciés, CR longues variables en nombre et en longueur, plusieurs référents proéminents et parfois un référent très proéminent
- **commyn1** : beaucoup de référents et de référents humains, mais peu qui sont assez ou très proéminents, moins de CR longues (entre 2 et 3) et moins longues que dans les autres textes

Analyse syntagmatique

VISUALISATION DE LA PROGRESSION DES CHAÎNES

Roland : exemple de progression



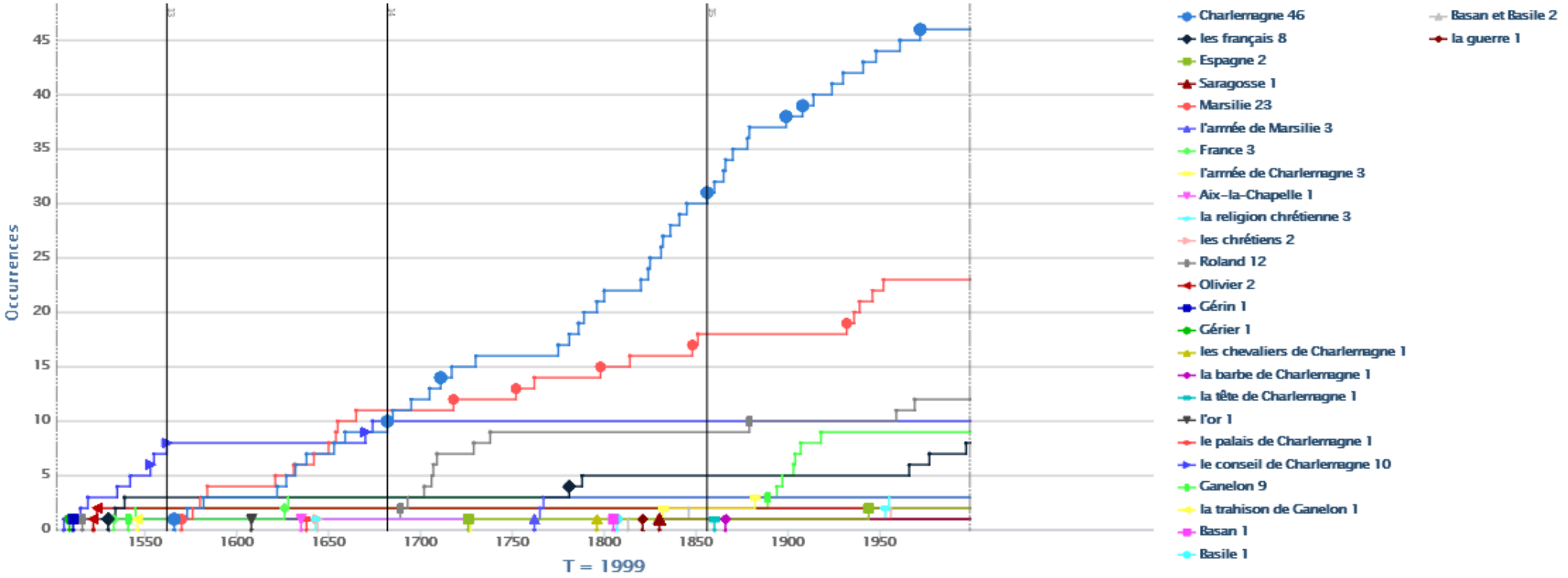
Topique (dis)continu (Givón 1983)

« la chaîne de référence d'un topique « continu » devrait être majoritairement composée de pronoms personnels, sur une distance « longue » (*i.e.* de 10 phrases consécutives, au moins selon T. Givón) alors que celles d'un topique « discontinu » du fait qu'elles entrent en compétition avec d'autres chaînes, seraient plus courtes et instancieraient davantage des SN que les chaînes de topiques continus » (Schnedecker & Landragin 2017 : 12)

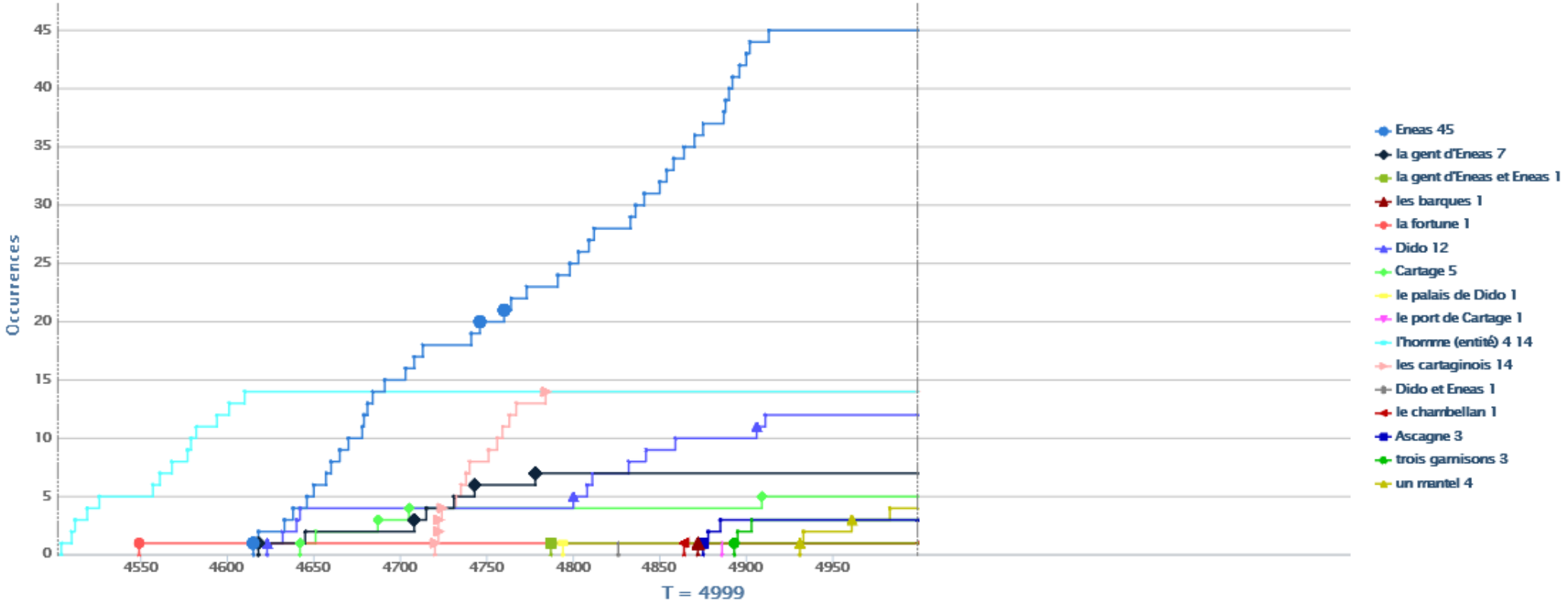
Progression à thème constant de l'ancien français (Combettes 2012)

« c'est la progression à thème constant, caractéristique de la narration dans la prose narrative en ancien français, qui assure, si l'on peut dire, la dépendance du second plan et sa faible importance, ne serait-ce que du point de vue quantitatif. **Ce type de progression permet en effet à un référent saillant, d'ordinaire "personnage principal", d'apparaître comme thème et comme sujet syntaxique dans des énoncés successifs, et de servir ainsi de relais dans le maintien de la cohérence du passage.** C'est par l'intermédiaire de cet actant, par l'emploi de procès de perception ou d'action que vont se trouver introduits les référents nouveaux, supports éventuels d'une description, d'un commentaire. **Le syntagme sujet n'étant pas obligatoirement exprimé, la première place de la proposition se trouve d'ordinaire remplie par un marqueur qui renforce le lien contextuel (si, or, lors, ...) et accentue l'impression d'une cohérence "étroite", resserrée.** » (Combettes 2012 : 7)

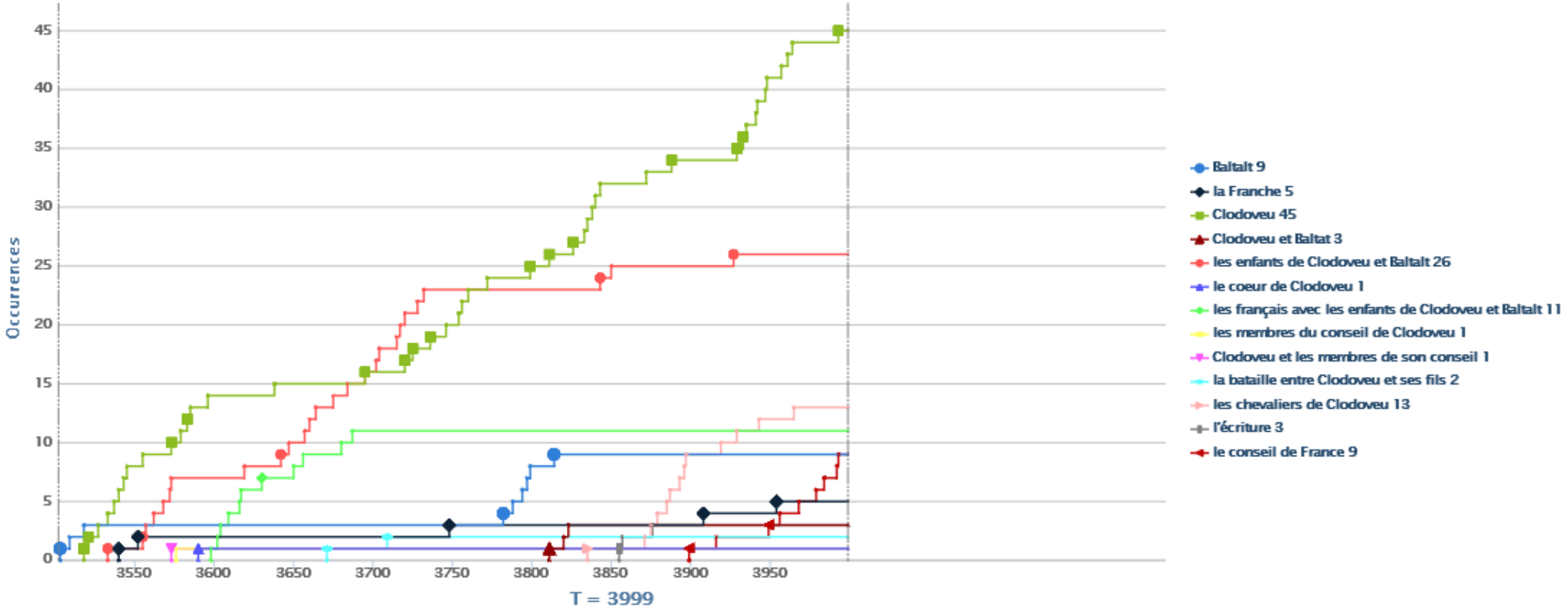
Roland : exemple de progression



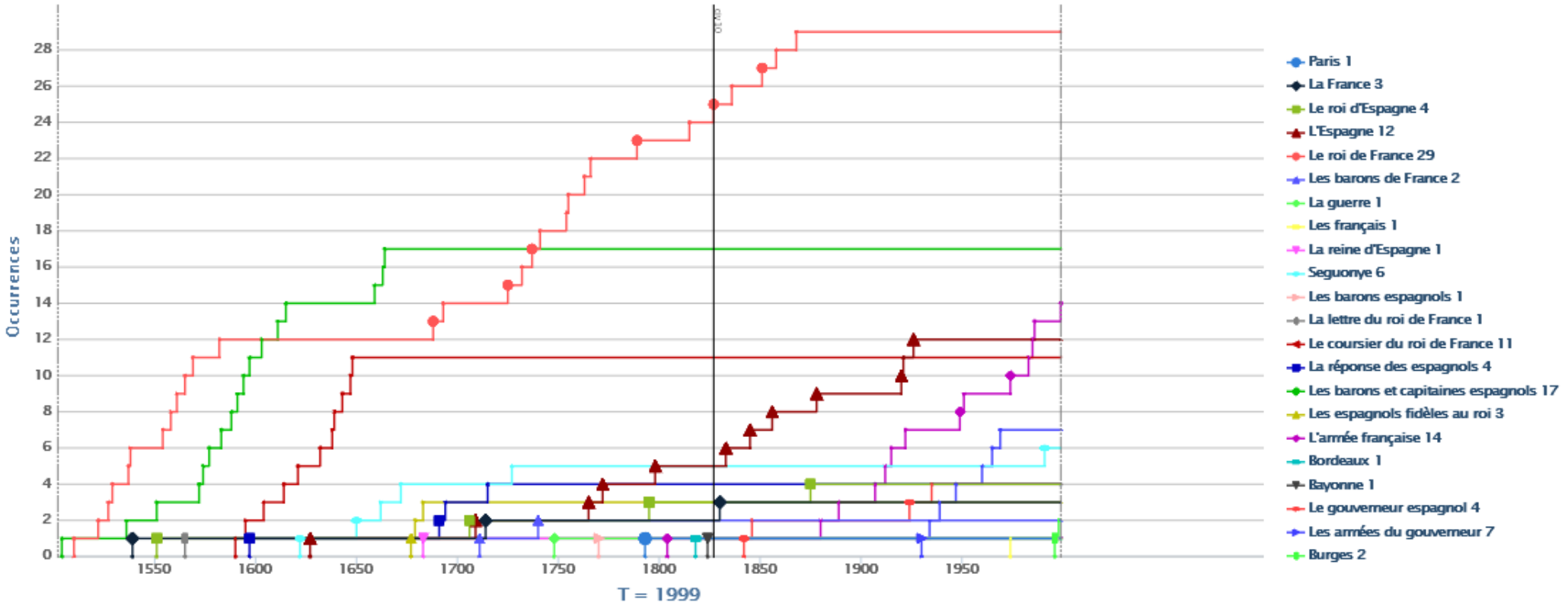
Eneas1 : exemple de progression



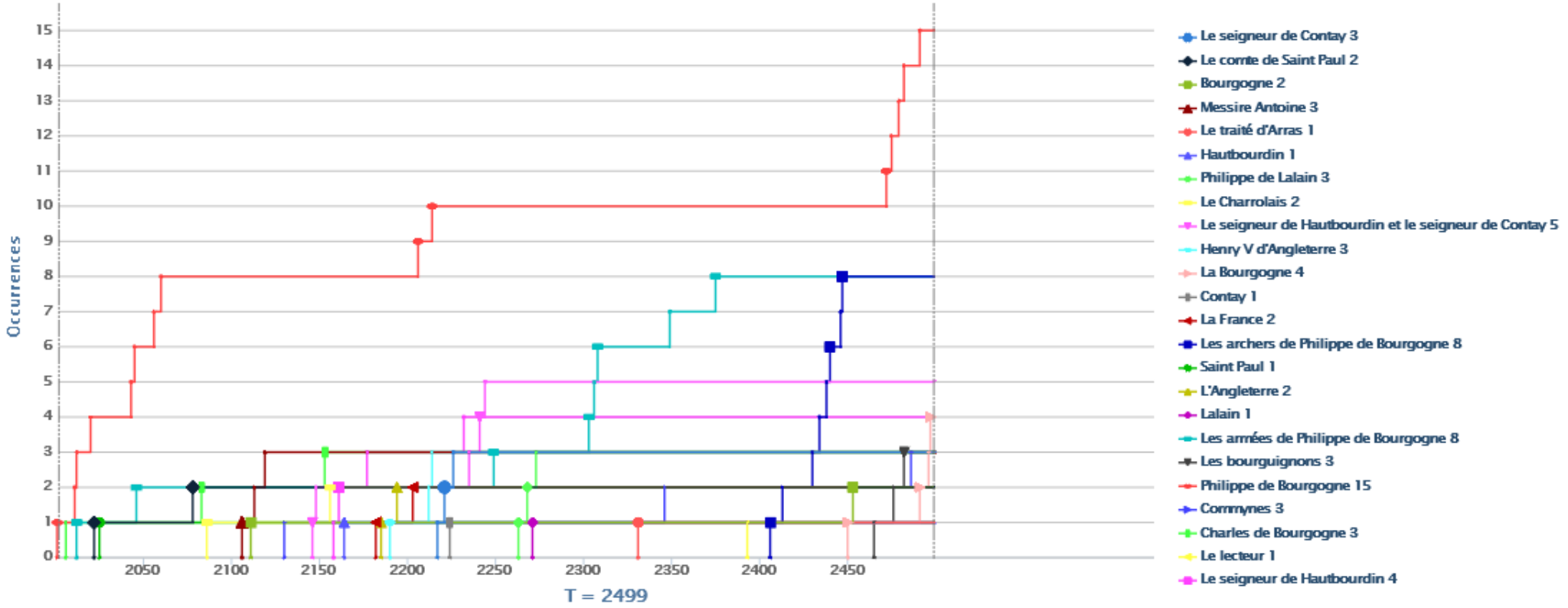
Sbtah1 : exemple de progression



Jehpar : exemple de progression



Commyn1 : exemple de progression



Progression des CR : profil des textes

- **roland** : quelques topiques continus de longue durée qui alternent avec des CR de densité moyenne ou forte qui s'interrompent et redémarrent assez souvent, beaucoup de référents humains concurrents
- **eneas1** : pas mal de topiques continus, grande densité, durée variable et parfois longue, on voit assez bien les topiques continus se succéder, assez peu de redémarrages de CR, peu de référents humains concurrents
- **sbath1** : beaucoup de topiques continus, longue durée, grande densité, peu de référents humains concurrents, peu de redémarrages de CR
- **jehpar** : pas mal de topiques continus, relative densité des CR, durée variable, on voit assez bien les topiques continus se succéder, pals mal de redémarrages de CR, beaucoup de référents humains concurrents
- **commyn1** : très peu de topiques continus et sur une courte durée, beaucoup de référents humains concurrents et dont les CR sont courtes

Roland : GN et topiques multiples

274 « Francs chevalers, dist li emperere Carles,

275 Car m'eslisez un barun de ma marche

276 Qu'a Marsiliun me portast mun message. »

277 Ço dist Rollant : « Ço ert Guenes, mis parastre. »

278 Dient Franceis : « Car il le poet ben faire ;

279 Se lui lessez, n'i trametrez plus saive. »

Roland : GN et changement thématique

485 Marsilies fut esculurez de l'ire,

486 Freint le seel, getet en ad la cire.

487 Guardet al bref, vit la raisun escrite :

488 « Carle me mandet, ki France ad en baillie,

489 Que me remembre de la dolur e l'ire,

490 Ço est de Basan e de sun frere Basilie

491 Dunt pris les chefs as puis de Haltoïe ;

492 Se de mun cors voeil aquiter la vie,

493 Dunc li envei mun uncle l'algalife ;

494 Altrement ne m'amerat il mie. »

495 Aprés parlat ses filz envers Marsilies,

496 E dist al rei : « Guenes ad dit folie.

Commynes : GN et topiques multiples

Encores disoit ledict Morvillier qu'il ne povoit penser qui avoit meu ledict conte de prendre ceste alliance avecques ledict duc de Bretaigne, sinon une pension que le roy luy avoit donné avec le gouvernement de Normandie que le roy luy avoit osté. Le lendemain, à l'assemblée et en la compagnie des dessusdicts, ledict conte de Charroloys, le genouil en terre sur ung carreau de veloux, parla à son père premier et commença de ce bastard de Rubempré, disant les causes estre justes et raisonnables de sa prinse et qu'il se monstreroit par procès.

Sbath1 : GN et progression thématique

Ne demora gaires aprés ches coses que li rois par le conseil de le sainte roine se feme fist assamler tous les rinches et tous les barons de se terre pour atorner en quel maniere li roiaumes de France seroit gardés et gouvernés tant qu'il fust repairiés de cel saint pelerinaige. Li baron qui resgarderent le frailleté d'umainne nature et qui douterent mout que leurs sires ne repairast jamais, vinrrent tout emsamle au roi et li prierent mout douchement qu'il coronnast son ainsné fil a roi anchois qu'il se meust pour le terre garder au conseil de le sainte roine se mere. Li rois Clodoveu, quant il vit que tuit s'acorderent a chou, s'en fist toute leur priere et puis ne demora gaires qu'il se mist a le voie et emprist le haut pelerinaige qu'il avoit tant desirré. Quant le roine Baltelt eut son seigneur convoié et eut pris congié a li et ele em plorant eut commandé s'ame et sen cors en le garde Nostre Seignor, si retorna entre li et ses enfans et demora el regne ou ele abandonna son cuer et son cors a geuner et a orisons vers Nostre Seigneur por le pais del regne et pour l'estat de sainte eglise et pour le salu de son seignor et de ses enfans.

Li jones rois ses fix, de qui li nons est mis en oubli, regna ainssi grant pieche…

Conclusion

Deux grands ensembles de textes

- ceux qui ont beaucoup de référents concurrents nombreux et peu continus (commyn1) ou qui ont beaucoup de topiques +/- continus avec des interruptions et redémarrages (roland, jehpar) et qui ont le plus d'expressions nominales ; ce sont aussi les textes qui ont des divisions internes plus explicites
- ceux qui ont beaucoup de topiques continus qui durent (sbath1) ou se succèdent (eneas1), avec une densité forte, peu de redémarrages et peu de référents concurrents et dans lesquels la dénomination des personnages semble permettre l'identification du thème autour duquel s'organise un segment de discours de taille assez réduite

Intérêt d'une approche discursive de la référence

- Nécessité de prendre en compte des facteurs multiples
- Intérêt des corpus annotés
- Création de méthodologies et d'outils adaptés