

```
POS_E_V_count
POS_E_C_count <- POS_E_
  rename(n.c
POS_E_ <- full_join(POS_E_V_count,
  POS_E_C_count,
  by = "id")
POS_E_ <- replace(POS_E_, is.na(POS_E_),
  POS_E_ <- POS_E_ %>% mutate(Nb_POS_E = 1)
  mutate(Tx_E = 1)
  mutate(Tx_C = 1)
  ## Reinject metadata
  POS_E_ <- merge(POS_E_,
```

Comparalem

Outil pour une phonétique
historique de corpus

Timothée Premat

Congrès DIACHRO XI | Madrid | 24/05/2024

Plan

1. Introduction

2. Comparalem

3. La variation -e/Ø des adverbes

4. Conclusion

Diapositives, script et données :

<https://phonodiachro.hypotheses.org/category/actualite>
(Moteur de recherche → « Timothée Premat » → Actualité)

Introduction

Introduction

La phonétique historique

- ▶ Discipline « sinistrée » (SÉGÉRAL et SCHEER 2015)
- ▶ N'a pas su s'approprier les outils numériques (id.)
 - ▶ Cf. chap. « Phonétique Historique » (SÉGÉRAL et SCHEER) dans la GGHF (MARCHELLO-NIZIA et al. 2020)

Introduction

La phonétique historique et les heuristiques numériques

- ▶ Quelques contre-exemples de phonétique historique de corpus :
 1. MARCHELLO-NIZIA (2015), RAINSFORD (2020)
 2. RAINSFORD (2010, 2011a,b)
 3. RAINSFORD et SCRIVNER (2014), RAINSFORD (2022)
- ▶ Méthode :
 1. Se restreindre à un lemme, ou à quelques textes
 2. Développer des outils d'annotation et d'analyse
 3. Développer un corpus enrichi en information grapho-phonologiques et métriques

Introduction

La phonétique historique et les heuristiques numériques

- ▶ Quelques contre-exemples de phonétique historique de corpus :
 1. MARCHELLO-NIZIA (2015), RAINSFORD (2020)
 2. RAINSFORD (2010, 2011a,b)
 3. RAINSFORD et SCRIVNER (2014), RAINSFORD (2022)
- ▶ Méthode :
 1. Se restreindre à un lemme, ou à quelques textes
 2. **Développer des outils** (d'annotation et) **d'analyse** ← **Méthode choisie**
 3. Développer un corpus enrichi en information grapho-phonologiques et métriques

Introduction

La phonétique historique et les heuristiques numériques

- ▶ Développer des outils pour l'analyse grapho-phonologique
- ▶ Comparalem : programme développé pour ma thèse
 - ▶ *La genèse de l'éisión* (PREMAT 2023)

Comparalem

Comparalem

Présentation

- ▶ Script R
- ▶ Qui permet :
 1. la détection des lemmes sujets à une variation graphique donnée
 2. la projection de cette variation dans le temps et dans l'espace
 3. des tests statistiques
- ▶ Objectif :
 - ▶ Étudier des variations grapho-phonologiques dans de grands corpus annotés

Comparalem

Fonctionnement général

- ▶ Le programme fonctionne en deux exécutions :
 1. Détection des lemmes sujets à la variation
 2. Analyse des taux de variantes pour chaque texte

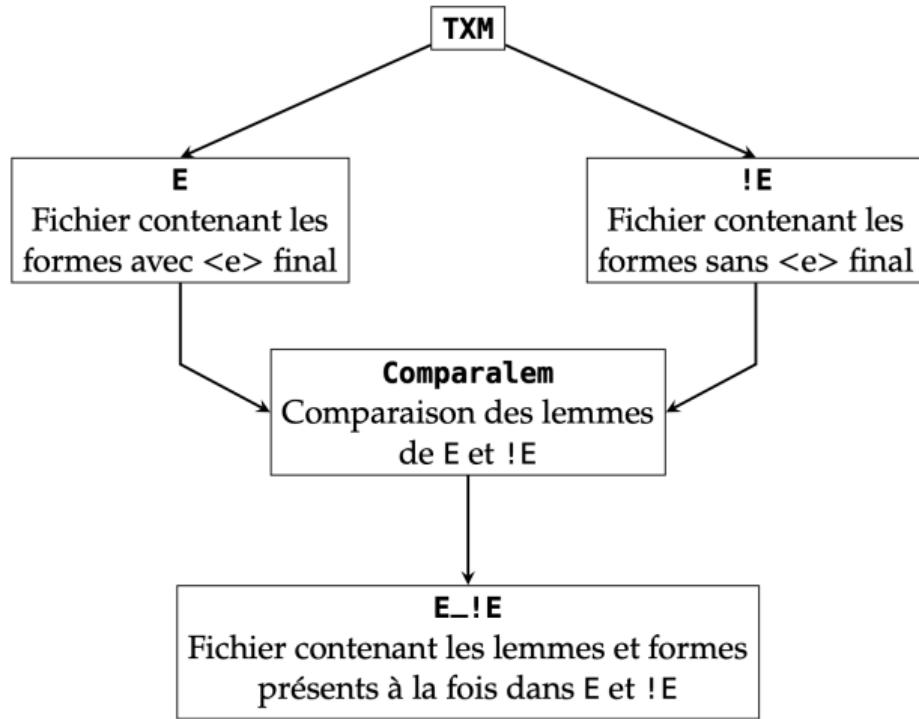


Figure 1 – Fonctionnement de Comparalem : étape 1.

(Exemple de l'analyse d'une variation -e/Ø)

Comparalem

Fonctionnement (1.b)

- ▶ Comparalem produit un fichier tabulaire avec tous les lemmes répondant à la variation graphique en question
- ▶ L'utilisateur valide chaque lemme dans le fichier tabulaire
- ▶ Il peut également refuser des occ. dans un autre fichier tabulaire
- ▶ Grâce à cela, le programme a ce qu'il lui faut pour la deuxième étape

Comparalem

Fonctionnement (2.a)

Deux possibilités :

- ▶ Analyse non contextuelle
 - les 2 fichiers d'entrée suffisent
- ▶ Analyse contextuelle
 - il faut 4 fichiers, ici :
 - 1 -e devant #C
 - 2 -e devant #V
 - 3 Ø devant #C
 - 4 Ø devant #V

Comparalem

Fonctionnement (2.a)

Deux possibilités :

- ▶ Analyse non contextuelle
 - les 2 fichiers d'entrée suffisent
- ▶ Analyse contextuelle
 - il faut 4 fichiers, ici :
 - 1 -e devant #C
 - 2 -e devant #V
 - 3 Ø devant #C
 - 4 Ø devant #V
 - ▶ Approche nécessaire pour le sandhi externe
 - ▶ Développée dans ma thèse (PREMAT 2023)
 - ▶ 'But ultime' Comparalem

Comparalem

Fonctionnement (2.b)

Le programme produit :

- ▶ Des graphiques de série temporelle (diachronie)
- ▶ Des cartes de la variation
- ▶ Des graphiques et statistiques pour mesurer l'effet du contexte
- ▶ Et d'autres facteurs (p.ex. a.n. ~ a.fr. cont.)

La variation -e/Ø des adverbes

La variation -e/Ø des adverbes

Problématique des adverbes

Les adverbes

- ▶ Les adverbes ont une variation -e/Ø
 - ▶ P.ex. *encor* ~ *encore*; *desor* ~ *desore*
 - ▶ Qui n'est pas celle des autres POS (PREMAT 2023)

La variation -e/Ø des adverbes

Problématique des adverbes

Les adverbes

- ▶ Les adverbes ont une variation -e/Ø
 - ▶ P.ex. *encor* ~ *encore*; *desor* ~ *desore*
 - ▶ Qui n'est pas celle des autres POS (PREMAT 2023)
- ▶ Et une variation -s/Ø (suff. '-s adverbial')
 - ▶ *encore* ~ *encores*; *onk* ~ *onques*
 - Ici, je vais ignorer les formes en -s

La variation -e/Ø des adverbes

Deux types d'adverbes à variation

- ▶ La variation des adverbes est connue, elle se maintient jusqu'aux XVI/XVII^e s.
 - ▶ PALSGRAVE (1530), SYLVIUS [DU BOIS] (1531), MEIGRET (1550), ESTIENNE (1569 [1557]), RONSARD (1566 [1565]), RAMÉE (1572), DEIMIER (1610), MALHERBE ([1862]) et VAUGELAS (1880 [1647])

La variation -e/Ø des adverbes

Deux types d'adverbes à variation

- ▶ La variation des adverbes est connue, elle se maintient jusqu'aux XVI/XVII^e s.
 - ▶ PALSGRAVE (1530), SYLVIUS [DU BOIS] (1531), MEIGRET (1550), ESTIENNE (1569 [1557]), RONSARD (1866 [1565]), RAMÉE (1572), DEIMIER (1610), MALHERBE ([1862]) et VAUGELAS (1880 [1647])
- ▶ Hypothèse traditionnelle : les formes sans -e étaient, à l'origine, pré-vocaliques (POPE 1966 [1934]; FOUCHÉ 1969 [1958]).

La variation -e/Ø des adverbes

Deux types d'adverbes à variation

- ▶ La variation des adverbes est connue, elle se maintient jusqu'aux XVI/XVII^e s.
 - ▶ PALSGRAVE (1530), SYLVIUS [DU BOIS] (1531), MEIGRET (1550), ESTIENNE (1569 [1557]), RONSARD (1866 [1565]), RAMÉE (1572), DEIMIER (1610), MALHERBE ([1862]) et VAUGELAS (1880 [1647])
- ▶ Hypothèse traditionnelle : les formes sans -e étaient, à l'origine, pré-vocaliques (POPE 1966 [1934]; FOUCHÉ 1969 [1958]).
- ▶ Mais MEYER-LÜBKE (1890) dit que c'est faux, y compris dans les premiers textes
- ▶ On peut le tester avec Comparalem! → analyse en fonction du contexte

La variation -e/Ø des adverbes

Deux types d'adverbes à variation

1. Ceux où le -e est de droit

- ▶ *ariere, deriere, encore, ensemble*, etc.

2. Ceux où le -e est analogique d'USQUE ou UNQUAM

- ▶ *adonque, jusque, presque, tresque*, etc.
- ▶ Il faut les traiter séparément
- ▶ Ici je ne traite que de ceux dont le -e est de droit

La variation -e/Ø des adverbes

Adverbes à -e de droit

- ▶ Corpus : NCA4 (STEIN, KUNSTMANN et GLESGEN 2006 ← DEES 1987)
- ▶ Lemmes utilisés : RNN tagger
- ▶ 19 867 occ. de 11 lemmes
 - ▶ 14 665 #C
 - ▶ 5202 #V
- ▶ 56% de -Ø
- ▶ ≠ données du résumé

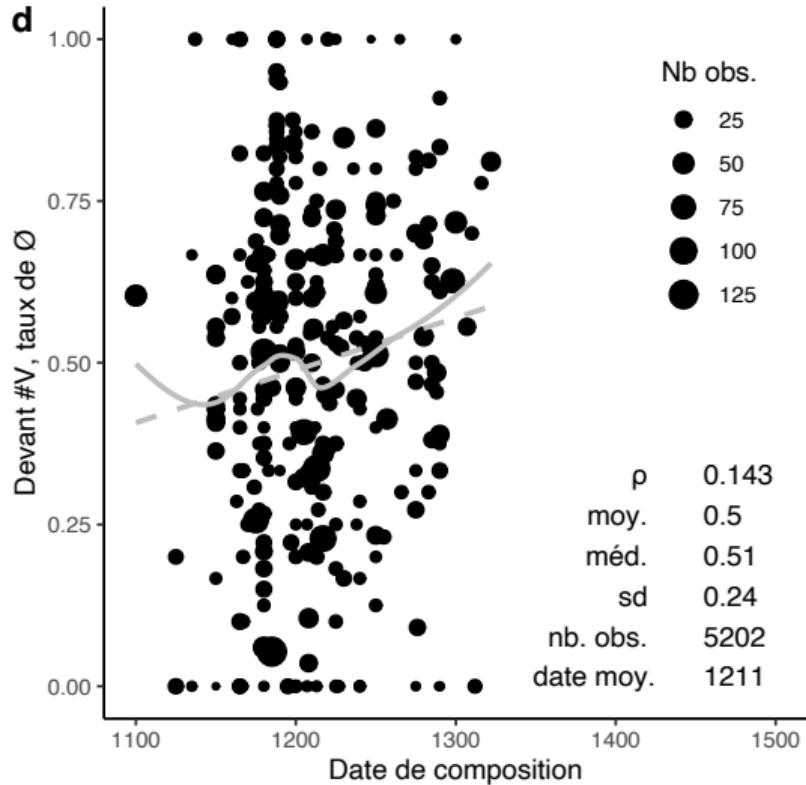
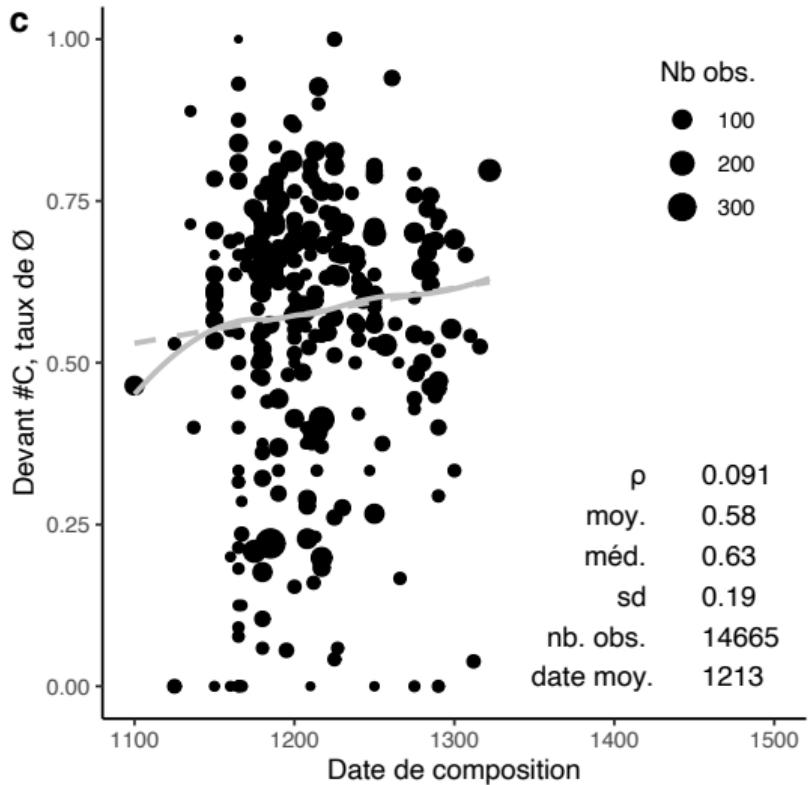
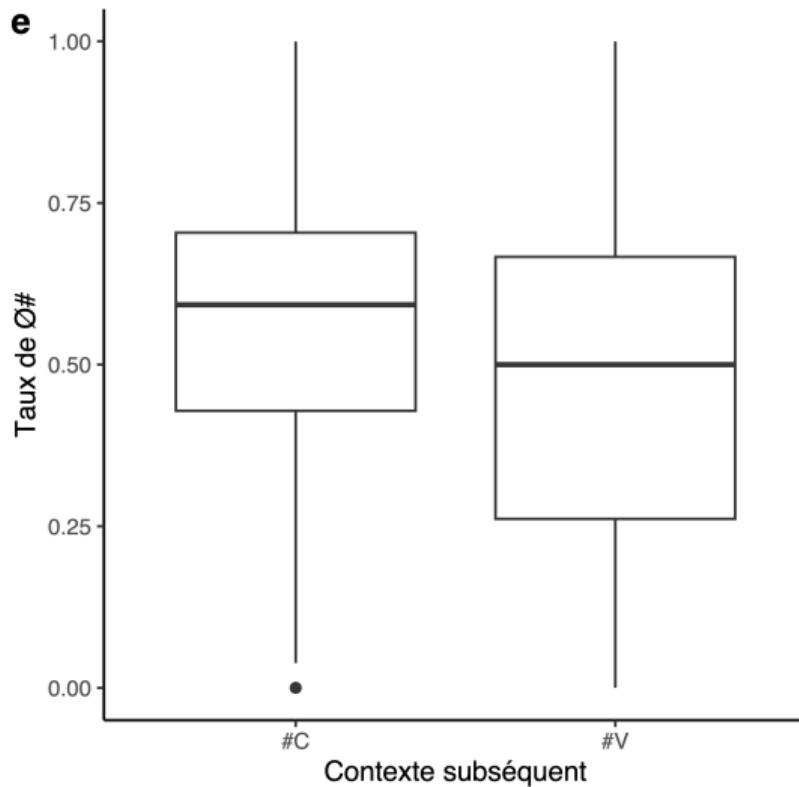


Figure 2 – Taux de Ø des ADV à -e de droit, *rnnlemma*, NCA.
 Diachronie contextuelle.



f

Test de Student

Indép. : $t = -3.16$; $p < 0.005$

Dép. : $t = -4.77$; $p < 0.005$

Figure 3 – Taux de Ø des ADV à -e de droit, *rnnlemma*, NCA.
Comparaison des deux contextes subséquents.

La variation -e/Ø des adverbes

Adverbes à -e de droit

- ▶ 56% de -Ø
- ▶ Plus de Ø devant #C que devant #V → ce n'est pas une élision
- ▶ La différence entre les deux contextes est significative mais faible
- ▶ Croissance globale des taux de Ø

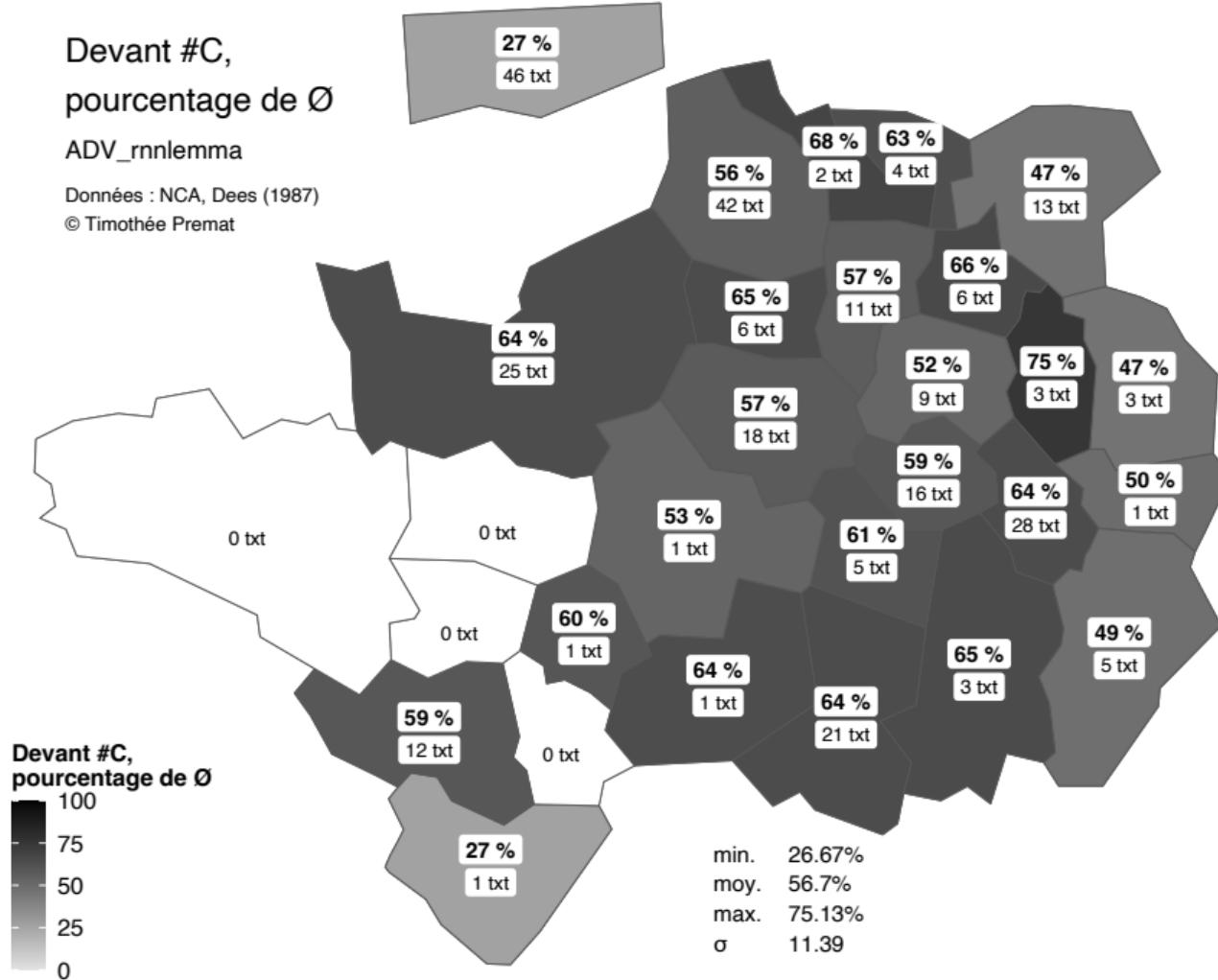
- ▶ Et en diatopie?

Devant #C,
pourcentage de Ø

ADV_rnnlemma

Données : NCA, Dees (1987)

© Timothée Premat

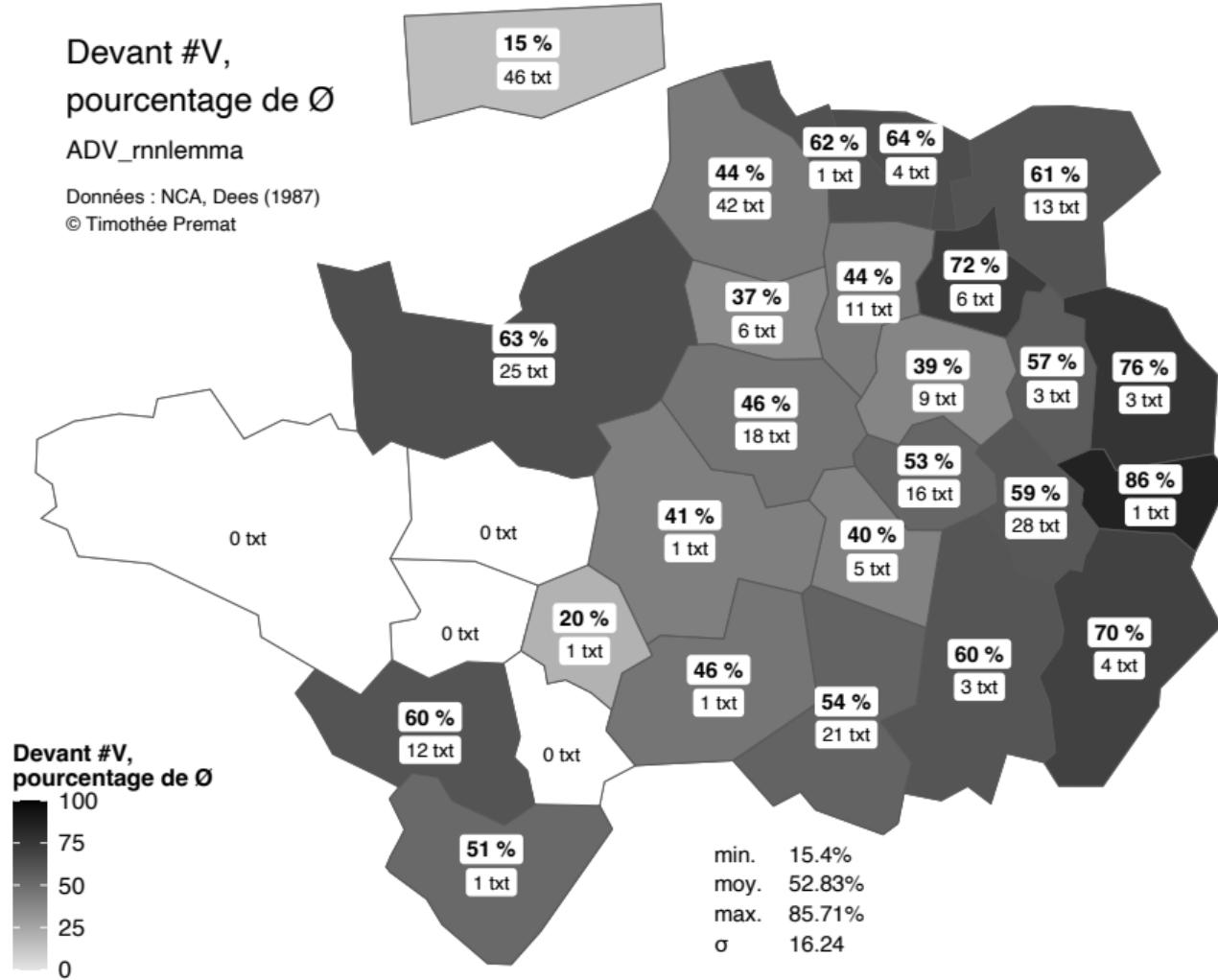


Devant #V,
pourcentage de Ø

ADV_rnlemma

Données : NCA, Dees (1987)

© Timothée Premat

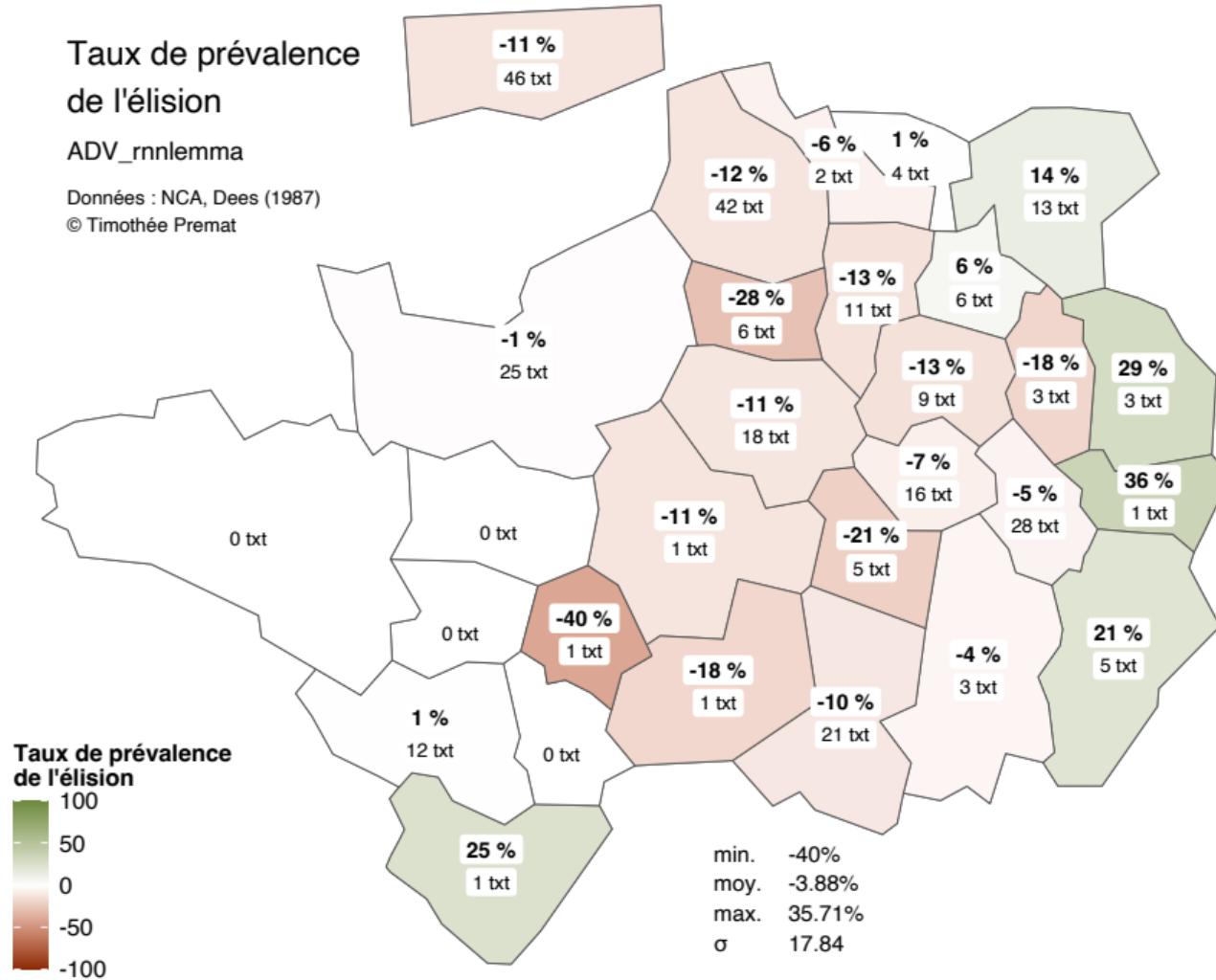


Taux de prévalence de l'élision

ADV_rnnlemma

Données : NCA, Dees (1987)

© Timothée Premat



rnlemma	nb.Ø	nb.E	frq.	tx.Ø
lors	3112	3	3115	>99%
mar	315	10	325	97%
onques	465	59	524	89%
or	5465	1534	6999	78%
dessure	435	146	581	75%
encore	1037	761	1798	58%
arrière	173	782	955	18%
derrière	28	182	210	13%
guerre	2	15	17	12%
ensemble	18	1364	1382	1%
mie	17	3944	3961	<1%

rnnlemma	nb.Ø	nb.E	frq.	tx.Ø	
lors	3112	3	3115	>99%	Apocope lexicalisée
mar	315	10	325	97%	
onques	465	59	524	89%	
or	5465	1534	6999	78%	← Var. connecteur/temp.
dessure	435	146	581	75%	
encore	1037	761	1798	58%	
arrière	173	782	955	18%	
derrière	28	182	210	13%	
guerre	2	15	17	12%	
ensemble	18	1364	1382	1%	Apocope/élision marginale
mie	17	3944	3961	<1%	

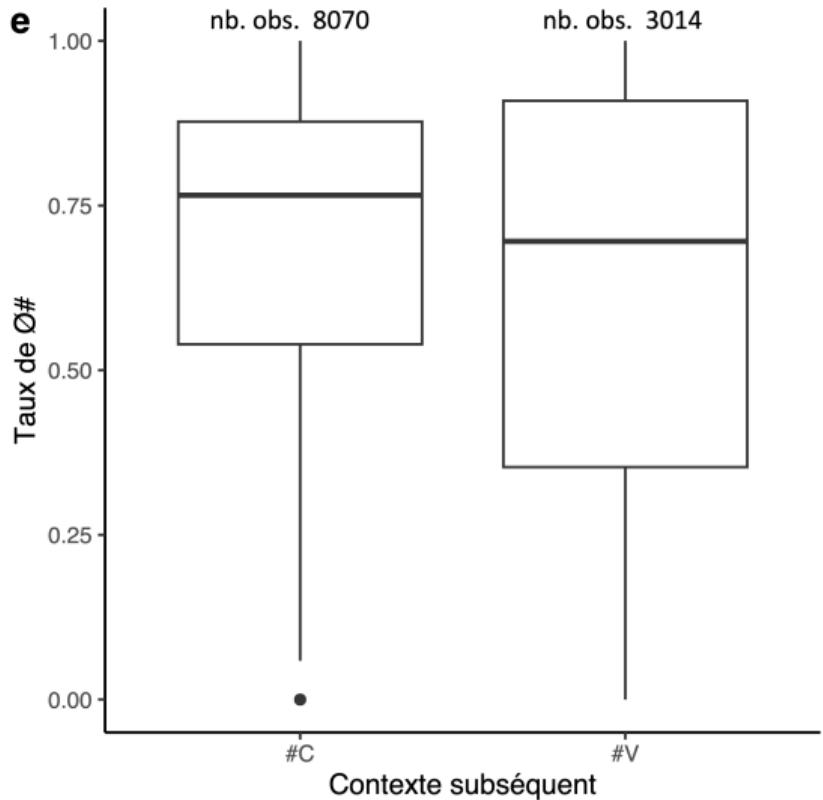


Figure 4 – Var. non marginale. Taux de Ø.
 $t = -2,58; p < 0,05$

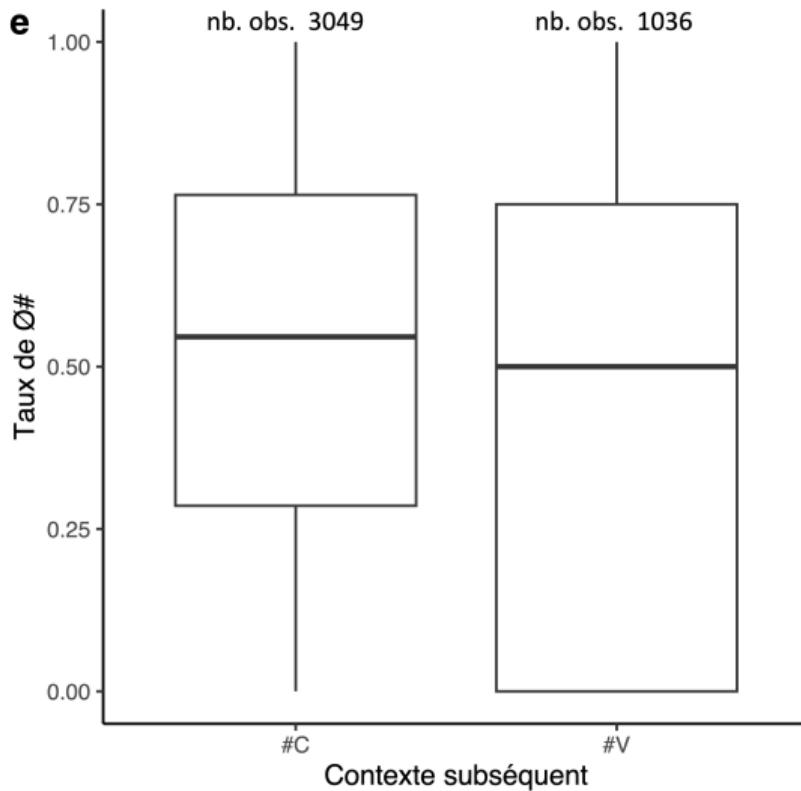


Figure 5 – Var. non marginale ($\neq ore$). Taux de Ø.
 $t = -2,41; p < 0,05$

La variation -e/Ø des adverbes

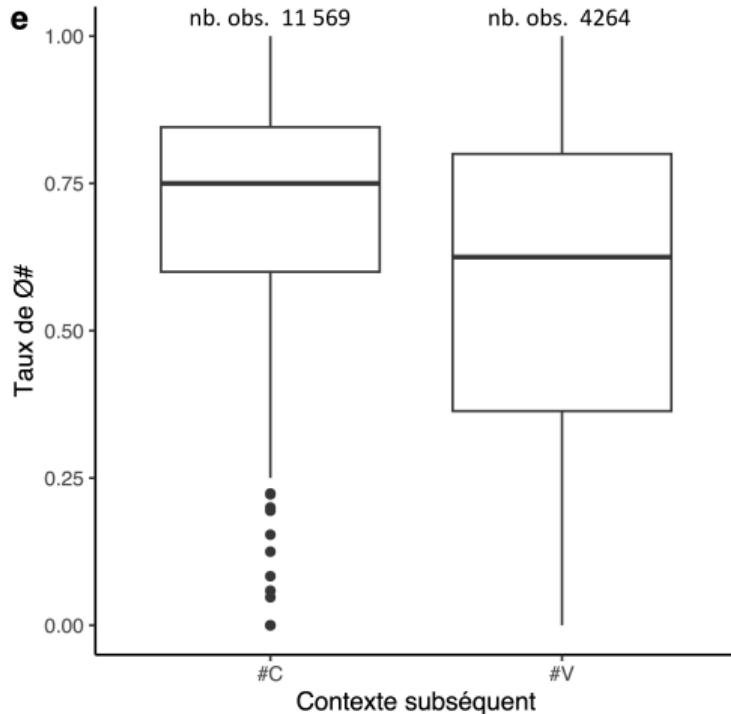
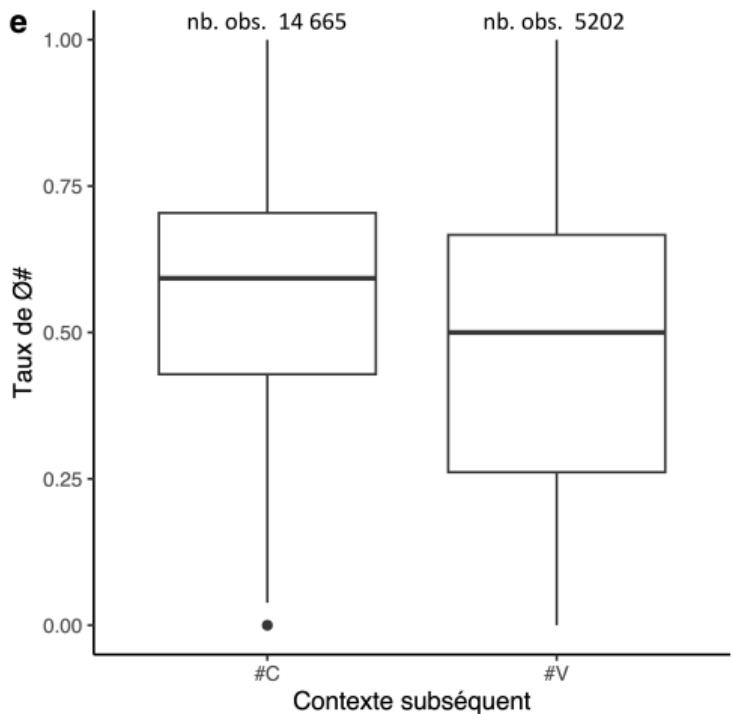
Adverbes à -e de droit

- ▶ Les taux varient
- ▶ Mais #C est toujours plus associé à Ø que #V
- ▶ Ce n'est pas une élision
- ▶ Cela contredit l'hypothèse traditionnelle

La variation -e/Ø des adverbes

Adverbes à -e de droit

- ▶ Les taux varient
- ▶ Mais #C est toujours plus associé à Ø que #V
- ▶ Ce n'est pas une élision
- ▶ Cela contredit l'hypothèse traditionnelle
- ▶ Est-ce que d'autres facteurs peuvent jouer?
 1. Lemmatisation : RNN vs TreeTagger
 2. Qualité de l'édition
 3. Prose/vers



La variation -e/Ø des adverbes

Adverbes à -e de droit

- ▶ Pas d'effet du lemmatiseur : les taux sont différents mais la situation demeure la même ($t < 0$)
- ▶ Effet de la qualité de l'édition ?
 - ▶ Quotation de la qualité de l'édition dans les métadonnées :
 - ▶ Diplomatique : ms1 > ms2 > ms3
 - ▶ Critique : cr1 > cr2 > cr3
 - ▶ GLESSGEN et GOUVERT (2007) : utiliser ms1-2 et cr1 pour la grapho-phonologie

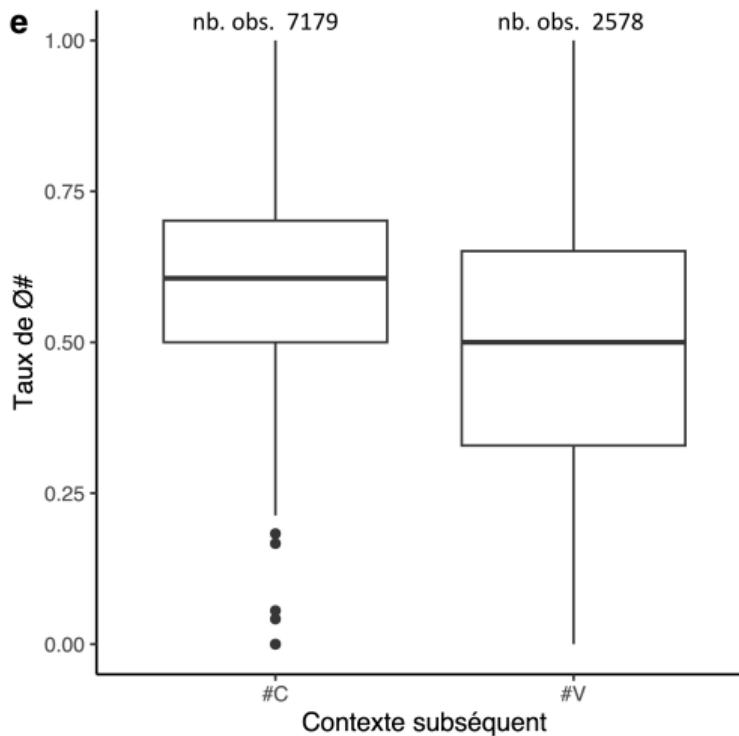


Figure 8 – ms1-2 & cr1. Taux de Ø.
 $t = -2.66; p < 0.05$

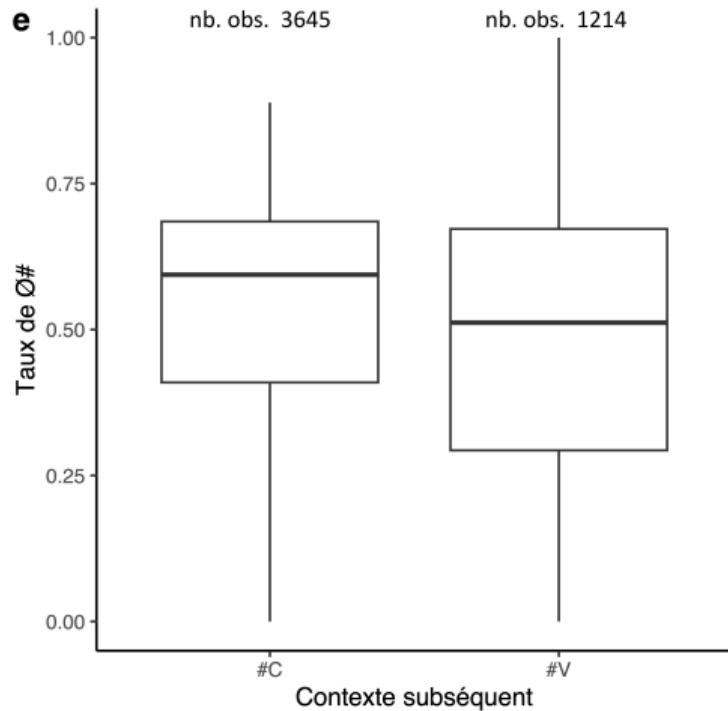


Figure 9 – ms3, cr, cr2-3. Taux de Ø.
 $t = -1; p = 0.321$

La variation -e/Ø des adverbes

Adverbes à -e de droit

- ▶ Peu d'effet de la qualité éditoriale
- ▶ L'absence de significativité ($p = 0.3$) peut venir d'un plus faible nombre de textes
- ▶ Effet du statut vers/prose ?

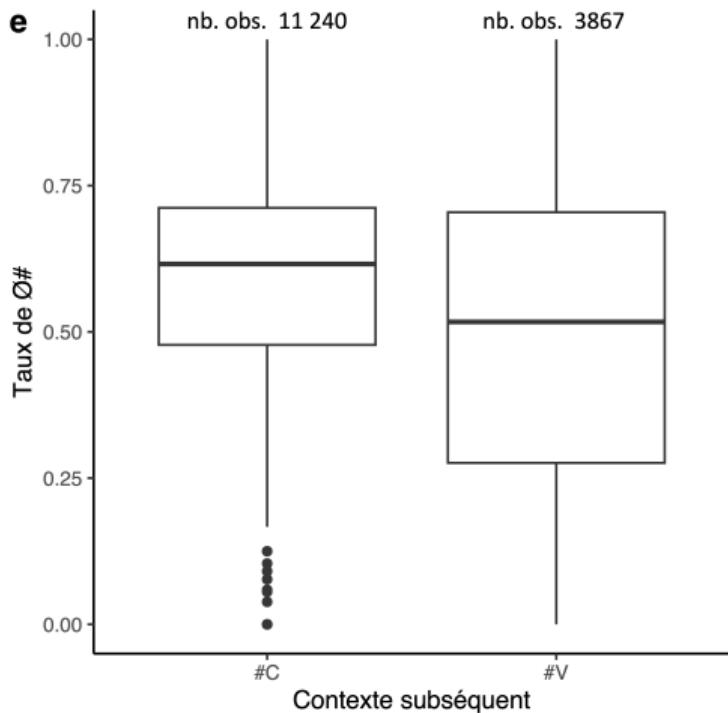


Figure 10 – Vers. Taux de Ø.

$t = -2.92; p < 0.005$

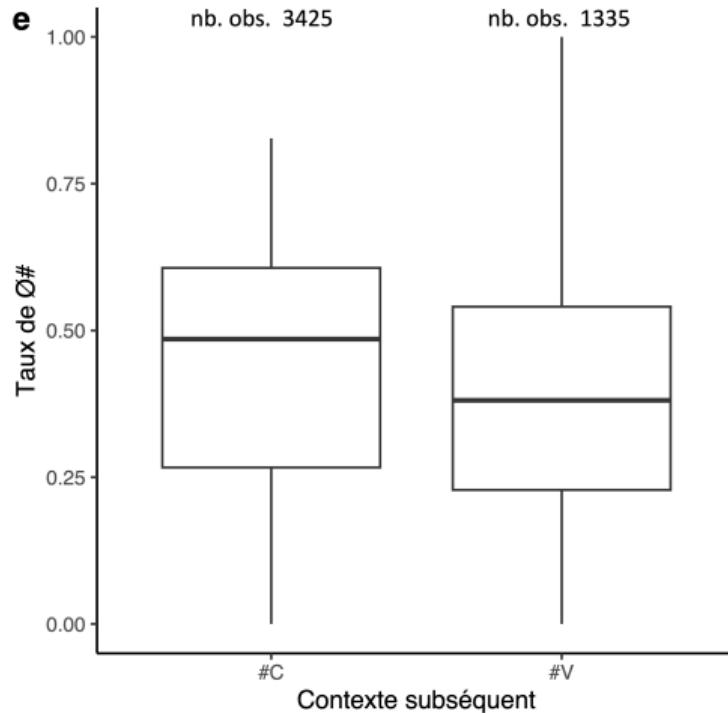


Figure 11 – Prose. Taux de Ø.

$t = -1.3; p = 0.198$

La variation -e/Ø des adverbes

Adverbes à -e de droit | Conclusion

- ▶ En prose : taux de Ø moins élevés et effet du contexte non significatif ($p = 0.2$)
- ▶ Mais pas d'effet du statut vers/prose

La variation -e/Ø des adverbes

Adverbes à -e de droit | Conclusion

- ▶ En prose : taux de Ø moins élevés et effet du contexte non significatif ($p = 0.2$)
- ▶ Mais pas d'effet du statut vers/prose
- ▶ Les résultats d'ensemble sont corroborés : ni le lemmatiseur, ni la qualité éditoriale, ni le statut vers/prose ne peuvent inverser la tendance.
- ▶ L'hypothèse traditionnelle n'est pas confirmée :
 - ▶ On ne voit d'effet positif du contexte #V
 - ▶ En synchronie de l'a.fr., ce n'est pas une élision
 - ▶ (MEYER-LÜBKE (1890) : ce n'était pas non plus le cas en t.a.fr.)

Conclusion

Conclusion

Comparalem

Comparalem permet :

- ▶ la détection des lemmes atteints par une variation graphique donnée
- ▶ la projection de cette variation dans le temps et dans l'espace, en tenant compte du contexte
- ▶ l'application de tests statistiques
- ▶ de manière semi-automatisée

Conclusion

Comparalem : usage

Comparalem permet :

- ▶ d'interroger des phénomènes non documentés dans la littérature (cf. PREMAT 2023)
- ▶ de ré-interroger des phénomènes décrits dans la littérature, pour les confirmer/remettre en question
- ▶ en prenant en compte de multiples facteurs de variation

Conclusion

Comparalem : futur

Objectif à court terme :

- ▶ développement d'une version stable et publique, mise à disposition librement
- ▶ pour participer à la réinvention de la phonétique historique, par un retour aux données, à grande échelle

Références

- DEES, Anthonij (1987). *Atlas des formes linguistiques des textes littéraires de l'ancien français.* Tübingen : M. Neimeyer Verlag.
- DEIMIER, Pierre de (1610). *L'Académie de l'art poétique.* Paris : J. de Bordeaux.
- ESTIENNE, Robert (1569 [1557]). *Traicté de la grāmaire Francoise.* 2^e éd. Paris : Robert Estienne.
- FOUCHÉ, Pierre (1969 [1958]). *Phonétique historique du français — Les voyelles.* 2^e éd. T. 2. 3 t. Paris : Klincksieck.
- GLESGEN, Martin-D. et Xavier GOUVERT (2007). « La base textuelle du *Nouveau Corpus d'Amsterdam* : Ancre diasystématique et évaluation philologique ». In : *Le Nouveau Corpus d'Amsterdam : actes de l'atelier de Lauterbad.* Sous la dir. de Pierre KUNSTMANN et Achim STEIN. Stuttgart : Steiner, p. 51-84.
- MALHERBE, François de ([1862]). *Œuvres.* Sous la dir. de M. L. LALANNE. Paris : Hachette, [url t. iv].
URL : <https://gallica.bnf.fr/ark:/12148/bpt6k5210r>.
- MARCHELLO-NIZIA, Christiane (2015). « De JE à J' en français : étapes vers l'élosion, interactions entre phonétique et syntaxe ». In : *Diachroniques* 5, p. 17-43.

- MARCHELLO-NIZIA, Christiane et al. (2020). *Grande Grammaire Historique du Français*. Berlin/Boston : Mouton de Gruyter.
- MEIGRET, Louis (1550). *Le tretté de la grammere françoëze*. Paris : Chrestien Wechel. URL : <https://gallica.bnf.fr/ark:/12148/btv1b8624665r>.
- MEYER-LÜBKE, Wilhelm (1890). *Grammaire des langues romanes*. Trad. par Eugène RABIET. T. 1. Paris : E. Welter.
- PALSGRAVE, Jehan (1530). *L'Esclarcissemest de la langue françoysse*. London : John Haukyns.
- POPE, Mildred Katherine (1966 [1934]). *From Latin to Modern French with Especial Consideration of Anglo-Norman*. 2^e éd. Manchester : Manchester University Press.
- PREMAT, Timothée (2023). *La genèse de l'éisión*. Nice : Université Côte d'Azur. URL : <https://theses.hal.science/tel-04574584>. Thèse de doctorat.
- RAINSFORD, Thomas (2010). « Rhythmic change in the medieval octosyllable and the development of group stress ». In : *Congrès Mondial de Linguistique Française - CMLF 2010*. DOI : 10.1051/cmlf/2010174.

RAINSFORD, Thomas (2011a). « Dividing lines: the changing syntax and prosody of the mid-line break in medieval French octosyllabic verse ». In : *Transactions of the Philological Society* 109.3, p. 265-283.

- (2011b). « The Emergence of Group Stress in Medieval French ». Thèse de doct. Cambridge : Cambridge University, St. John's College.
- (2020). « Syllable structure and prosodic words in Early Old French ». In : *Papers in Historical Phonology* 5, p. 63-89. DOI : 10.2218/pihph.5.2020.4433.
- (2022). « Old Gallo-Romance (OGR) Corpus : annotation phonologique et métrique des plus anciens textes gallo-romans ». In : *SHS Web Conf.* 138.

RAINSFORD, Thomas et Olga SCRIVNER (2014). « Metrical annotation for a verse treebank ». In : *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT 13)*. Sous la dir. de Verena HENRICH et al. Tübingen : Département de Linguistique, Université de Tübingen.

RAMÉE, Pierre de la (1572). *Grammaire*. Paris : André Wechel.

RONSARD, Pierre de (1866 [1565]). « Abbrégé de l'art poetique françois ». In : *Œuvres complètes*. Sous la dir. de Prosper BLANCHEMIN. Paris : A. Franck, [première édition Paris : Gabriel Buon]. URL (1565, facs.) : <https://gallica.bnf.fr/ark:/12148/bpt6k4601j>.

SÉGÉRAL, Philippe et Tobias SCHEER (2015). « Présentation ». In : *Diachroniques* 5, p. 7-15.

STEIN, Achim, Pierre KUNSTMANN et Martin-D. GLESGGEN (2006). *Nouveau Corpus d'Amsterdam. Corpus informatique de textes littéraires d'ancien français (ca 1150-1350)*. établi par Anthonij Dees (Amsterdam : 1987), remanié par Achim Stein, Pierre Kunstmann et Martin-D. Gleßgen, Stuttgart : Institut für Linguistik/Romanistik.

SYLVIUS [DU Bois], Jacques (1531). *In linguam gallicam Isagoge, una cum eiusdem grammatica latino-gallica, ex Hebræis, Græcis et Latinis authoribus*. Paris : Robert Estienne. URL : <https://gallica.bnf.fr/ark:/12148/bpt6k50952q>.

VAUGELAS, Claude Favre de (1880 [1647]). *Remarques sur la langue françoise*. 2 t. Versailles/Paris : Cerf et fils/J. Baudry [réédition contenant un supplément (nouvelles remarques posthumes et observations) ; première édition Paris : La veuve Jean Camusat et Pierre Le Petit].