Impact du type de référent sur la composition des chaînes référentielles

Silvia Federzoni, Lydia-Mai Ho-Dac et Cécile Fabre
CLLE, CNRS et Université de Toulouse Jean Jaurès

Plan

Les chaînes de référence

Chaînes de référence

Suite d'expressions d'un texte entre lesquelles l'interprétation établit une identité de référence (Corblin, 1995)

George H.W. Bush s'inquiéta de la volonté du régime irakien de se doter d'armes de destruction massive, [...] Se sentant dans une situation de plus en plus délicate, **le président** comprit que la guerre était inévitable, et que son issue déterminerait **son** propre avenir politique. **Il** estimait en effet que la mauvaise conduite des opérations pouvait **lui** coûter cher [...]

Maillons : expressions linguistiques porteuses d'une **valeur instructionnelle**.

Leur succession contribue à créer des **liens de cohésion**
(Halliday & Hasan, 1976)

Les chaînes de référence

Chaînes de référence

Suite d'expressions d'un texte entre lesquelles l'interprétation établit une identité de référence (Corblin, 1995)

George H.W. Bush s'inquiéta de la volonté du régime irakien de se doter d'armes de destruction massive, [...] Se sentant dans une situation de plus en plus délicate, **le président** comprit que la guerre était inévitable, et que son issue déterminerait **son** propre avenir politique. **Il** estimait en effet que la mauvaise conduite des opérations pouvait **lui** coûter cher [...]

Chaînes de référence : méchanisme fondamental dans **l'organisation** et **l'interprétation** du discours

Les chaînes de référence : différentes approches

- En linguistique cognitive : théorie de l'accessibilité (Ariel, 2001) et théorie du centrage (Walker, Joshi, & Prince, 1998)
  - Principes qui régissent l'interprétation et la construction de la continuité référentielle

- En psycholinguistique (Gundel, Hedberg, & Zacharski, 2019 ; Kaiser & Fedele, 2019 ; Roberts, 2019 ; Salazar Orvig, 2019)
  - Facteurs qui influencent le choix des expressions référentielles (Fossard et al., 2018 ; Vogels, Krahmer, & Maes, 2013)

Les chaînes de référence : différentes approches

- En linguistique descriptive (Charolles, 2002 ; Cornish, 2000 ; Kleiber, 1994, 2002 ; Schnedecker, 2005)
  - Description des variations selon les genres textuels (Schnedecker, 2014 ; Schnedecker & Landragin, 2014 ; Schnedecker & Longo, 2012)
  - Observations fines, petits corpus, prise en compte d'un seul type de référent ou peu de paramètres pour caractériser les chaînes
- En traitement automatique du langage (TAL) (Landragin, 2018 ; Longo, 2013 ; Poesio, Pradhan, Recasens, Rodriguez, & Versley, 2016 ; Wilkens, Oberle, Landragin, & Todirascu, 2020)
  - Détection automatique des chaînes de référence et des maillons (De Marneffe, Recasens, & Potts, 2015 ; Mitkov, 2014 ; Recasens & Hovy, 2010)
  - Observations systématiques, diffusion de gros corpus annotés
- Corpus : AnnoDis (2010), AnCor (2014), Democrat (2019), E-Calm (2019-2021)

Les chaînes de référence : différentes approches

- En linguistique descriptive (Charolles, 2002 ; Cornish, 2000 ; Kleiber, 1994, 2002 ; Schnedecker, 2005)
  - Description des variations selon les genres textuels (Schnedecker, 2014 ; Schnedecker & Landragin, 2014 ; Schnedecker & Longo, 2012)
  - Observations fines, petits corpus, prise en compte d'un seul type de référent ou peu de paramètres pour caractériser les chaînes

- En traitement automatique du langage (TAL) (Landragin, 2018 ; Longo, 2013 ; Poesio et al., 2016 ; Wilkens et al., 2020)
  - Détection automatique des chaînes de référence et des maillons (De Marneffe et al., 2015 ; Mitkov, 2014 ; Recasens & Hovy, 2010)
  - Observations systématiques, diffusion de gros corpus annotés

- Corpus : AnnoDis (2010), AnCor (2014), Democrat (2019), E-Calm (2019-2021)

Les chaînes de référence : différentes approches

- En linguistique descriptive (Charolles, 2002 ; Cornish, 2000 ; Kleiber, 1994, 2002 ; Schnedecker, 2005)
  - Description des variations selon les genres textuels (Schnedecker, 2014 ; Schnedecker & Landragin, 2014 ; Schnedecker & Longo, 2012)
  - Observations fines, petits corpus, prise en compte d'un seul type de référent ou peu de paramètres pour caractériser les chaînes

- En traitement automatique du langage (TAL) (Landragin, 2018 ; Longo, 2013 ; Poesio et al., 2016 ; Wilkens et al., 2020)
  - Détection automatique des chaînes de référence et des maillons (De Marneffe et al., 2015 ; Mitkov, 2014 ; Recasens & Hovy, 2010)
  - Observations systématiques, diffusion de gros corpus annotés

- Corpus : AnnoDis (2010), AnCor (2014), Democrat (2019), E-Calm (2019-2021)

Objectif

Typologie des chaînes de référence

Mettre au jour des **types** de chaînes en fonction de leur **rôle discursif**

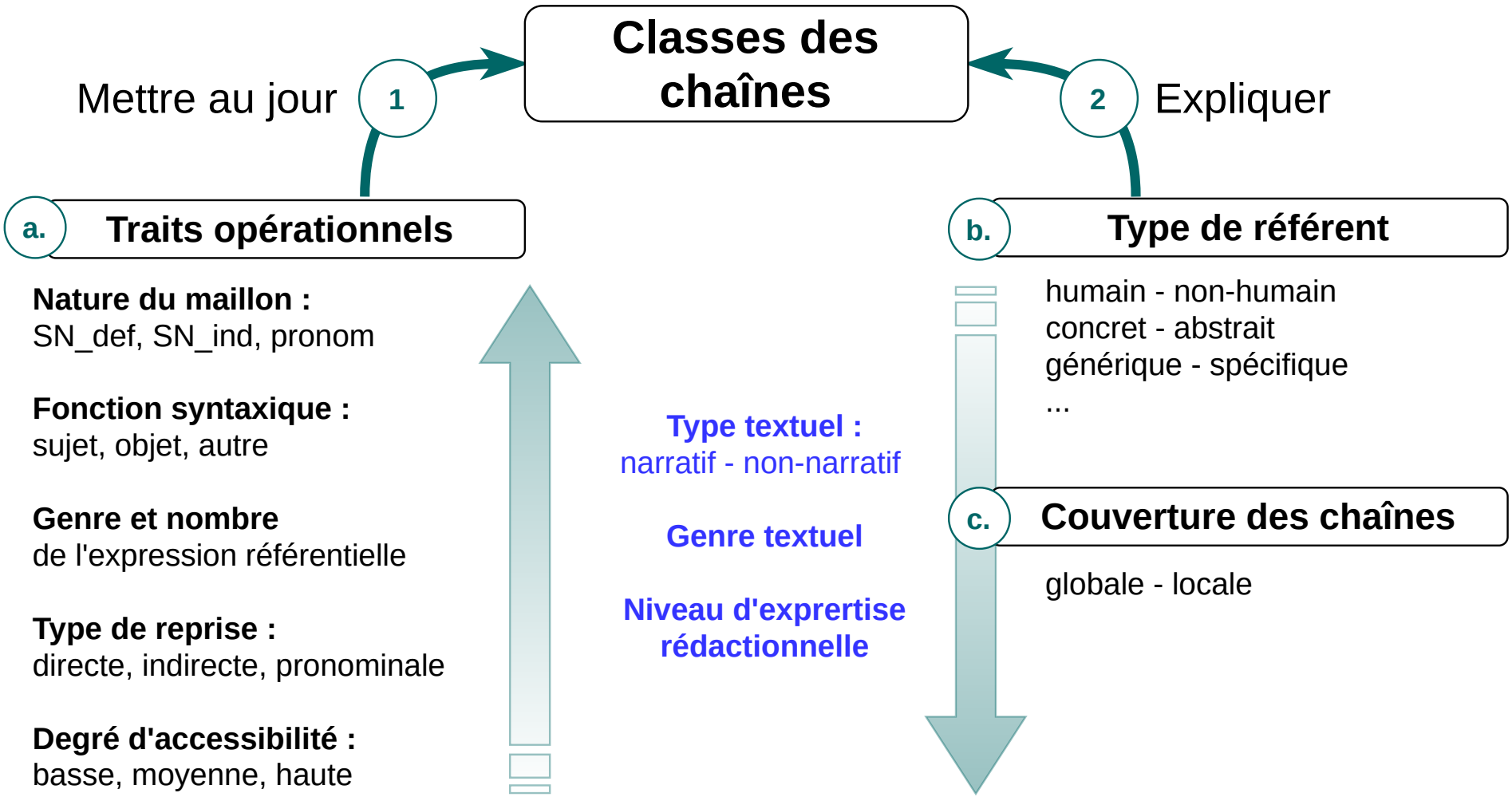
Proposer une description **systématique** et **exhaustive** :

- de la **composition**
- de la **variété**
- de la **complexité**

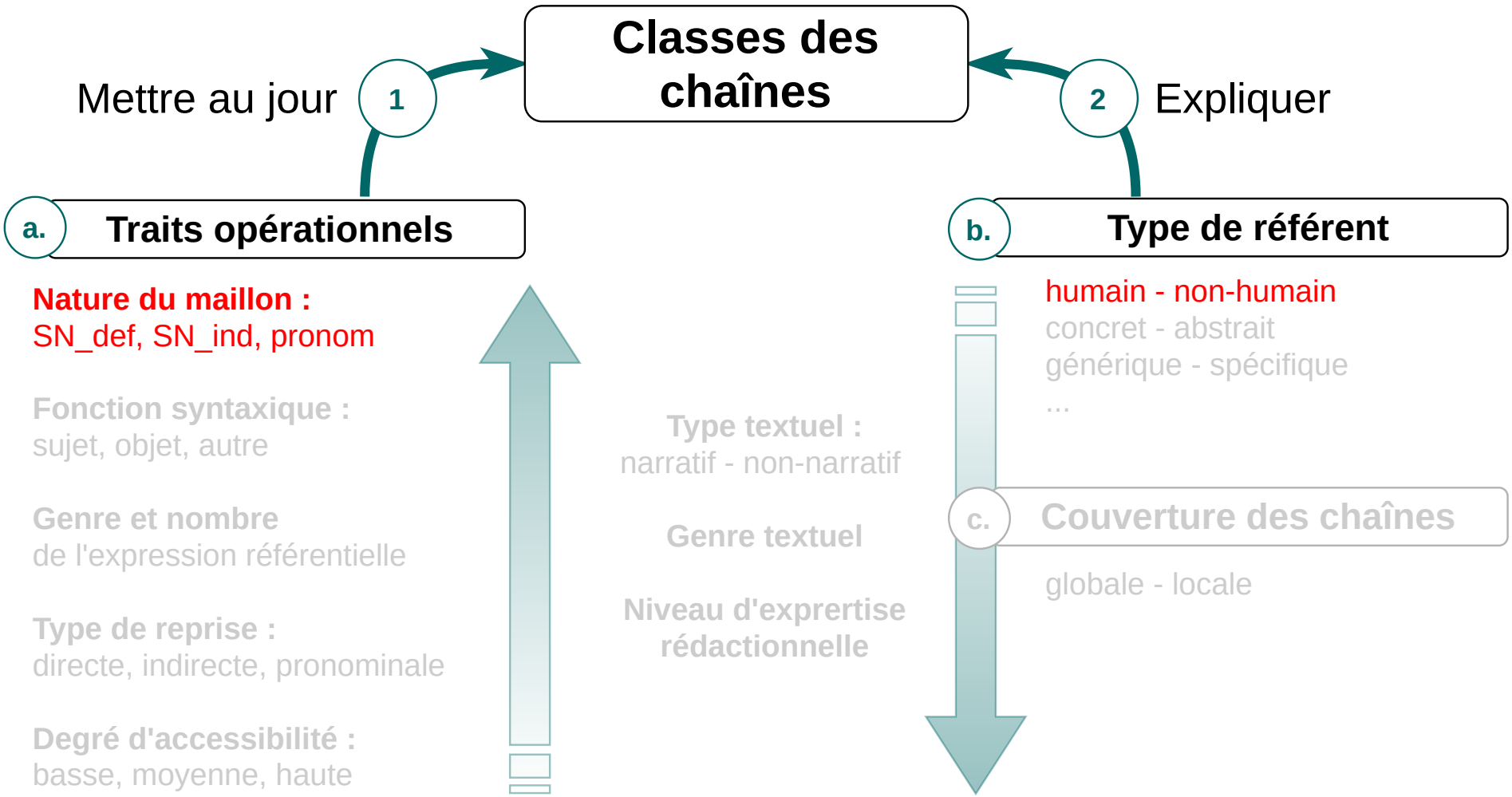
des chaînes de référence

Mettre en place une **méthode automatique** permettant une **analyse systématique** d'un grand nombre de **traits** à partir de **ressources diversifiées**

Méthode de description



Méthode de description



Plan

La ressource AnnoDis (Péry-Woodley, Afantenos, Ho-Dac, & Asher, 2011)

- Ressource annotée en **chaînes topicales** $\Rightarrow$ assimilées aux chaînes de référence (Federzoni, Ho-Dac, & Rebeyrolle, 2020)

  - textes longs structurés (87 textes, 7655 mots/texte)
  - entièrement annotés (581 CR, 3456 maillons)
  - non narratifs (expositifs)
  - 3 genres textuels (rapports, articles scientifiques, textes encyclopédiques)

Vers une typologie des chaînes

Objectif : décrire la composition des chaînes

Mettre au jour des classes de chaînes de référence en observant les **enchaînements** les plus fréquents

- Compléter les analyses traditionnelles (Corblin, 1987 ; Cornish, 1998 ; Longo, 2013 ; Salles, 2015)
- Prendre en compte la **complexité** et la **variété** de **composition** des chaînes

Vers une typologie des chaînes

Objectif : décrire la composition des chaînes

Mettre au jour des classes de chaînes de référence en observant les **enchaînements** les plus fréquents

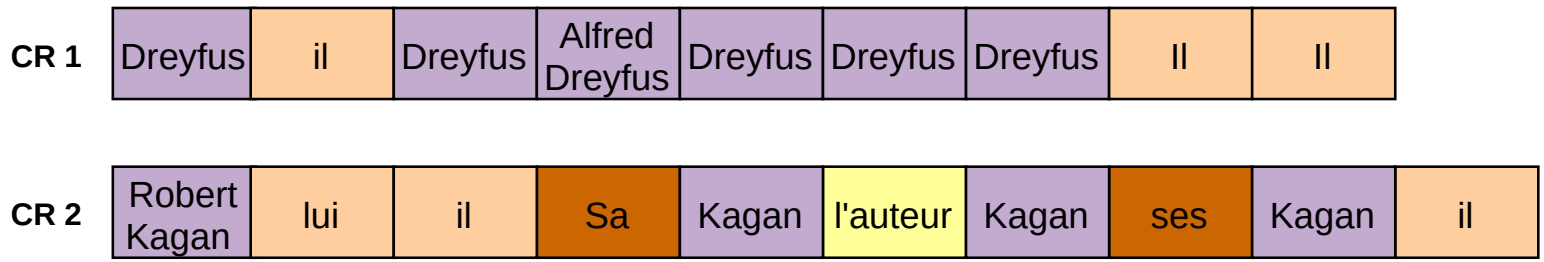
- Compléter les analyses traditionnelles (Corblin, 1987 ; Cornish, 1998 ; Longo, 2013 ; Salles, 2015)
- Prendre en compte la **complexité** et la **variété** de **composition** des chaînes

Vers une typologie des chaînes

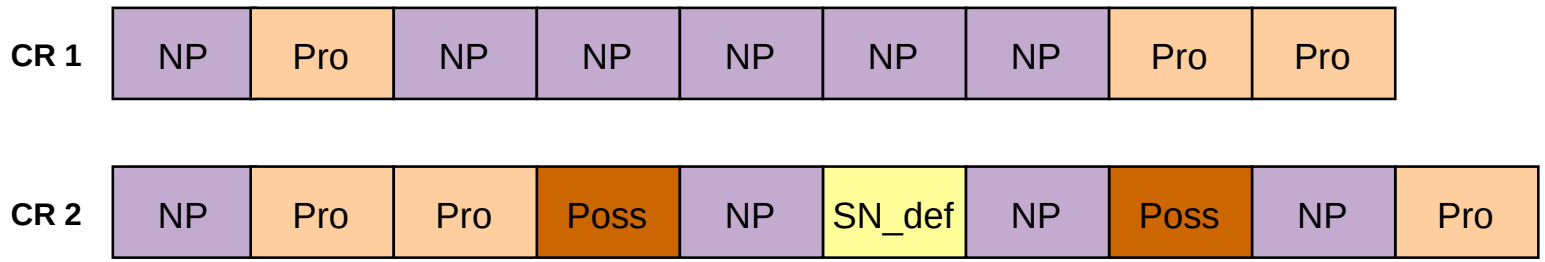
Objectif : décrire la composition des chaînes

Mettre au jour des classes de chaînes de référence en observant les **enchaînements** les plus fréquents

- Compléter les analyses traditionnelles (Corblin, 1987 ; Cornish, 1998 ; Longo, 2013 ; Salles, 2015)
- Prendre en compte la **complexité** et la **variété** de **composition** des chaînes



Vers une typologie des chaînes



Nature maillons	#	%
SN_def	1198	34,66
SN_dem	272	7,87
SN_ind	126	3,65
SN_sansDET	64	1,85
NP	442	12,79
Poss	182	5,27
Pro	1026	29,69
Autre	146	4,22
Total	3456	100

Trait utilisé : nature des maillons

- La fréquence des catégories grammaticales : variations les plus fortes
- Informations les plus riches sur la typologie des chaînes : degré d'homogénéité

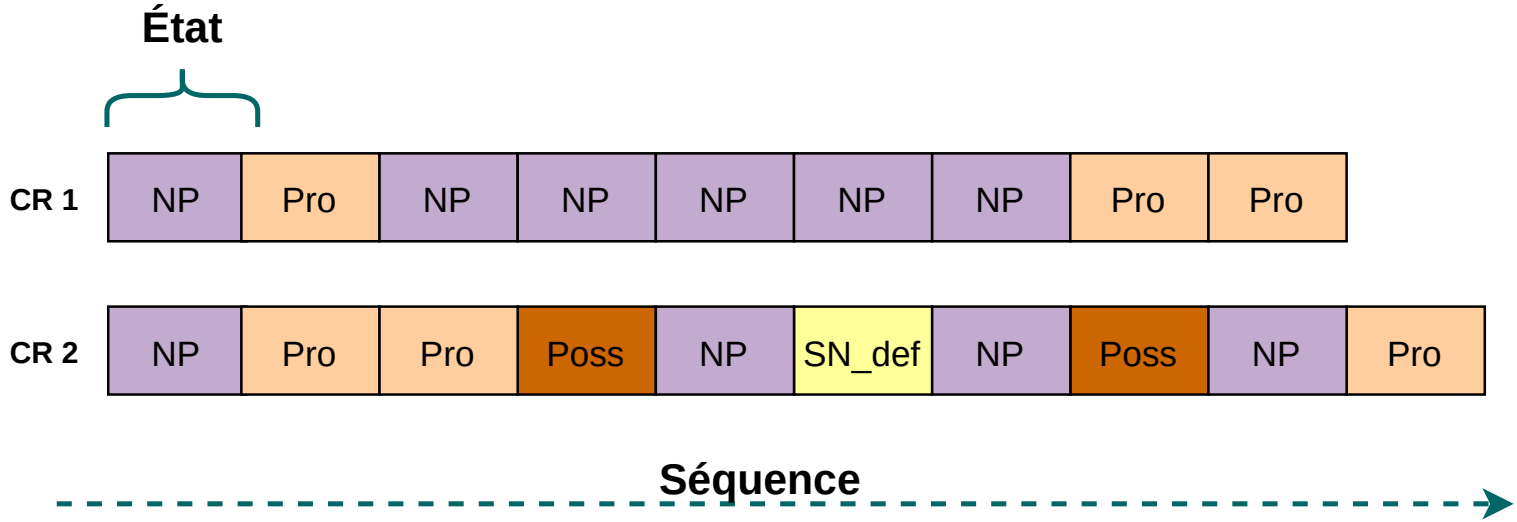
(Obry, Glikman, Guillot-Barbance, & Pincemin, 2017)

Vers une typologie des chaînes

Méthode : analyse des séquences (Quiniou, Cellier, Charnois, & Legallois, 2012)

Identifier les régularités, les ressemblances, construire des typologies de « séquences-types » (Robette, 2011)

Séquence : liste ordonnée d'**états** ou d'événements (Brzinsky-Fay, Kohler, & Luniak, 2006)



Méthode - analyse de séquences

TraMineR (Gabadinho, Ritschard, Studer, & Müller, 2009) :

- Visualisation spécifique pour l'analyse des séquences
- Méthodes de classification automatique
- Méthodes statistiques

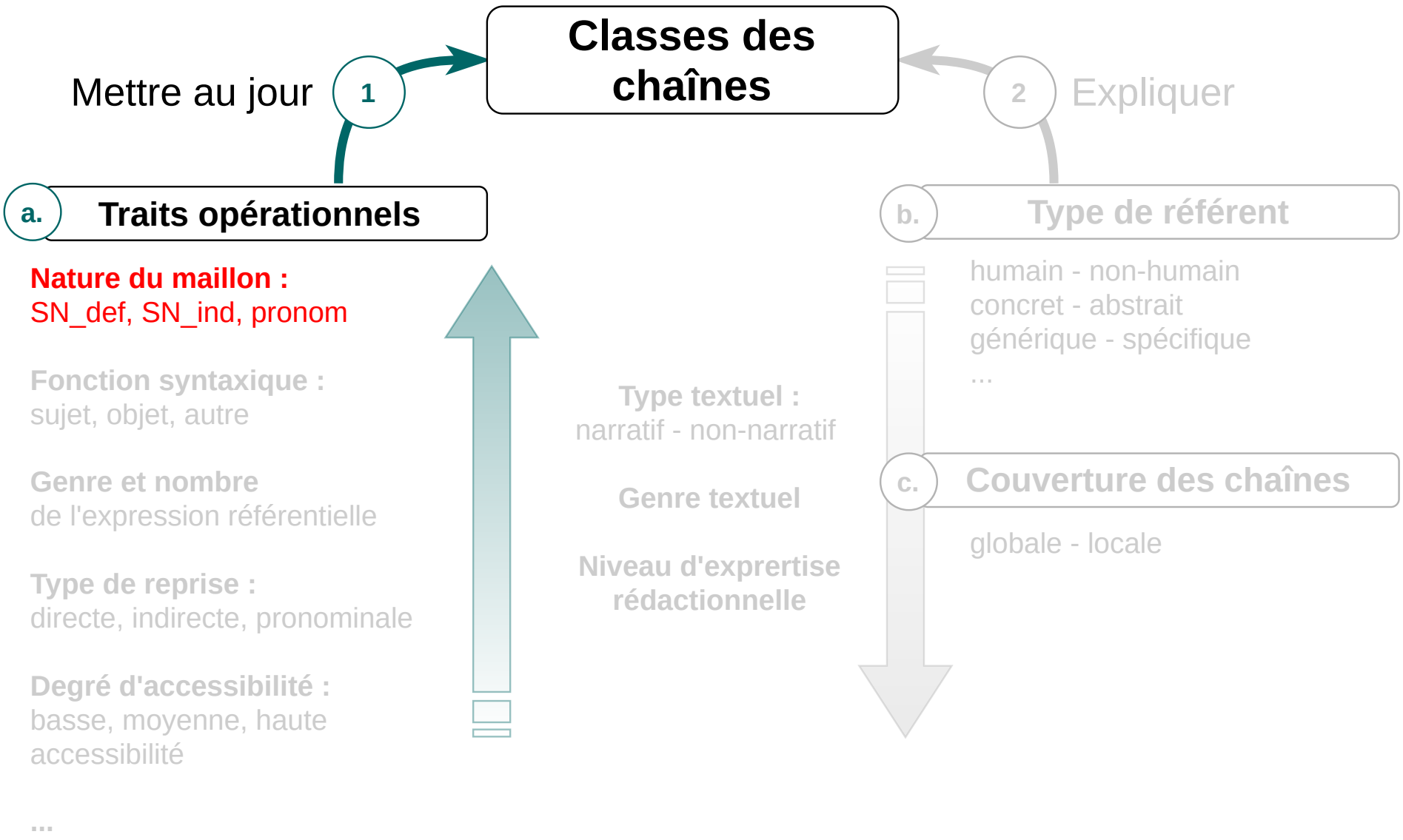
Paramètres :

① Méthode de classification : clustering hiérarchique (méthode de 'Ward')

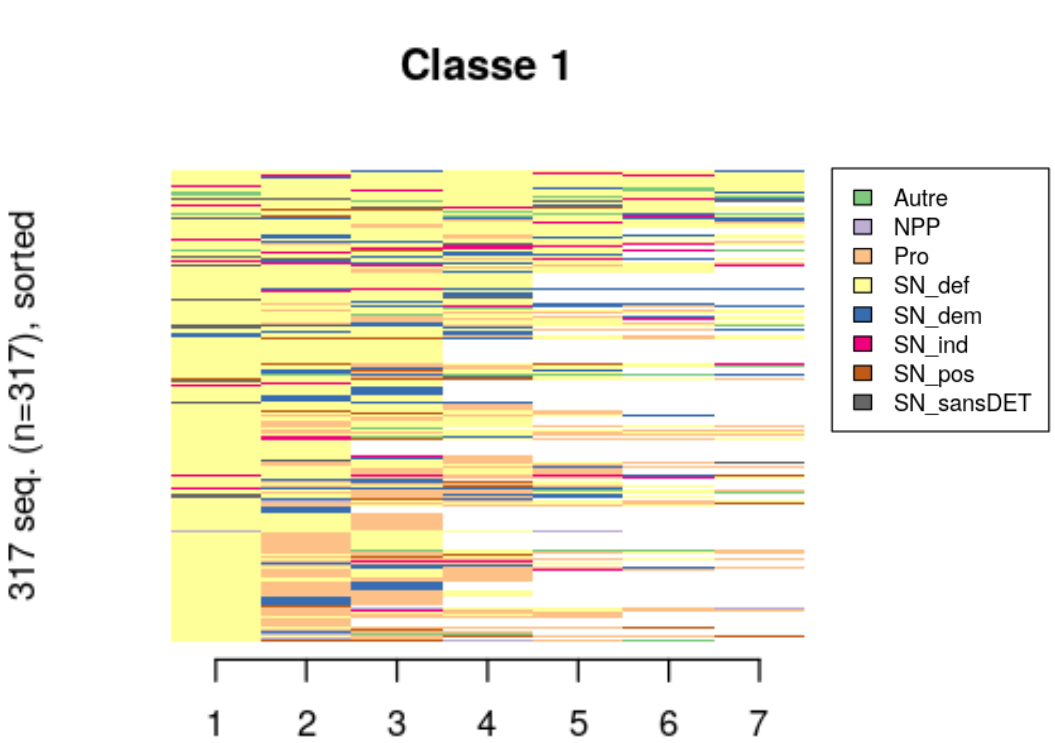
② Taille des séquences : entre 2 et 7 (73,15% des CR ont moins de 7 maillons)

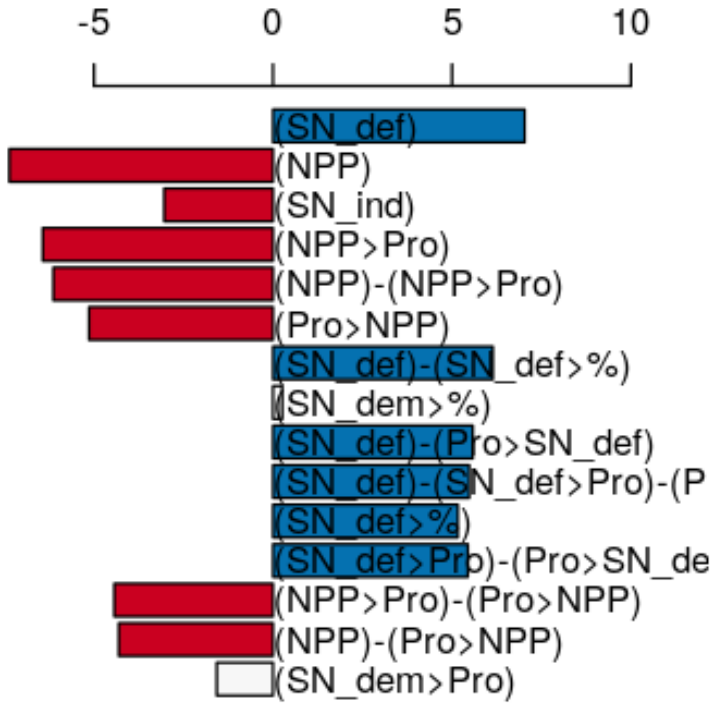
③ Nombre de classes : 3 classes (observation de 2,3 et 5 classes)

Mettre au jour des classes



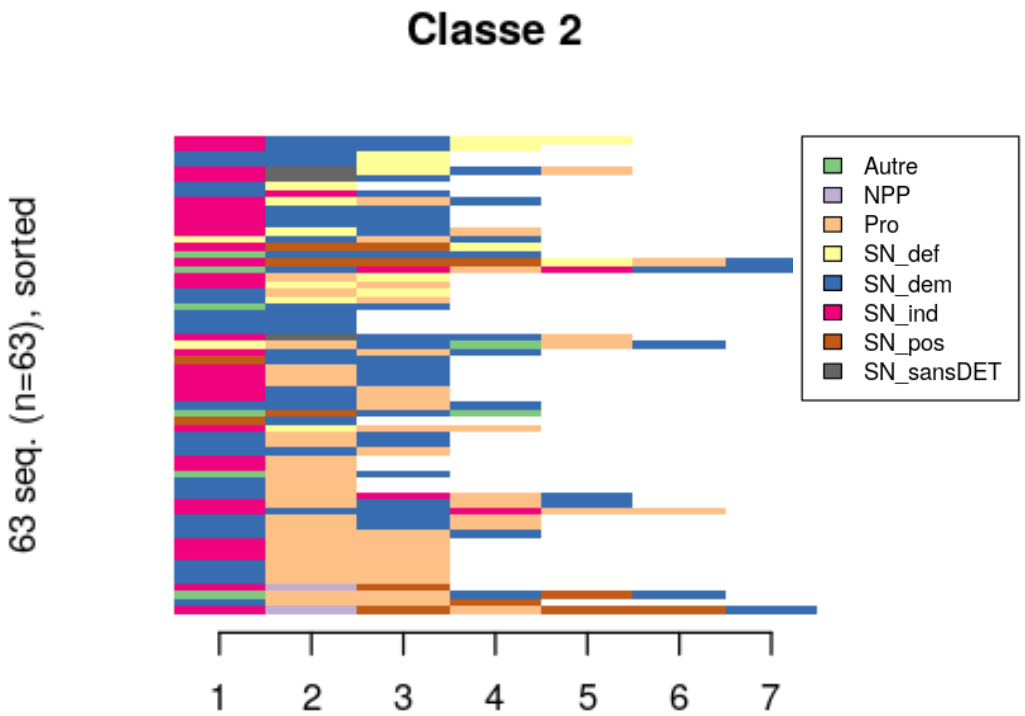
Classe TraMineR 1

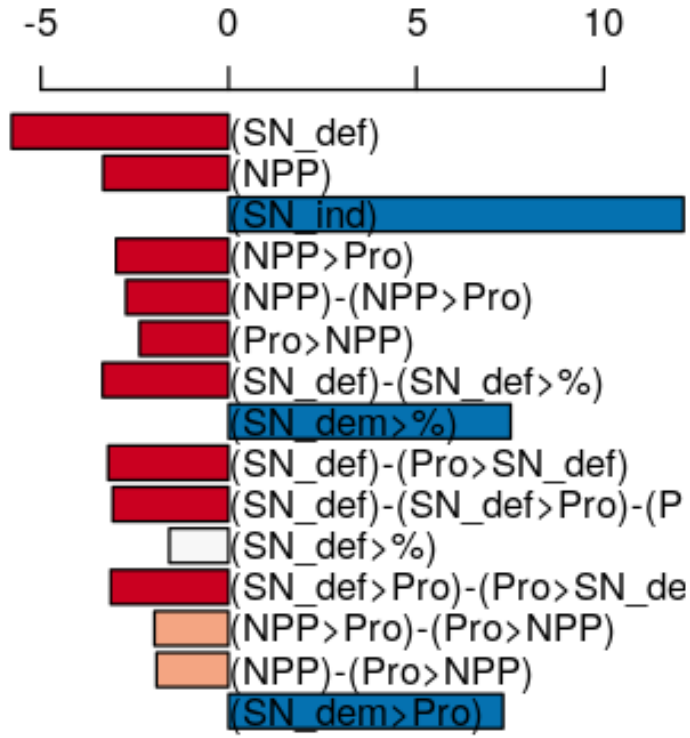




le vignoble champenois *s'étendait sur quelques [...]* *le vignoble connaît [...] Après les fléau du phylloxéra et de la Grande guerre,* *le vignoble* *s'est réduit à 12 000 hectares. Aujourd'hui, en 2007,* *le vignoble champenois* *s'étend sur 32 341 hectares.*

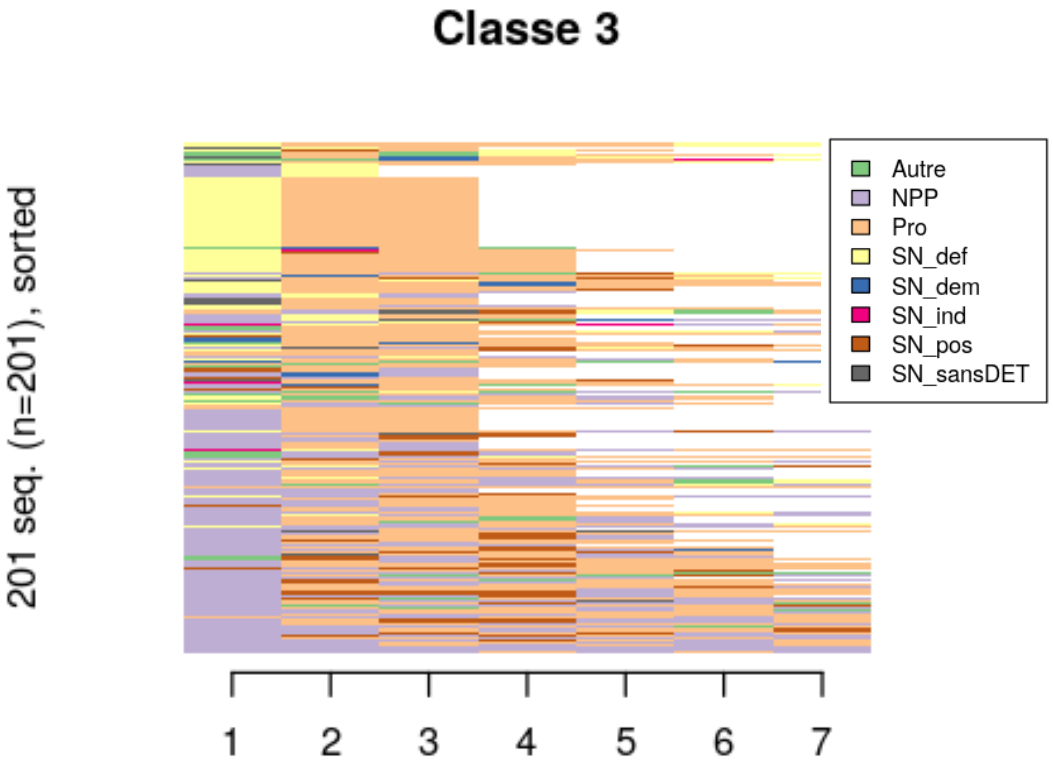
Classe TraMineR 2

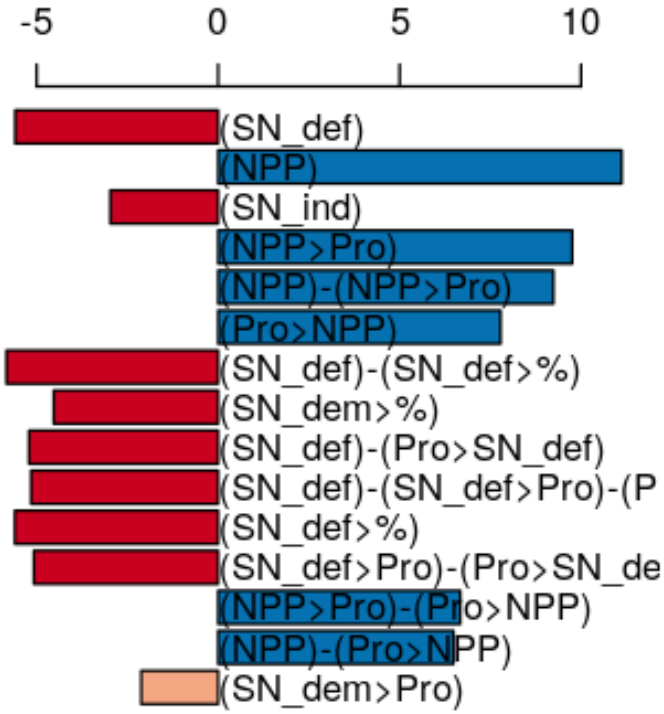




[...] « le " communicatif " présente ainsi fréquemment une connotation oppositionnelle (sinon contradictoire) avec le linguistique. Ceci est particulièrement crucial lorsqu'on traite de la " compétence de communication " [...] Cette dichotomie pose problème au psychologue [...].

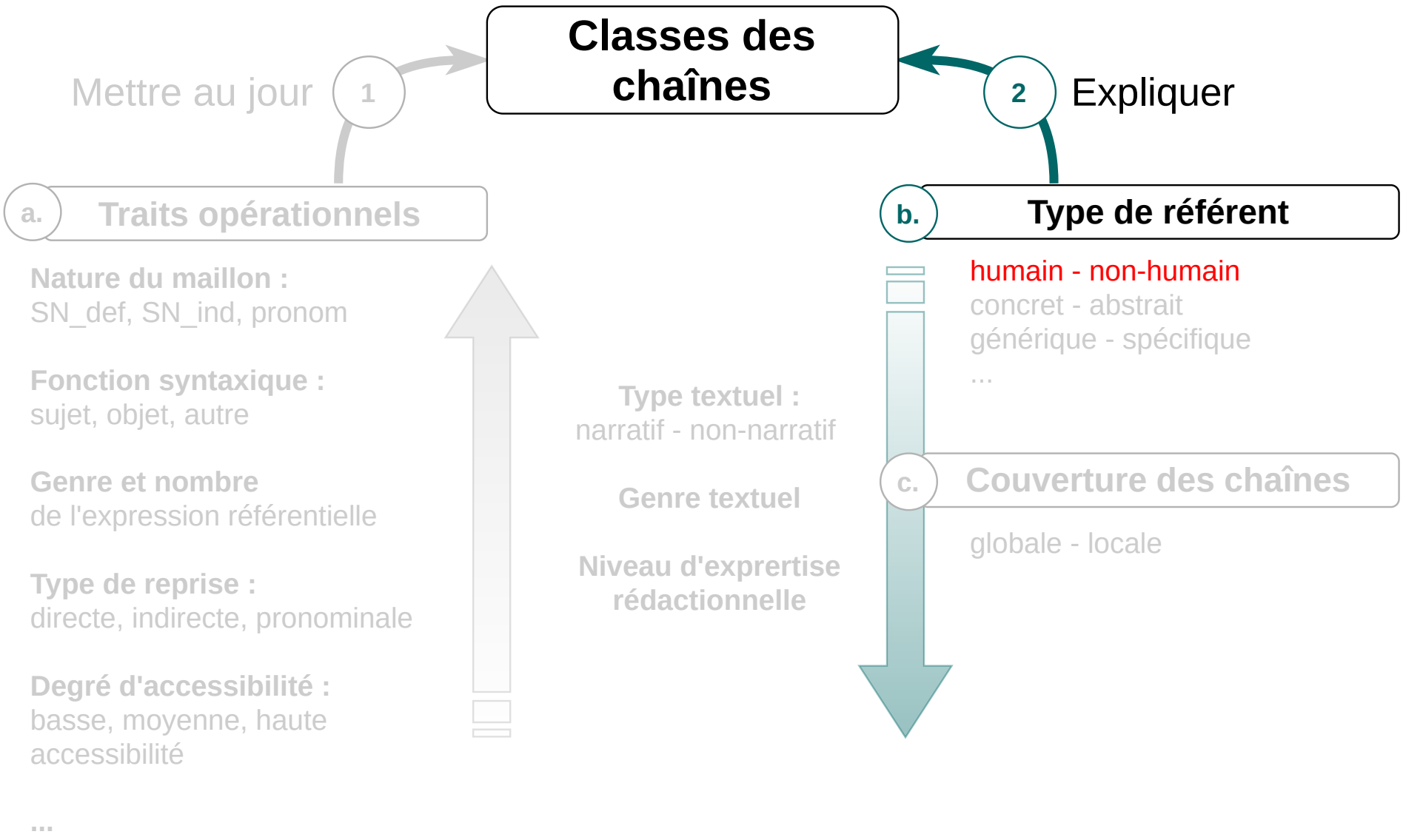
Classe TraMineR 3



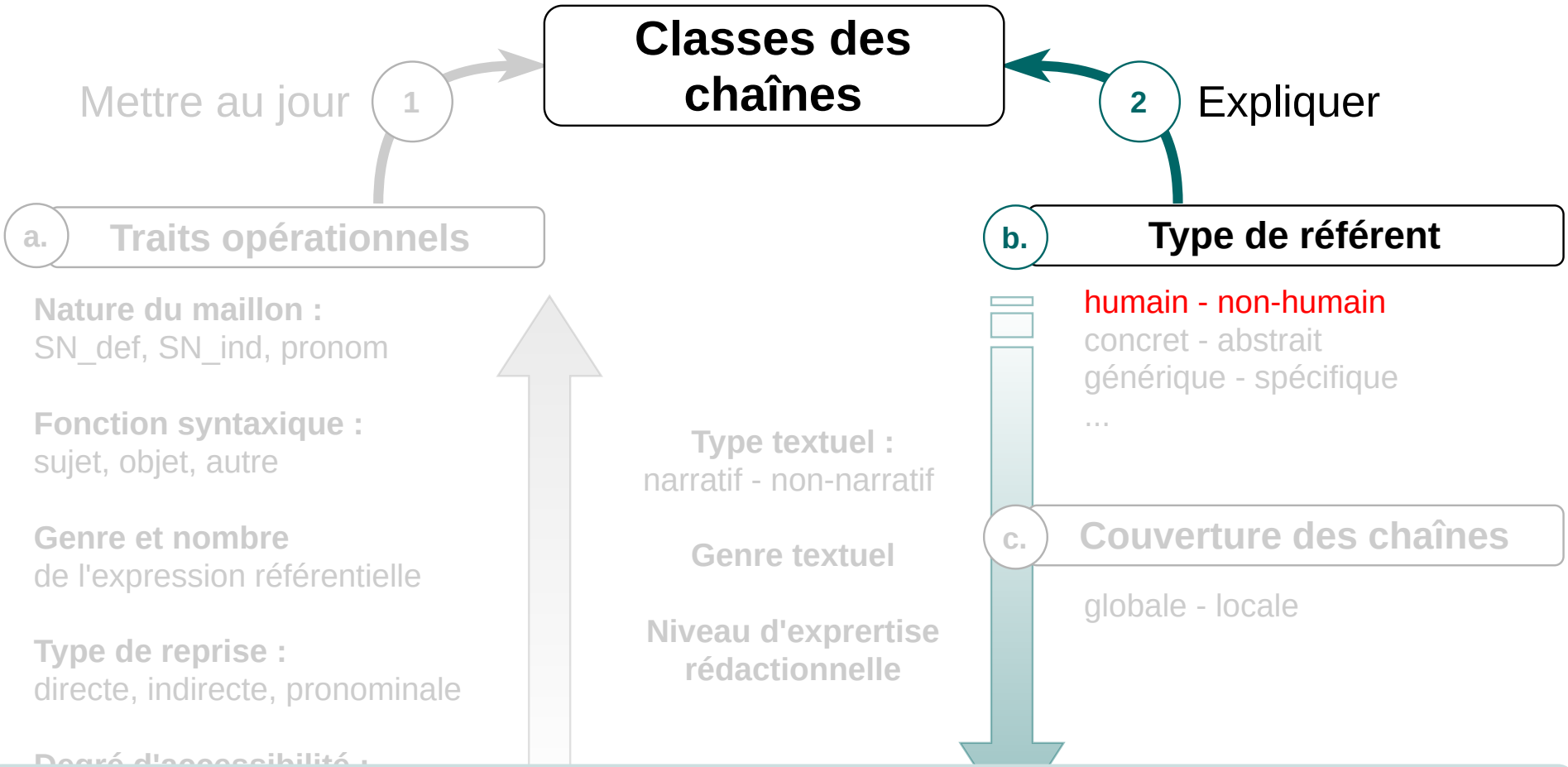


Chez F. de Saussure *, l'analogie [...],* il *pose que les facteurs de trouble [...]. Pour* lui *, cette tendance à l'irrégularité est heureusement contre-balancée par l'analogie [...]. Comme H. Paul,* il *ramène le concept au calcul de l'équation de la quatrième proportionnelle.*

Expliquer les Classes TraMineR



Expliquer les Classes TraMineR

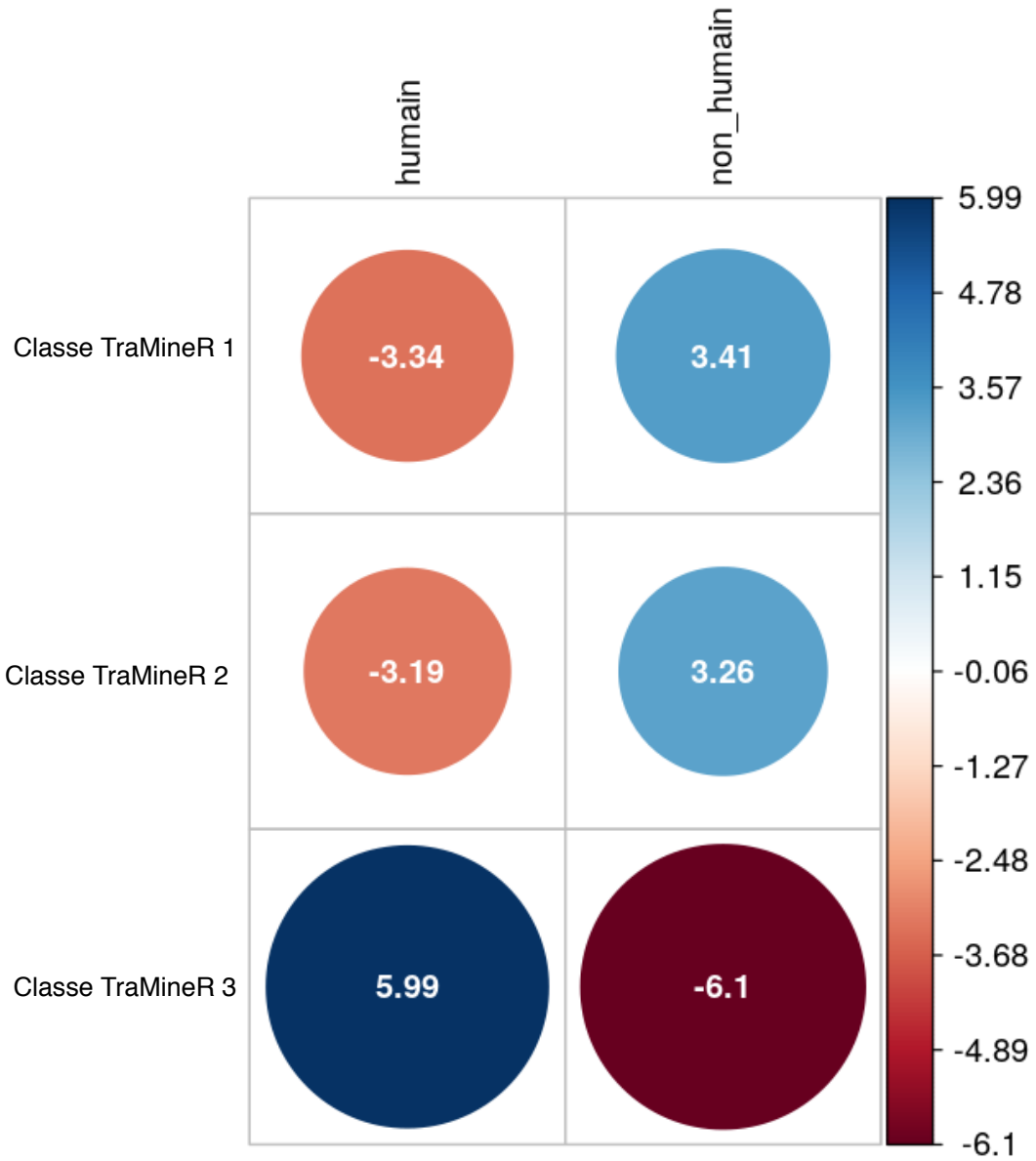


Type de référent : annotation semi-automatique

...

Projeter le type de référent

Conclusion : liaison significative entre les classes et le type de référent (test du khi-deux)



Analyser les classes TraMineR

	Composition	**Human**		**Non-human**	
		#	**%**	**#**	**%**
Classe TraMineR 1	SN_def++	119	37,54	**198**	**62,46**
Classe TraMineR 2	SN_ind SN_dem	14	22,22	**49**	**77,78**
Classe TraMineR 3	NP+ Pro+	**163**	**81,09**	38	18,91

Analyser les classes TraMineR

	Composition	Human		Non-human	
		#	%	#	%
Classe TraMineR 1	SN_def++	119	37,54	**198**	**62,46**
Classe TraMineR 2	SN_ind SN_dem	14	22,22	**49**	**77,78**
Classe TraMineR 3	NP+ Pro+	**163**	**81,09**	38	18,91

Patron : SN_def{2:7} (24 CR)

Référent non humain

le vignoble champenois s'étendait sur quelques [...] le vignoble connaît [...] Après les fléau du phylloxéra et de la Grande guerre, le vignoble s'est réduit à 12 000 hectares. Aujourd'hui, en 2007, le vignoble champenois s'étend sur 32 341 hectares.

Analyser les classes TraMineR

	Composition	Human		Non-human	
		#	%	#	%
Classe TraMineR 1	SN_def++	119	37,54	**198**	**62,46**
Classe TraMineR 2	SN_ind SN_dem	14	22,22	**49**	**77,78**
Classe TraMineR 3	NP+ Pro+	**163**	**81,09**	38	18,91

Patron : SN_def{2:7} (22 CR)

Référent humain

En effet, les parlementaires ont le pouvoir de bloquer les propositions du président [...] le Congrès peut même orienter positivement les choix de l'Exécutif [...], quand les parlementaires imposèrent au président d'adopter des sanctions à l'égard de l'Afrique du Sud [...] Dans l'ensemble, le Congrès dispose de prérogatives [...].

Analyser les classes TraMineR

	Composition	Human		Non-human	
		#	%	#	%
Classe TraMineR 1	SN_def++	119	37,54	**198**	**62,46**
Classe TraMineR 2	SN_ind SN_dem	14	22,22	**49**	**77,78**
Classe TraMineR 3	NP+ Pro+	**163**	**81,09**	38	18,91

Patron : SN_def{1:2} > Pro{1:6} (42 CR)

Référent humain

Le prisonnier n'est en rien au courant des événements qui se déroulent à des milliers de kilomètres de lui. Ni des complots ourdis pour que jamais il ne puisse revenir, [...]. À la fin de l'année 1898, il apprend avec stupéfaction la dimension réelle de l'Affaire, dont il ne sait rien [...].

Analyser les classes TraMineR

	Composition	Human		Non-human	
		#	%	#	%
Classe TraMineR 1	SN_def++	119	37,54	**198**	**62,46**
Classe TraMineR 2	SN_ind SN_dem	14	22,22	**49**	**77,78**
Classe TraMineR 3	NP+ Pro+	**163**	**81,09**	38	18,91

Patron : SN_def{1:2} > Pro{1:6} (24 CR)

Référent non humain

[...] le navire *, nommé Titan, [...] est présenté comme insubmersible grâce à ses 19 compartiments étanches. De fait,* il *ne dispose que du nombre minimum de canots de sauvetage requis par la loi.* Il *heurte un iceberg, , coule et la majorité des passagers sont victimes du naufrage.*

Plan

Conclusion et perspectives

- Méthode proposée valide pour mettre au jour des classes de chaînes

Prochaines étapes :

1. Prendre en compte d'autres traits opérationnels :
   - Fonction syntaxique : sujet, objet, autre
   - Degré d'informativité du maillon
   - Type de reprise : directe, indirecte, pronominale
   - Interdistance (Rousier-Vercruyssen & Landragin, 2019)
   - Instabilité des chaînes (Rousier-Vercruyssen & Landragin, 2019)
2. Combiner plusieurs traits opérationnels
3. Projeter d'autres types de référent :
   - concret - abstrait, générique - spécifique
   - distinctions plus fines : individus particuliers - humain collectifs
4. Reproduire la même analyse sur d'autres corpus : Democrat (Landragin, 2015) et E-CaIm (Garcia-Debanc, Ho-Dac, Bras, & Rebeyrolle, 2017)
   - type textuel
   - niveaux d'expertise rédactionnelle

Conclusion et perspectives

- Méthode proposée valide pour mettre au jour des classes de chaînes

Prochaines étapes :

① Prendre en compte d'autres traits opérationnels :
- Fonction syntaxique : sujet, objet, autre
- Degré d'informativité du maillon
- Type de reprise : directe, indirecte, pronominale
- Interdistance (Rousier-Vercruyssen & Landragin, 2019)
- Instabilité des chaînes (Rousier-Vercruyssen & Landragin, 2019)

② Combiner plusieurs traits opérationnels

③ Projeter d'autres types de référent :
- concret - abstrait, générique - spécifique
- distinctions plus fines : individus particuliers - humain collectifs

④ Reproduire la même analyse sur d'autres corpus : Democrat (Landragin, 2015) et E-Calm (Garcia-Debanc et al., 2017)
- type textuel
- niveaux d'expertise rédactionnelle

Conclusion et perspectives

- Méthode proposée valide pour mettre au jour des classes de chaînes

Prochaines étapes :

① Prendre en compte d'autres traits opérationnels :
- Fonction syntaxique : sujet, objet, autre
- Degré d'informativité du maillon
- Type de reprise : directe, indirecte, pronominale
- Interdistance (Rousier-Vercruyssen & Landragin, 2019)
- Instabilité des chaînes (Rousier-Vercruyssen & Landragin, 2019)

② Combiner plusieurs traits opérationnels

③ Projeter d'autres types de référent :
- concret - abstrait, générique - spécifique
- distinctions plus fines : individus particuliers - humain collectifs

④ Reproduire la même analyse sur d'autres corpus : Democrat (Landragin, 2015) et E-Calm (Garcia-Debanc et al., 2017)
- type textuel
- niveaux d'expertise rédactionnelle

Conclusion et perspectives

- Méthode proposée valide pour mettre au jour des classes de chaînes

Prochaines étapes :

① Prendre en compte d'autres traits opérationnels :
- Fonction syntaxique : sujet, objet, autre
- Degré d'informativité du maillon
- Type de reprise : directe, indirecte, pronominale
- Interdistance (Rousier-Vercruyssen & Landragin, 2019)
- Instabilité des chaînes (Rousier-Vercruyssen & Landragin, 2019)

② Combiner plusieurs traits opérationnels

③ Projeter d'autres types de référent :
- concret - abstrait, générique - spécifique
- distinctions plus fines : individus particuliers - humain collectifs

④ Reproduire la même analyse sur d'autres corpus : Democrat (Landragin, 2015) et E-Calm (Garcia-Debanc et al., 2017)
- type textuel
- niveaux d'expertise rédactionnelle


Référence
Chaînes
Nature
Nature
Maillons
Référence
Nature
Maillons
Nature
Nature
Longueur
Longueur
Maillons
Merci
Longueur
Entité
Entité
Référent
Nature
Entité
Merci
Chaînes
Merci
Merci
Chaînes
Référence
Référent
Nature
Référent
Longueur
Nature
Chaînes
Entité
Merci
Entité
Longueur
Longueur
Chaînes
Longueur
Chaînes

Ariel, M. (2001). Accessibility theory : An overview. *Text representation : Linguistic and psycholinguistic aspects*, *8*, 29–87.

Brzinsky-Fay, C., Kohler, U., & Luniak, M. (2006). Sequence analysis with stata. *The Stata Journal*, *6*(4), 435–460.

Charolles, M. (2002). *La référence et les expressions référentielles en français*. Editions Ophrys.

Corblin, F. (1987). *Indéfini, défini et démonstratif droz.* Genève.

Corblin, F. (1995). *Les formes de reprise dans le discours. anaphores et chaînes de référence*. Presses Universitaires de Rennes.

Cornish, F. (1998). Les chaînes topicales : leur rôle dans la gestion et la structuration du discours. *Cahiers de grammaire*, *23*(1), 9–40.

Cornish, F. (2000). L'accessibilité cognitive des référents, le centrage d'attention, et la structuration du discours : une vue d'ensemble. *Verbum*, *22*(1), 7–30.

De Marneffe, M.-C., Recasens, M., & Potts, C. (2015). Modeling the lifespan of discourse entities with application to coreference resolution. *Journal of Artificial Intelligence Research*, *52*, 445–475.

Federzoni, S., Ho-Dac, L.-M., & Rebeyrolle, J. (2020, juillet). Les chaînes topicales dans la ressource ANNODIS. In *CMLF2020 : 7e Congrès*

Mondial de Linguistique Française. Montpellier, France. Consulté sur `https://hal.archives-ouvertes.fr/hal-02890989`

Fossard, M., Achim, A. M., Rousier-Vercruyssen, L., Gonzalez, S., Bureau, A., & Champagne-Lavau, M. (2018). Referential choices in a collaborative storytelling task : Discourse stages and referential complexity matter. *Frontiers in psychology*, *9*, 176.

Gabadinho, A., Ritschard, G., Studer, M., & Müller, N. S. (2009). Mining sequence data in r with the traminer package : A user's guide. *Geneva : Department of Econometrics and Laboratory of Demography, University of Geneva*.

Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M., & Rebeyrolle, J. (2017). Vers l'annotation discursive de textes d'élèves. *Corpus*(16).

Gundel, J. K., Hedberg, N., & Zacharski, R. (2019). Cognitive status and the form of referring expressions in discourse. In J. Gundel & B. Abbott (Eds.), *The oxford handbook of reference* (pp. 67–99). New York : Oxford University Press.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in english*. Routledge.

Kaiser, E., & Fedele, E. (2019). Reference resolution : A psycholonguistic perspective. In J. Gundel & B. Abbott (Eds.), *The oxford handbook*

of reference (pp. 309–336). New York : Oxford University Press.
Kleiber, G. (1994). Anaphores et pronoms. *Louvain-la-Neuve, Duculot*.
Kleiber, G. (2002). Marqueurs référentiels et théorie du centrage. *Linx. Revue des linguistes de l'université Paris X Nanterre*(47), 107–119.
Landragin, F. (2015). Description, modélisation et détection automatique des chaînes de référence (democrat). *Bulletin de l'AFIA*(92), 11–15.
Landragin, F. (2018, juillet). Étude de la référence et de la coréférence : rôle des petits corpus et observations à partir du corpus mc4. *Corpus*, *18*.
Longo, L. (2013). *Vers des moteurs de recherche "intelligents" : un outil de détection automatique de thèmes. méthode basée sur l'identification automatique des chaînes de référence* (Theses). Université de Strasbourg. (thèse réalisée dans le cadre d'une CIFRE (convention industrielle de formation par la recherche) avec la société RBS (Ready Business System))
Mitkov, R. (2014). *Anaphora resolution*. Routledge.
Obry, V., Glikman, J., Guillot-Barbance, C., & Pincemin, B. (2017). Les chaînes de référence dans les récits brefs en français : étude diachronique (xiiie-xvie s.). *Langue française*(3), 91–110.

Poesio, M., Pradhan, S., Recasens, M., Rodriguez, K., & Versley, Y. (2016). Annotated corpora and annotation tools. , 97–140.

Péry-Woodley, M.-P., Afantenos, S., Ho-Dac, L.-M., & Asher, N. (2011). La ressource annodis, un corpus enrichi d'annotations discursives. *Traitement Automatique des Langues*, *52*(3), 71-101.

Quiniou, S., Cellier, P., Charnois, T., & Legallois, D. (2012, juin). Fouille de données pour la stylistique : cas des motifs séquentiels émergents. *Journées Internationales d'Analyse Statistique des Données Textuelles (JADT'12)*, 821-833. Consulté sur `https://hal.archives-ouvertes.fr/hal-00675586`

Recasens, M., & Hovy, E. (2010). Coreference resolution across corpora : Languages, coding schemes, and preprocessing information. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1423–1432).

Roberts, C. (2019). Contextual influences on reference. In J. Gundel & B. Abbott (Eds.), *The oxford handbook of reference* (pp. 260–280). New York : Oxford University Press.

Robette, N. (2011). *Explorer et décrire les parcours de vie : les typologies de trajectoires*. CEPED. Consulté sur

https://halshs.archives-ouvertes.fr/halshs-01016125
Rousier-Vercruyssen, L., & Landragin, F. (2019). Interdistance et instabilité au sein des chaînes de référence : indices textuels ? *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*(25).
Salazar Orvig, A. (2019). Reference and referring expressions in first language acquisition. In J. Gundel & B. Abbott (Eds.), *The oxford handbook of reference* (pp. 283–308). New York : Oxford University Press.
Salles, M. (2015). Chaînes de référence : la deuxième mention. l'exemple des entités inanimées dans les narrations littéraires. *Travaux de linguistique*(2), 111–133.
Schnedecker, C. (2005). Les chaînes de référence dans les portraits journalistiques : éléments de description. *Travaux de linguistique*(2), 85–133.
Schnedecker, C. (2014). Chaînes de référence et variations selon le genre. *Langages*(3), 23–42.
Schnedecker, C., & Landragin, F. (2014, septembre). Les chaînes de

référence : présentation. *Langages*(195), 3–22.

Schnedecker, C., & Longo, L. (2012, juillet). Impact des genres sur la composition des chaînes de référence : le cas des faits divers. In *3ième congrès mondial de linguistique française* (p. 1957-1972). Lyon, France.

Vogels, J., Krahmer, E., & Maes, A. (2013). When a stone tries to climb up a slope : the interplay between lexical and perceptual animacy in referential choices. *Frontiers in psychology*, *4*, 154.

Walker, M. A., Joshi, A. K., & Prince, E. F. (1998). *Centering theory in discourse*. Oxford University Press.

Wilkens, R., Oberle, B., Landragin, F., & Todirascu, A. (2020). French coreference for spoken and written language. In *Language Resources and Evaluation Conference (LREC 2020)* (p. 80-89). Marseille, France.