Mining annotated corpora for coreference chains patterns

Silvia Federzoni, Lydia-Mai Ho-Dac and Cécile Fabre
CLLE, CNRS & University of Toulouse

DisCorX 2021
Discourse in corpus and experimental data:
Bridging the methodological gap

18th March, 2021

Plan

Coreference chains

Coreference chain

After George W. Bush is sworn in, **Bill Clinton** will head to New York. **The President** has said **he** and **his** wife, now a New York senator will spend weekends at their house in Chappaqua. **Mr. Clinton** will also spend time at **his** presidential library in Arkansas. **He** says **he** will come to Washington, "every now and then."

Coreference chains: discourse structures that group together several clauses around a common **referent**

Coreference chains

Coreference chain

After George W. Bush is sworn in, **Bill Clinton** will head to New York. **The President** has said **he** and **his** wife, now a New York senator will spend weekends at their house in Chappaqua. **Mr. Clinton** will also spend time at **his** presidential library in Arkansas. **He** says **he** will come to Washington, "every now and then."

Coreference chains: discourse structures that group together several clauses around a common **referent**

Referent: extra-linguistic entity
Entity types: human, organisation, event, generic, specific

Coreference chains

Coreference chain

After George W. Bush is sworn in, **Bill Clinton** will head to New York. **The President** has said **he** and **his** wife, now a New York senator will spend weekends at their house in Chappaqua. **Mr. Clinton** will also spend time at **his** presidential library in Arkansas. **He** says **he** will come to Washington, "every now and then."

Mentions: linguistics expressions signalling the **referent** and carrying an **instructional value**

Coreference chains

Coreference chain

After George W. Bush is sworn in, **Bill Clinton** will head to New York. **The President** has said **he** and **his** wife, now a New York senator will spend weekends at their house in Chappaqua. **Mr. Clinton** will also spend time at **his** presidential library in Arkansas. **He** says **he** will come to Washington, "every now and then."

Mentions

Mentions: their succession contributes to the creation of **cohesive ties** (Halliday & Hasan, 1976)

Coreference chains: fundamental mechanism in **the organisation** and **interpretation of discourse**

Objective

Typology of coreference chains

Identify **classes** of chains according to their **discursive role**.

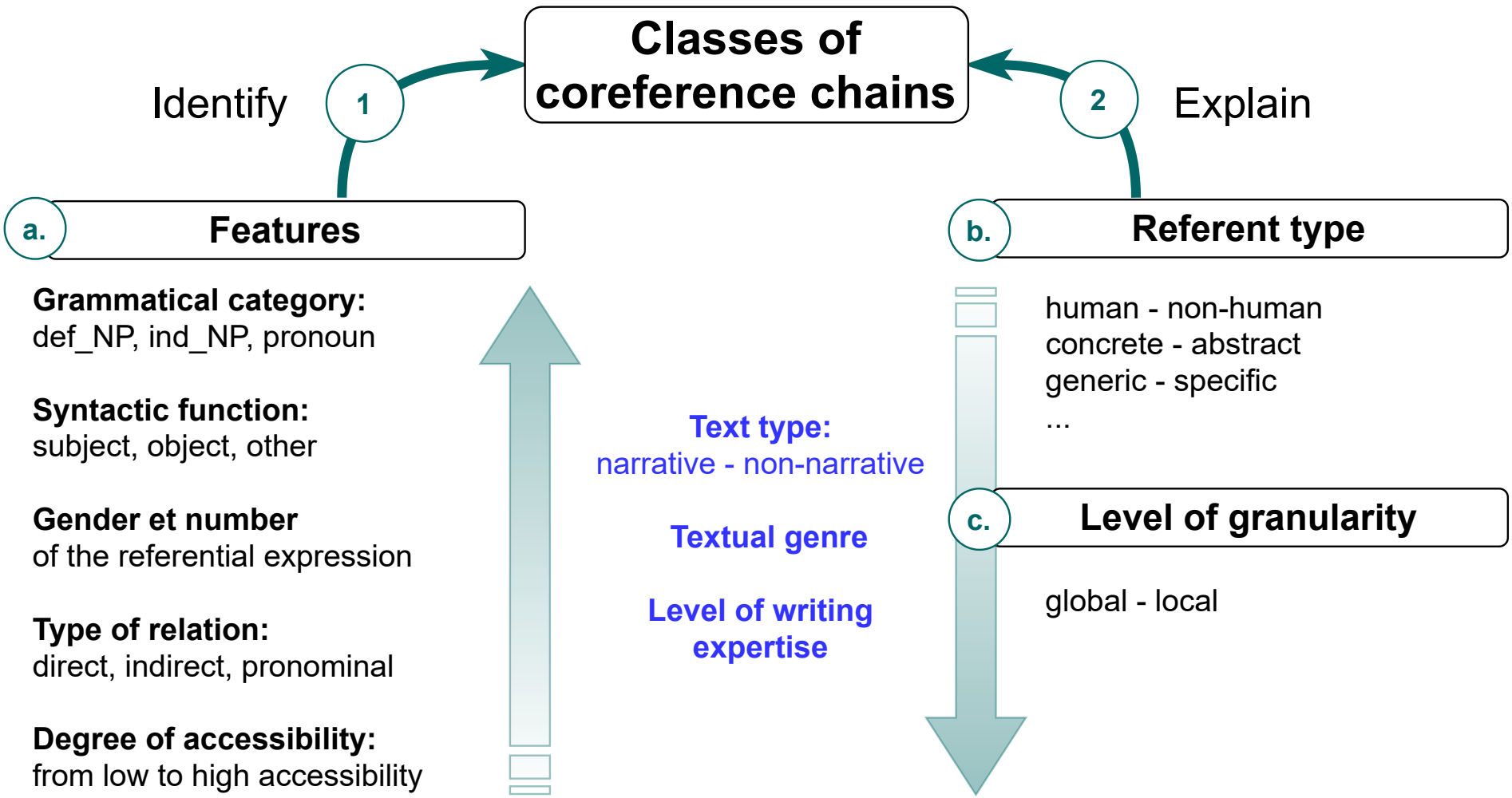
Provide a **systematic** and **exhaustive** description of:

- the **composition**
- the **variety**
- the **complexity**

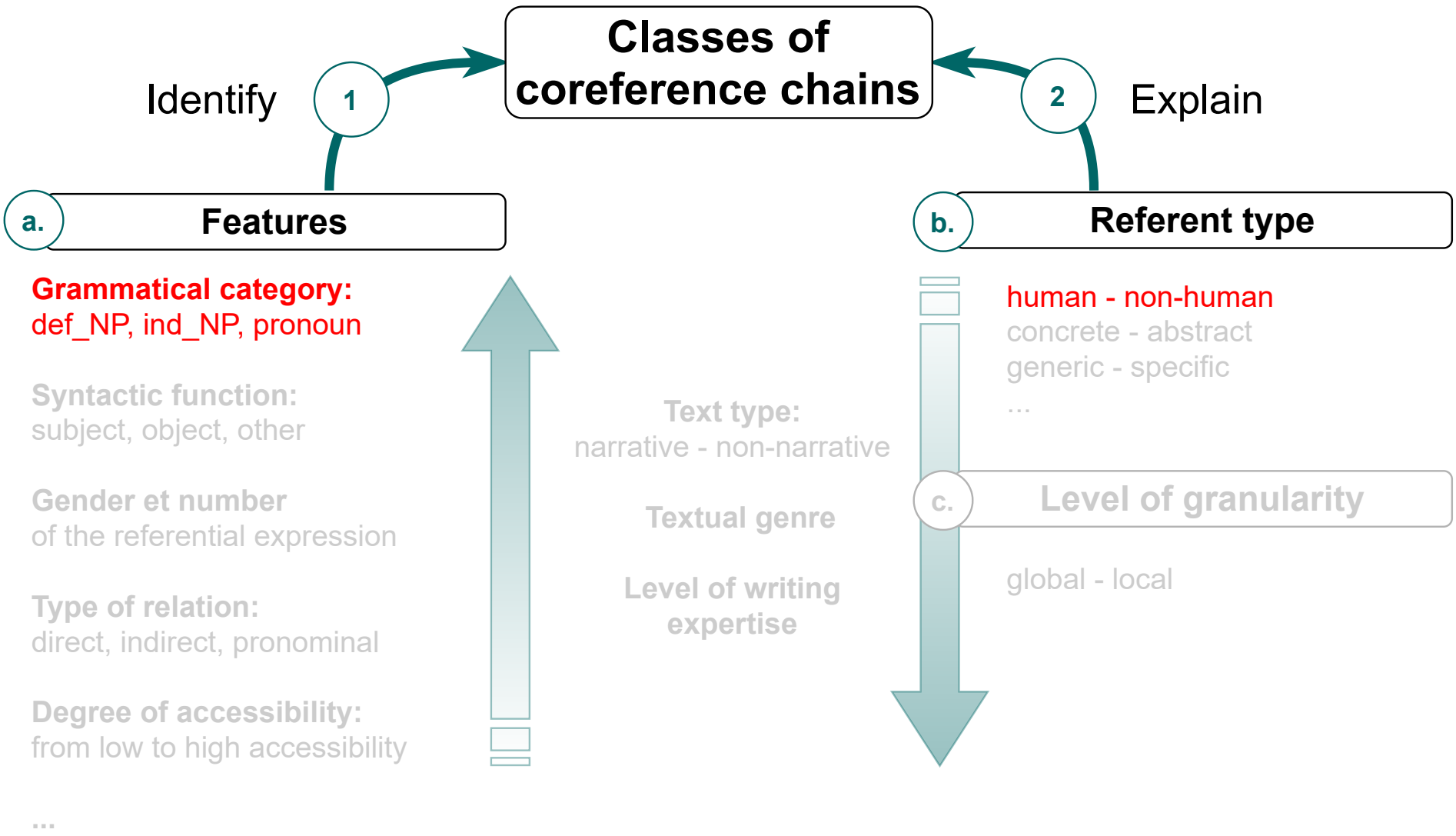
of the coreference chains

Propose an **automatic method** based on the **systematic** analysis of a large range of features

Method of description



Method of description



Plan

The AnnoDis corpus (Péry-Woodley, Afantenos, Ho-Dac, & Asher, 2011)

- A written French corpus annotated for **high level coreference chains** (topical chains)

  - long structured texts (87 texts, 7655 words/text)
  - fully annotated (581 chains, 3456 mentions)
  - non-narrative texts
  - 3 textual genres (reports, scientific articles, encyclopedic texts)

Towards a typology of coreference chains

Objective: describe the composition of coreference chains

Identify classes of coreference chains by observing the most frequent **sequences** of mentions.

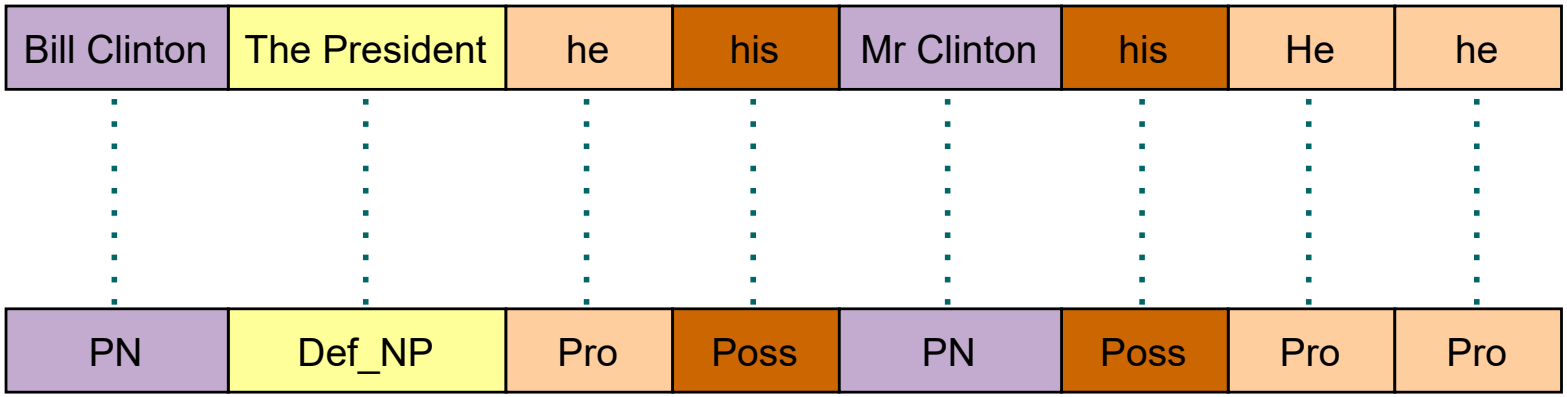
Towards a typology of coreference chains

Objective: describe the composition of coreference chains

Identify classes of coreference chains by observing the most frequent **sequences** of mentions.

After George W. Bush is sworn in, **Bill Clinton** will head to New York. **The President** has said **he** and **his** wife, now a New York senator will spend weekends at their house in Chappaqua. **Mr. Clinton** will also spend time at **his** presidential library in Arkansas. **He** says **he** will come to Washington, "every now and then."

Towards a typology of coreference chains



Grammatical category	#	%
Def_NP	1198	34.66
Dem_NP	272	7.87
Ind_NP	126	3.65
NoDet_NP	64	1.85
PN	442	12.79
Poss	182	5.27
Pro	1026	29.69
Other	146	4.22
Total	3456	100

Feature used:
grammatical category

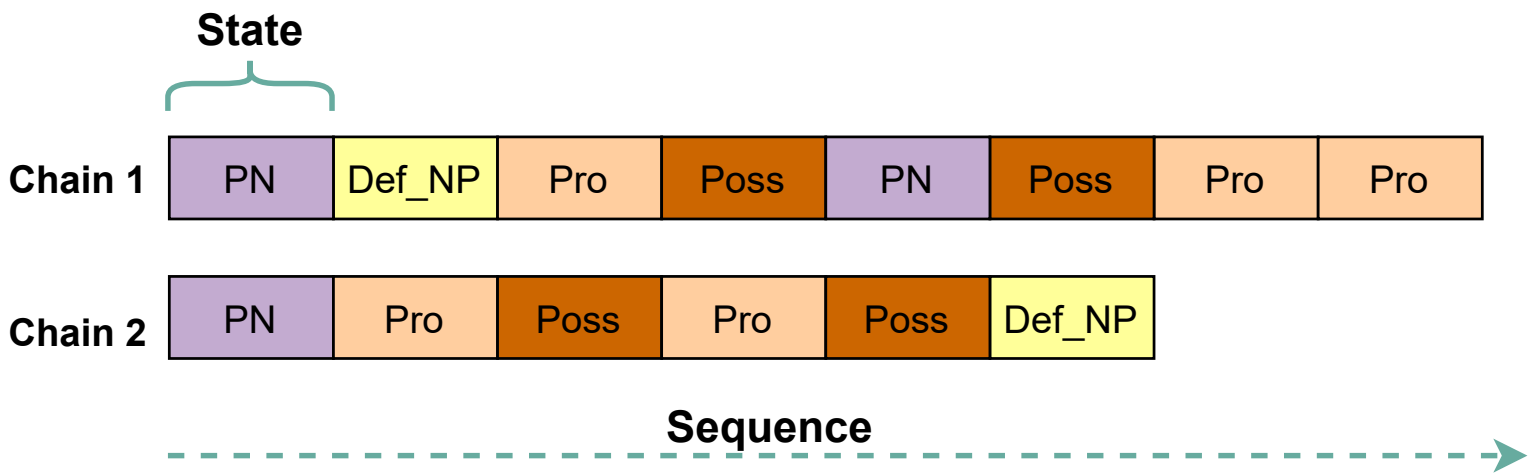
- Frequency of grammatical categories: strongest variations
- Richest information on chain typology: degree of homogeneity

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

Method - sequence analysis (Quiniou, Cellier, Charnois, & Legallois, 2012)

Identify regularities, similarities, build typologies of "typical sequences" (Robette, 2011)

Sequences: ordered list of **states** or **events** (Brzinsky-Fay, Kohler, & Luniak, 2006)



Method - sequence analysis

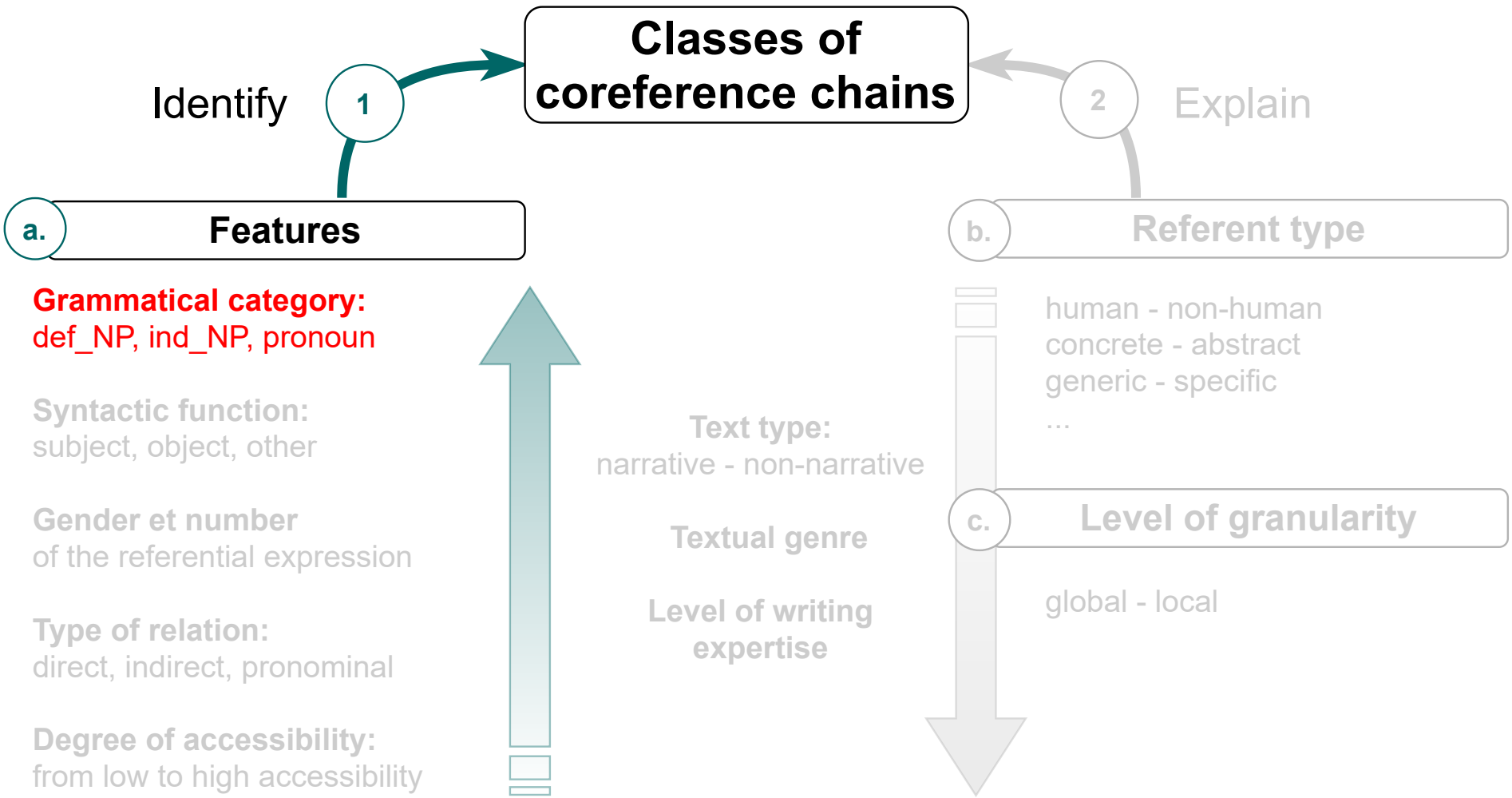
Method - sequence analysis (Quiniou et al., 2012)

Identify regularities, similarities, build typologies of "typical sequences" (Robette, 2011)

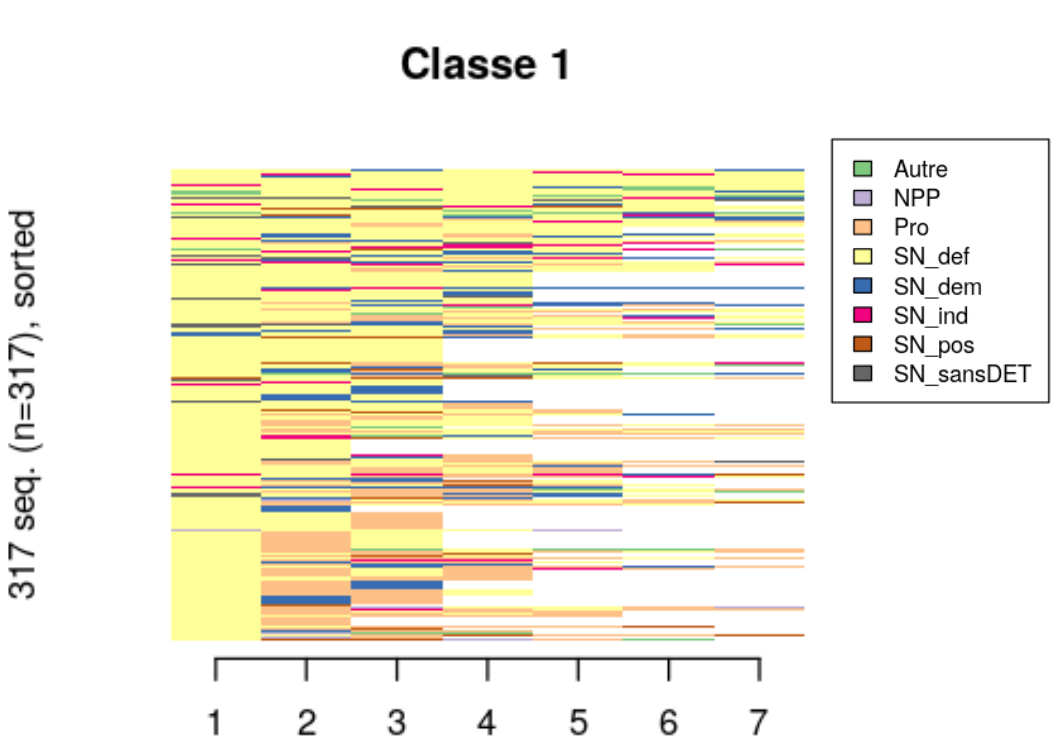
TraMineR (Gabadinho, Ritschard, Studer, & Müller, 2009):

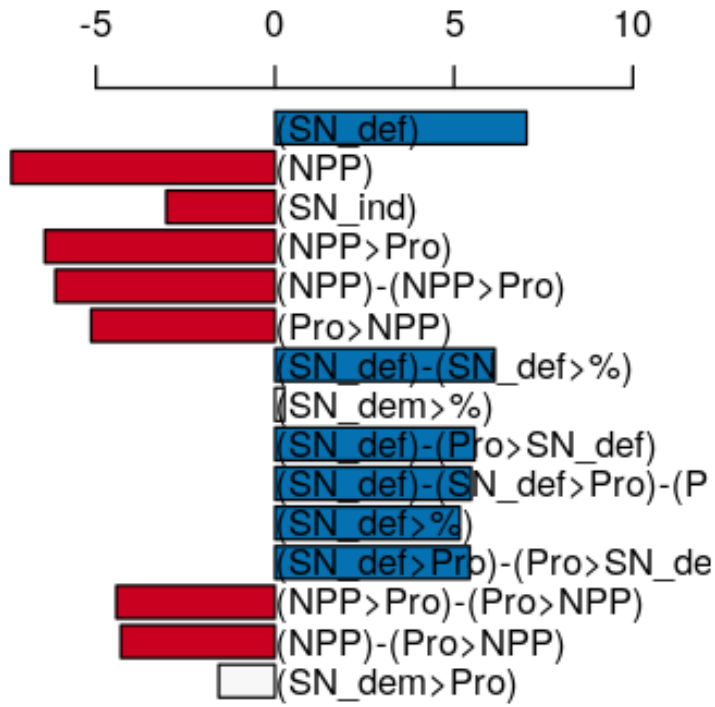
- Clustering method: hierarchical clustering ('Ward' linkage criterion)
- Sequence length: from 2 to 7 mentions (73.15% chains have less than 7 mentions)
- Number of clusters: 3 clusters

Identify classes



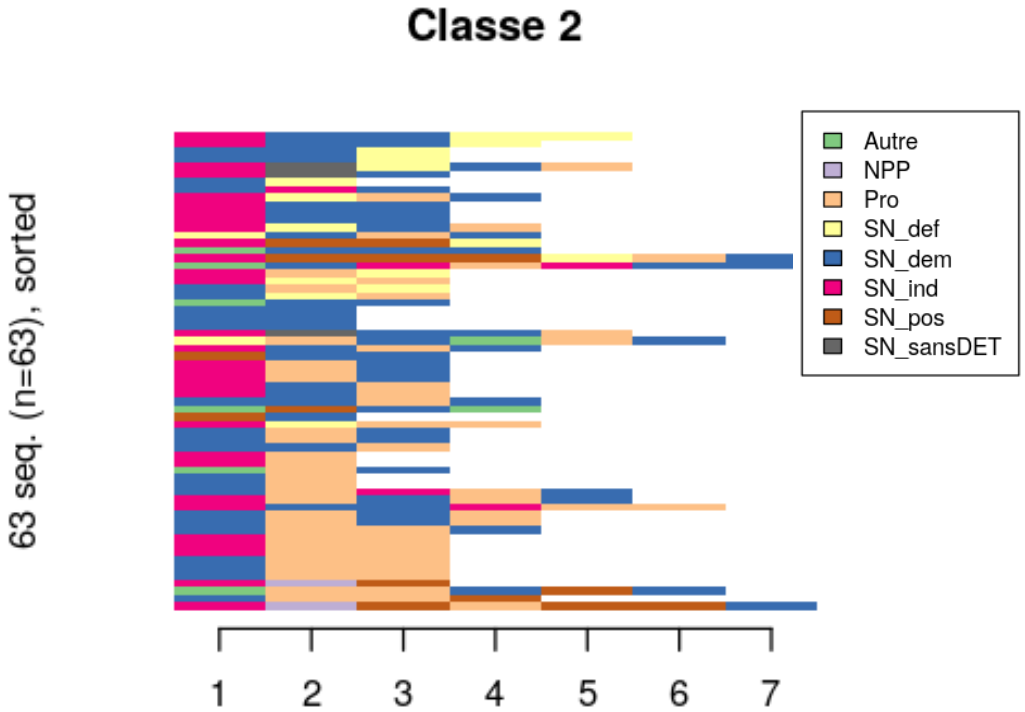
TraMineR cluster 1

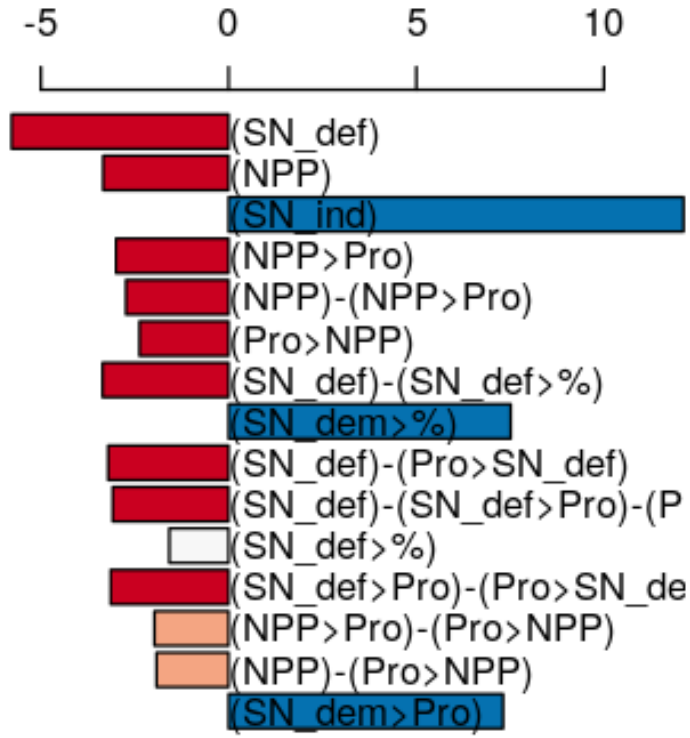




le vignoble champenois ***{the Champagne vineyard}*** *s'étendait sur quelques [...]* *le vignoble* ***{the vineyard}*** *connaît [...] Après les fléau du phylloxéra et de la Grande guerre,* *le vignoble* ***{the vineyard}*** *s'est réduit à 12 000 hectares. Aujourd'hui, en 2007,* *le vignoble champenois*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

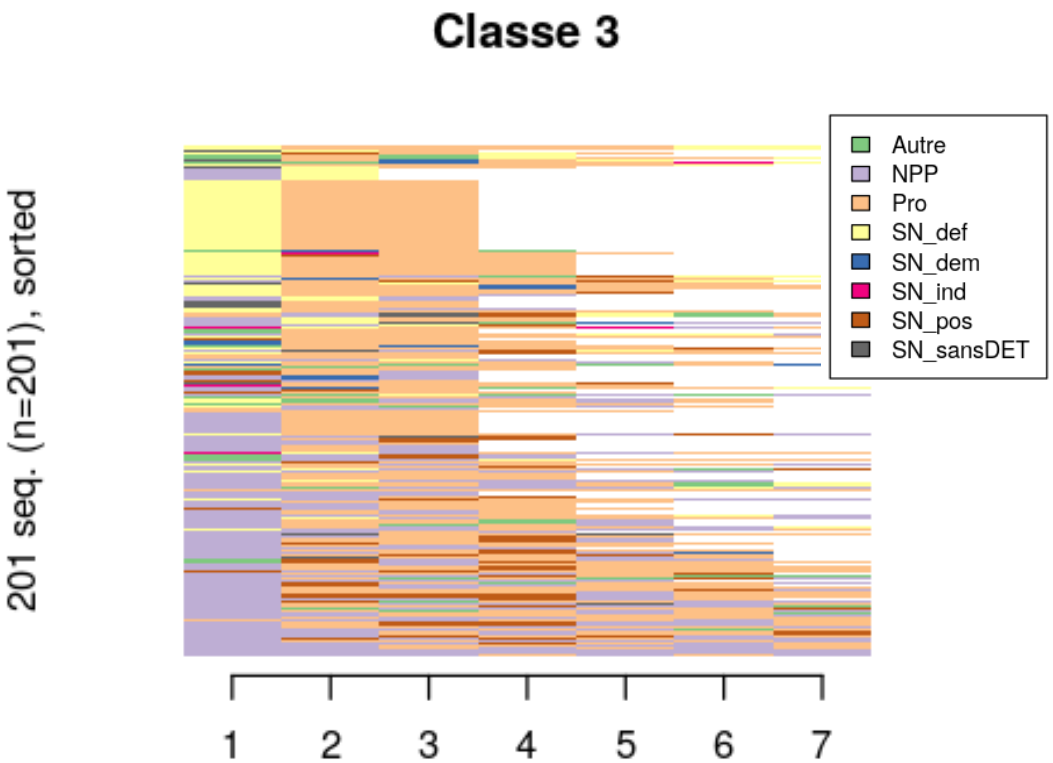
TraMineR cluster 2

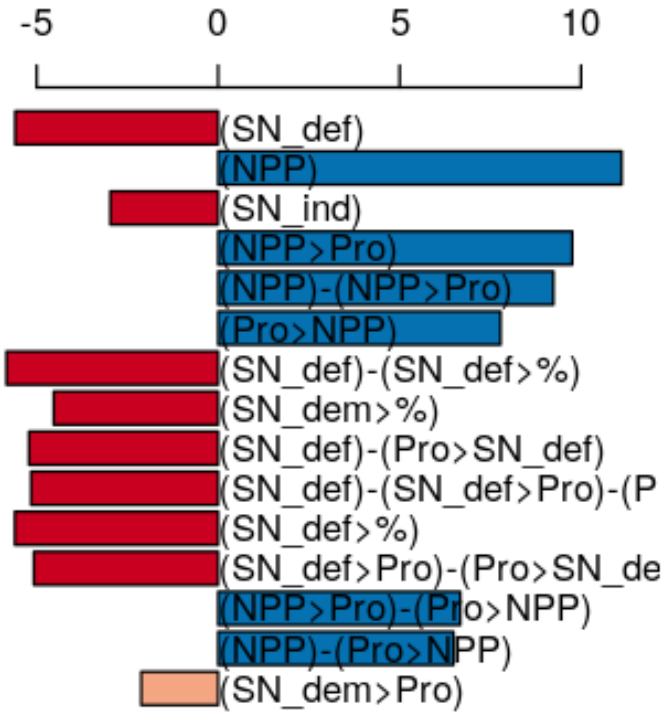




[...] « le " communicatif " présente ainsi fréquemment une connotation oppositionnelle ***{an oppositional connotation}*** *(sinon contradictoire) avec le linguistique.* ***Ceci {This}*** *est particulièrement crucial lorsqu'on traite de la " compétence de communication " [...]* ***Cette dichotomie {This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

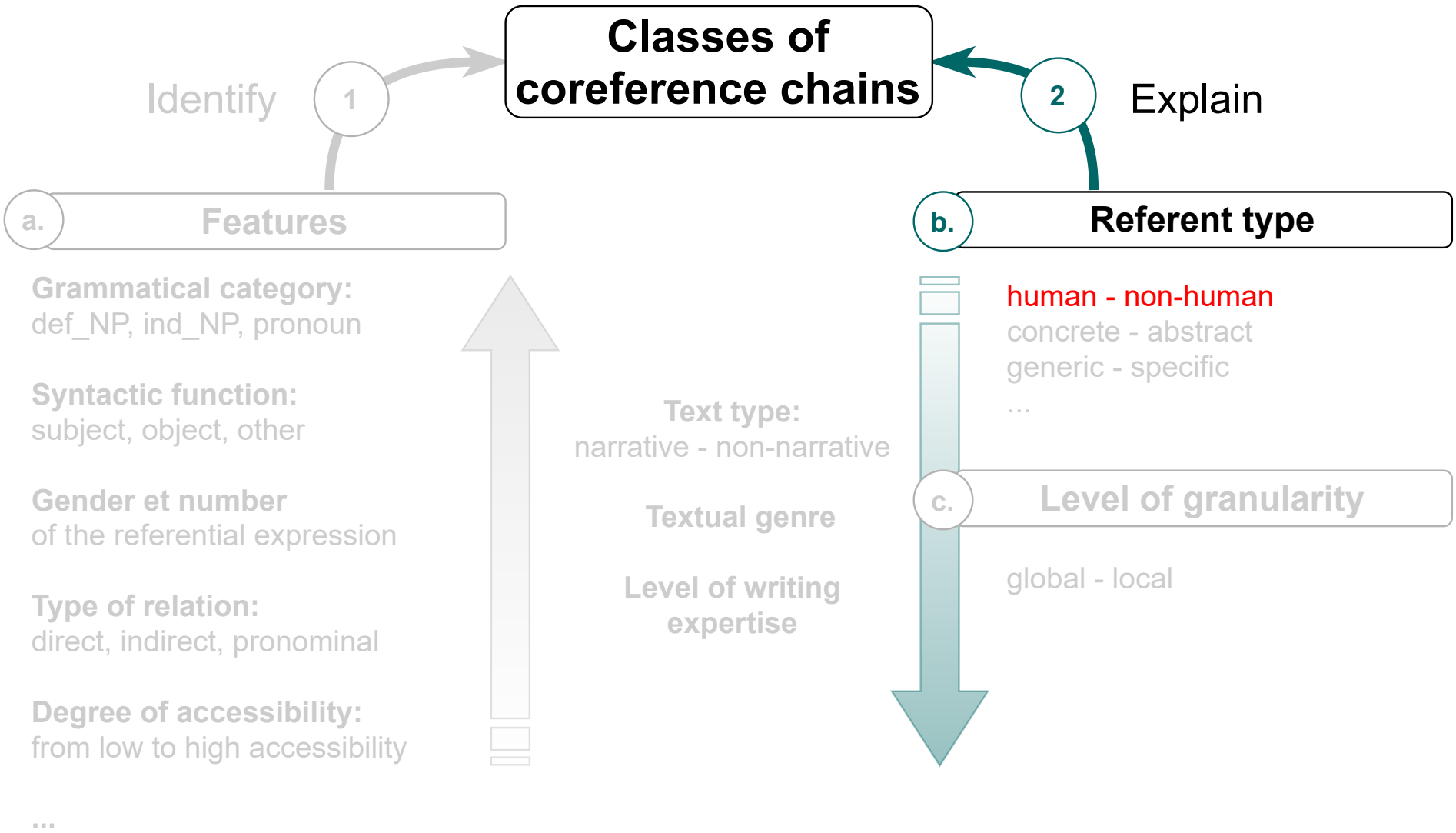
TraMineR cluster 3



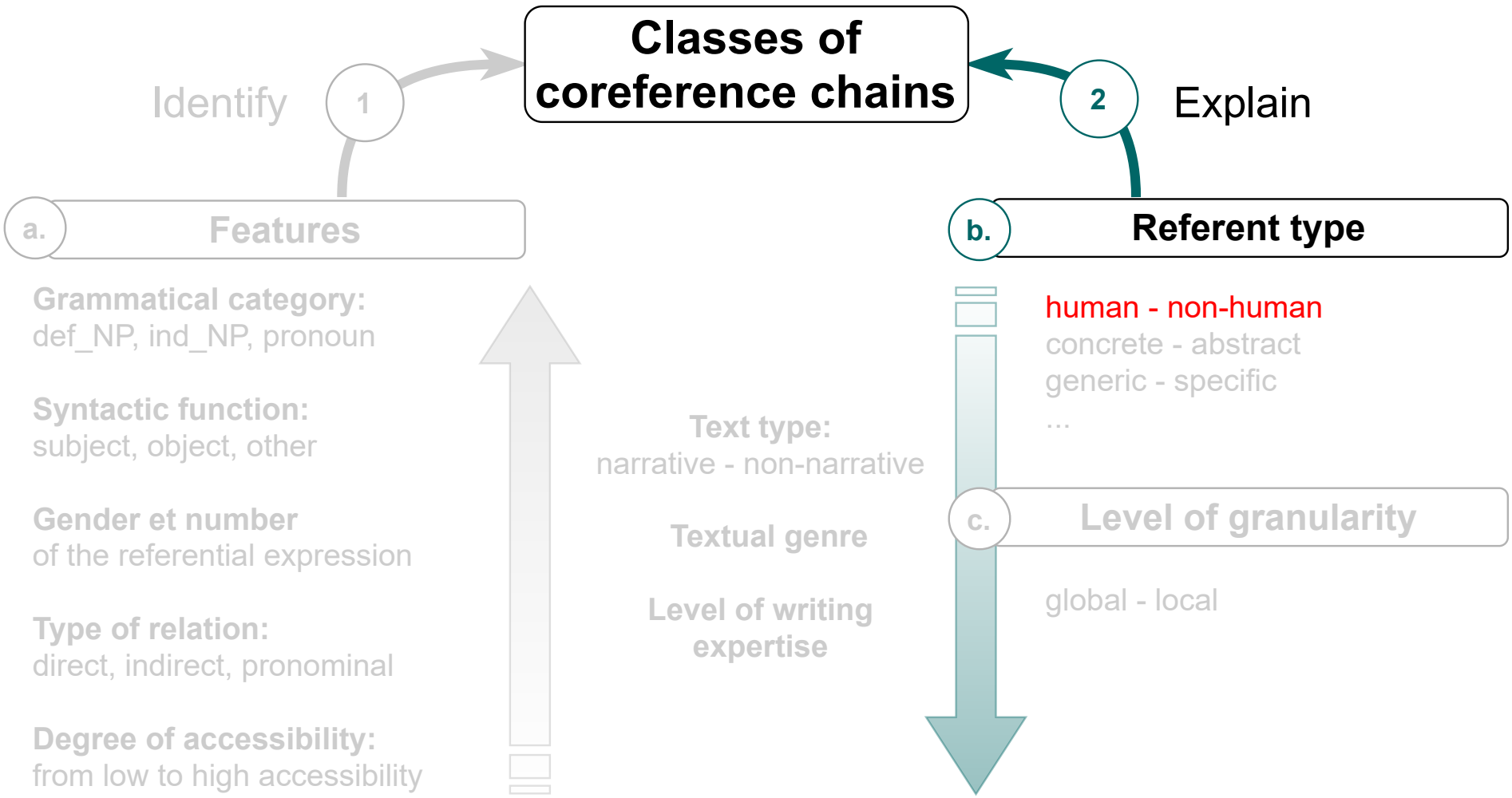


Chez F. de Saussure {F. de Saussure}, *l'analogie [...],* il {he} *pose que les facteurs de trouble [...]. Pour* lui {him}, *cette tendance à l'irrégularité est heureusement contrebalancée par l'analogie [...]. Comme H. Paul,* il {he} *ramène le concept au calcul de l'équation de la quatrième proportionnelle.*

Explain TraMineR clusters



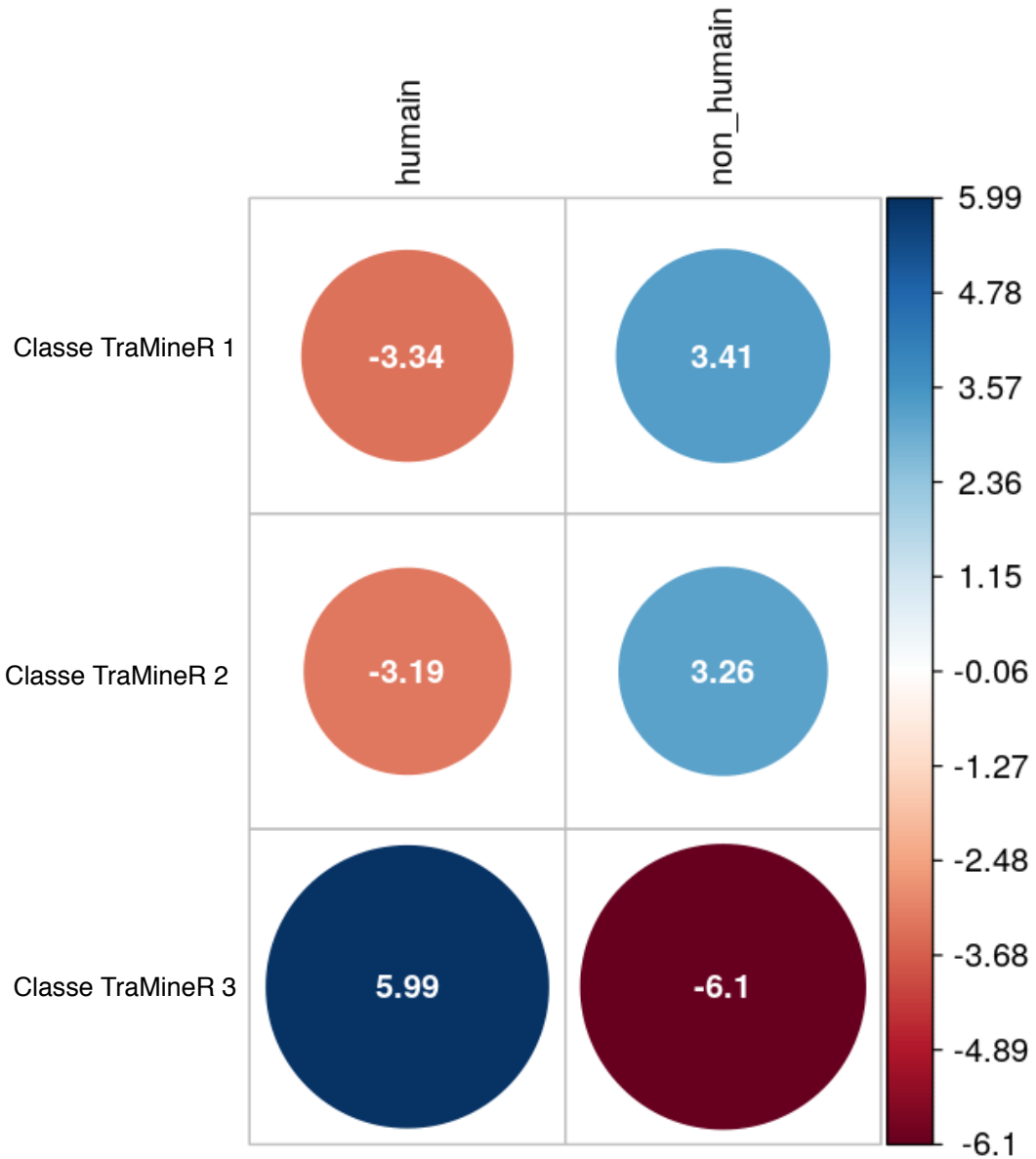
Explain TraMineR clusters



Referent type : semi-automatic annotation

Project referent type

Conclusion: significant relationship between classes and referent type (chi-squared test)



Analyse TraMineR clusters

	Composition	**Human**		**Non-human**	
		#	**%**	**#**	**%**
TraMineR cluster 1	Def_NP++	119	37.54	**198**	**62.46**
TraMineR cluster 2	Ind_NP Dem_NP	14	22.22	**49**	**77.78**
TraMineR cluster 3	Proper_N Pro	**163**	**81.09**	38	18.91

Analyse TraMineR clusters

	Composition	Human		Non-human	
		#	%	#	%
TraMineR cluster 1	Def_NP++	119	37.54	**198**	**62.46**
TraMineR cluster 2	Ind_NP Dem_NP	14	22.22	**49**	**77.78**
TraMineR cluster 3	Proper_N Pro	**163**	**81.09**	38	18.91

Pattern : Def_NP{1:7} (24 RC)

Non-human referents

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

Analyse TraMineR clusters

	Composition	Human		Non-human	
		#	%	#	%
TraMineR cluster 1	Def_NP++	119	37.54	**198**	**62.46**
TraMineR cluster 2	Ind_NP Dem_NP	14	22.22	**49**	**77.78**
TraMineR cluster 3	Proper_N Pro	**163**	**81.09**	38	18.91

Pattern : Def_NP{1:7} (22 RC)

Human referents

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

Analyse TraMineR clusters

	Composition	Human		Non-human	
		#	%	#	%
TraMineR cluster 1	Def_NP++	119	37.54	**198**	**62.46**
TraMineR cluster 2	Ind_NP Dem_NP	14	22.22	**49**	**77.78**
TraMineR cluster 3	Proper_N Pro	**163**	**81.09**	38	18.91

Pattern : Def_NP{1:2} > Pro{1:6} (42 RC)

Human referents

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

Analyse TraMineR clusters

	Composition	Human		Non-human	
		#	%	#	%
TraMineR cluster 1	Def_NP++	119	37.54	**198**	**62.46**
TraMineR cluster 2	Ind_NP Dem_NP	14	22.22	**49**	**77.78**
TraMineR cluster 3	Proper_N Pro	**163**	**81.09**	38	18.91

Pattern : Def_NP{1:2} > Pro{1:6} (24 RC)

Non-human referents

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

Plan

Conclusion and perspectivess

- Method proposed is successful for identifying classes of coreference chains

Next steps:

1. Take into account other features:
   - Syntactic function: subject, object, other
   - Gender and number
   - Type of relation: direct, indirect, pronominal
   - Inter-distance (Rousier-Vercruyssen & Landragin, 2019)
   - Chains instability (Rousier-Vercruyssen & Landragin, 2019)
2. Combine different features
3. Explain classes in according to other referent types:
   - concrete - abstract
   - generic - specific
4. Apply this method to other corpora: Democrat (Landragin, 2015) and E-Calm corpora (Garcia-Debanc, Ho-Dac, Bras, & Rebeyrolle, 2017)
   - text type
   - level of writing expertise

Conclusion and perspectivess

- Method proposed is successful for identifying classes of coreference chains

Next steps:

1. Take into account other features:
   - Syntactic function: subject, object, other
   - Gender and number
   - Type of relation: direct, indirect, pronominal
   - Inter-distance (Rousier-Vercruyssen & Landragin, 2019)
   - Chains instability (Rousier-Vercruyssen & Landragin, 2019)
2. Combine different features
3. Explain classes in according to other referent types:
   - concrete - abstract
   - generic - specific
4. Apply this method to other corpora: Democrat (Landragin, 2015) and E-Calm corpora (Garcia-Debanc et al., 2017)
   - text type
   - level of writing expertise

Conclusion and perspectivess

- Method proposed is successful for identifying classes of coreference chains

Next steps:

1. Take into account other features:
   - Syntactic function: subject, object, other
   - Gender and number
   - Type of relation: direct, indirect, pronominal
   - Inter-distance (Rousier-Vercruyssen & Landragin, 2019)
   - Chains instability (Rousier-Vercruyssen & Landragin, 2019)
2. Combine different features
3. Explain classes in according to other referent types:
   - concrete - abstract
   - generic - specific
4. Apply this method to other corpora: Democrat (Landragin, 2015) and E-Calm corpora (Garcia-Debanc et al., 2017)
   - text type
   - level of writing expertise

Conclusion and perspectivess

- Method proposed is successful for identifying classes of coreference chains

Next steps:

1. Take into account other features:
   - Syntactic function: subject, object, other
   - Gender and number
   - Type of relation: direct, indirect, pronominal
   - Inter-distance (Rousier-Vercruyssen & Landragin, 2019)
   - Chains instability (Rousier-Vercruyssen & Landragin, 2019)
2. Combine different features
3. Explain classes in according to other referent types:
   - concrete - abstract
   - generic - specific
4. Apply this method to other corpora: Democrat (Landragin, 2015) and E-Calm corpora (Garcia-Debanc et al., 2017)
   - text type
   - level of writing expertise

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.

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).

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, *21*(2), 203–225.

Landragin, F. (2015). Description, modélisation et détection automatique des chaînes de référence (democrat). *Bulletin de l'AFIA*(92), 11–15.

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.

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, June). Fouille de données pour la stylistique: cas des motifs séquentiels émergents. *Journées Internationales d'Analyse Statistique des Données Textuelles (JADT'12)*, 821-833. Retrieved from `https://hal.archives-ouvertes.fr/hal-00675586`

Robette, N. (2011). *Explorer et décrire les parcours de vie: les typologies de trajectoires*. CEPED. Retrieved from `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).