Coreference Chains Categorization by Sequence Clustering

Silvia Federzoni and **Lydia-Mai Ho-Dac** and **Cécile Fabre**
CLLE, CNRS & University of Toulouse
Toulouse, France
{silvia.federzoni, hodac, cecile.fabre}@univ-tlse2.fr

Abstract

The diversity of coreference chains is usually tackled by means of global features (length, types and number of referring expressions, distance between them, etc.). In this paper, we propose a novel approach that provides a description of their composition in terms of sequences of expressions. To this end, we apply sequence analysis techniques to bring out the various strategies for introducing a referent and keeping it active throughout discourse. We discuss a first application of this method to a French written corpus annotated with coreference chains. We obtain clusters that are linguistically coherent and interpretable in terms of reference strategies and we demonstrate the influence of text genre and semantic type of the referent on chain composition.

1 Introduction

Coreference chains are discourse structures built upon a set of referential expressions (or mentions) denoting a common discourse entity (Corblin, 1995; Poesio et al., 2016a). They provide a fundamental mechanism for text interpretation and contribute to cohesion between clauses (Halliday and Hasan, 1976). Linguistic analysis and automatic detection of coreference chains are still a challenge, due to their complexity and the diversity of their composition (Recasens et al., 2011). In particular, a large diversity of linguistic expressions may be used to mention a discourse entity, such as proper nouns, pronouns, possessives, definite or demonstrative noun phrases, in various syntactic positions. This diversity is usually tackled by studying the global characteristics of the coreference chains (Nedoluzhko and Lapshinova-Koltunski, 2016; Kunz and Lapshinova-Koltunski, 2015), e.g. the number of mentions, their type, the distance between them or by focusing only on the characteristics of the first two mentions.

We propose a method that complements these approaches by considering chains as linear sequences of mentions. This makes it possible to identify categories of coreference chains in terms of chaining, which gives new insights for linguistic characterization and provides knowledge about variational dimensions that can help to improve coreference resolution systems (Lapshinova-Koltunski and Kunz, 2020).

We resort to sequence analysis techniques which are generally used in social sciences to build typologies of "typical sequences" to study life-course trajectories, family histories, professional career paths. In such studies, sequences are used to model the chronology of states or events (Studer and Ritschard, 2014; Brzinsky-Fay et al., 2006). We apply these techniques to coreference chains categorization by considering each mention as a state in the chain chronology, and by characterizing mentions with features traditionally used for studying referring expressions and resolving coreference such as mention types and syntactic functions, relations between mentions, degree of accessibility (Ariel, 2001; Walker, 2000; Poesio et al., 2016b).

This paper presents a first step in which only mention types are considered. We apply sequence analysis to a French written corpus annotated with topical chains. We discuss the results in two steps: first we show that the obtained clusters are linguistically interpretable; then we demonstrate that text genre and semantic nature of the referent influence the coreference chain composition.

2 Description of the experiment

2.1 Data

We use a written French corpus, the AnnoDis corpus, annotated with topical chains, which correspond to coreference chains that are built upon a prominent or topical element (Asher et al., 2017). It provides 581 chains annotated in full long structured texts and organized in three sub-corpora pertaining to various non-narrative genres : geopoli-

Proceedings of the 2nd Workshop on Computational Approaches to Discourse, pages 52–57
November 10–11, 2021.

tics reports (GEOP), linguistics articles (LING) and encyclopedic texts (WIK2) (Péry-Woodley et al., 2011)[1], see Table 1.

	Words	Chains	Mentions
GEOP	266,000	234	1,125
LING	169,000	87	478
WIK2	231,000	260	1,853
AnnoDis	666,000	581	3,456

Table 1: Number of words, coreference chains and mentions in the AnnoDis corpus.

Among the wide range of linguistic features that can be considered for characterizing mentions this first experiment focuses on the grammatical category only, which provides the simplest information on chain typology. Each mention of the coreference chains is labeled by one of the 8 types listed in Table 2 (definite, demonstrative or indefinite NPs, NPs without determiners, proper nouns, possessives, pronouns, other).

Mention type	**GEOP**	**LING**	**WIK2**	**AnnoDis**
Def. NP	499	185	514	1,198
Dem. NP	115	49	108	272
Ind. NP	56	21	49	126
NoDet NP	14	6	44	64
Proper N.	63	41	338	442
Possessives	59	9	114	182
Pronouns	288	142	596	1,026
Other	31	25	90	146
Total	1,125	1,853	478	3,456

Table 2: Distribution of mention types in the AnnoDis corpus and his sub-corpora

2.2 Sequence analysis

We carry out the sequence analysis using the `TraMineR` toolbox (Gabadinho et al., 2009, 2011), that brings together various features designed to handle sequential data. This allows us to identify and visualize sequences, but also to implement clustering and statistical methods to identify categories of sequences.

Sequences length Length variation may have a strong impact on the clustering. According to the distribution of the data in our corpus, we decided to limit the variation in sequence length by taking into account the first seven mentions only. This means that 21% of the chains have been cut.

Optimal matching The similarity between the pairs of sequences is measured by using the optimal matching based on the Levenshtein distance. The optimal matching between two sequences is the minimum cost to transform one into the other by taking into account both the substitution and insertion or deletion operations. We chose this method because it can be applied to sequences of unequal lengths e.g. chains of two or seven mentions (Studer and Ritschard, 2014; Gabadinho et al., 2011). To compute the substitution-cost matrix we used the "TRATE" method, which determines the costs from the estimated transition rates (Lesnard and Saint Pol, 2006; Gabadinho et al., 2011).

Hierarchical clustering Since we have no hypothesis regarding the optimal number of clusters, we use hierarchical clustering and apply the Ward's linkage criterion which calculate the variance of clusters. We observed the results with different values, from 2 to 5. In this paper, we discuss in more details the results we obtained with 3 and 5 clusters.

3 Results

The first application of the clustering method was experimented on the 581 sequences provided by the AnnoDis corpus. Each sequence was automatically clustered in three and five classes according to the grammatical category of the mentions.

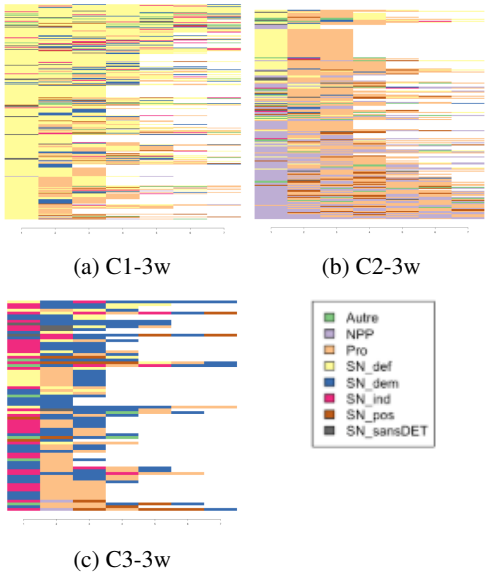
3.1 Description of the clusters

3-way clustering: The three classes of chains clearly differ according to the type of the first mention. Figure 1 gives a global overview of the three classes. Each color corresponds to a grammatical category of the mentions. Within each graph, the different chains are sorted according to their similarity, by using the multidimentional scaling (Popper and Heymann, 1996; Yeturu, 2020), in order to better visualize the heterogeneity of the classes.

The first cluster (C1-3w[2]) groups together the largest number of sequences (317) among which 89.6% (284) start with a definite NP (in light yellow). Typically, there are other definite NPs as next mentions.

[1]The AnnoDis corpus is available at: `http://redac.univ-tlse2.fr/corpus/annodis/`

[2]We use this notation to indicate the number of the cluster and whether it is part of 3-way clustering (3w) or 5-way clustering (5w).



(a) C1-3w (b) C2-3w

(c) C3-3w

Figure 1: 3-way clustering

Example (1) illustrates an excerpt of a C1-3w chain in which all mentions (underlined) are definite NPs (namely: *the Champagne vineyard* [...] *the vineyard* [...] *the vineyard* [...] *the Champagne vineyard*).

(1) *le vignoble champenois s'étendait sur quelques [...] le vignoble connaît [...] le vignoble s'est réduit à 12 000 hectares. Aujourd'hui, en 2007, le vignoble champenois s'étend sur 32 341 hectares.*

Cluster C2-3w groups together 201 sequences among which 50.7% (102) start with a proper noun (in light violet) and 28.8% (58) with a definite NP (in light yellow). The next mentions are mainly proper nouns or pronouns (in light orange). In contrast with cluster C1-3w, there are very few definite NPs as next mentions. Example (2) shows a typical C2-3w chain (*F. de Saussure* [...] *he* [...] *him* [...] *he*).

(2) *Chez F. de Saussure, l'analogie [...], il pose [...]. Pour lui, cette tendance [...]. Comme H. Paul, il ramène le concept [...]*

Finally, cluster C3-3w gathers 63 sequences, among which 47.6% (30) begin with an indefinite NP (in fuchsia) as first mention and 36.5% (23) with a demonstrative NP (in blue), mostly followed by other demonstrative NPs or pronouns (in light orange). Example (3) illustrates a C3-3w chain which starts with an indefinite NP (*an oppositional connotation*, followed by two demonstratives (*This*, *this dichotomy*).

(3) *[...] présente ainsi fréquemment une connotation oppositionnelle (sinon contradictoire) avec le linguistique. Ceci est particulièrement crucial [...] Cette dichotomie pose problème au psychologue [...].*

A preliminary interpretation of these results suggests that these three classes of coreference chains may be pointing at three different strategies for introducing and maintaining the referent. Chains in cluster C2-3w (proper nouns followed by pronouns) seem typically associated to human referents. In addition, the sequence of pronouns may also indicate that the referent is likely to be the topic of discourse. Chains in C3-3w are rather used for presenting ideas or concepts. Demonstratives NPs are likely to encapsulate large portions of text presenting such propositional content under an abstract anaphora (e.g. *this question*).

5-way clustering: We give here a quick overview of the results obtained with 5 clusters. Chains with indefinite and demonstrative NPs (C3-3w) are still clustered in a class (C2-5w). The 5-way clustering provides a different distribution of the chains that begin with a definite NP (C1-3w) and a proper noun (C2-3w).

More precisely, we see that a finer categorization is provided for the chains that begin mostly with a definite NP. Chains in C1-5w (45 sequences) are mainly composed by definite NPs with little variation in categories of the next mentions (as previously exemplified in (1)). In contrast, chains in C3-5w (272 sequences) present a greater variety of next mention types, while cluster C4-5w (96 sequences) mainly exhibits pronouns as next mentions. Finally, cluster C5-5w (105 sequences), begins mostly with proper nouns, next mentions being mainly proper nouns or pronouns.

This categorization highlights more clearly the homogeneity of the chains, as suggested by Obry et al. (2017). It also points at a difference within chains that are typically associated to human referents, namely C4-5w and C5-5w. The strong presence of pronouns indicates that the referent is not in competition with other referents, while the alternation between proper nouns and pronouns or possessives suggests that the referent is not sufficiently accessible because of competition between referents or long-distance coreferential relation.

What follows reports a first experiment to back up this qualitative analysis.

3.2 Interpretation of the clusters

Chain composition can vary according to language, mode, genre or register (Grishina and Stede, 2015; Kunz et al., 2016), text type (narrative or non-narrative) or the semantic nature of referent (Longo, 2013). In this study, we focus on two parameters: text genre (corresponding to the three sub-corpora) and the semantic nature of referent. The chi-squared test was used to observe correlations between these two parameters and the classes of chains.

Text genre

3-way clustering: The chi-squared test highlights a significant relationship (df = 4, p-value = 2.575e-06) between classes and text genres of the corpus. The observation of the Pearson's residual highlights a negative correlation between C3-3w and the WIK2 sub-corpus, a positive correlation between C2-3w and the WIK2 sub-corpus and a negative one between C2-3w and the GEOP sub-corpus. There are no other correlations.

Following these first results, we can say that chains composed by proper nouns followed by pronouns are more frequent in encyclopedic texts than in geopolitical reports. In geopolitical texts we find a majority of chains composed by definite noun phrases followed by pronouns.

5-way clustering: The chi-squared test highlights a significant relationship (df = 8, p-value = 1.945e-05) between classes and text genres of the corpus. The observation of the Pearson's residual highlights correlations similar to that observed for the 3-way clustering: a negative correlation between C2-5w and the WIK2 sub-corpus, a positive correlation between cluster C5-5w and the WIK2 sub-corpus and a negative one between cluster C5-5w and the GEOP sub-corpus. There are no other correlations. The differences between C4-5w and C5-5w are not dependant on the text genres.

Semantic nature of referent A semi-automatic annotation has been carried out to distinguish chains referring to human or non-human in the AnnoDis corpus.

3-way clustering: The correlation between the classes and the human or non-human nature of the referent is even clearer. The chi-squared test highlights a significant relationship between classes and referent types (df = 2, p-value < 2.2e-16), with a positive correlation between C2-3w and chains referring to humans and between C1-3w and C3-3w and chains referring to non-humans.

	Human	**Non-human**
GEOP	136	98
LING	22	65
WIK2	138	122
AnnoDis	296	285

Table 3: Number of human and non-human chains

5-way clustering: The chi-squared test highlights a significant relationship between classes and referent types (df = 4, p-value < 2.2e-16), with a positive correlation between C4-5w and C5-5w and chains referring to humans. There is also a positive correlation between C2-5w and C3-5w and chains referring to non-humans.

4 Discussion and perspectives

This application of sequence analysis to coreference chain description provides clusters that are linguistically coherent and interpretable and that reflect different strategies for introducing and maintaining the referent. This method allows for fine-grained analyses of the mentions chaining and demonstrate that text genres and the semantic nature of referent influence the chain composition.

Next steps will consist in refining the model in several ways. First, we will take into account more fine-grained features, for example by distinguishing the different types of pronouns (personal, demonstratives, indefinite, possessives). This may prove useful to highlight further differences in the composition of chains, potentially also related to the referent types. We will also test other features, such as syntactic functions, relations between mentions, degree of accessibility, coupled with different settings for sequence clustering and a larger variety of referent types (e.g. concrete referent vs abstract referent, generic vs specific, individuals vs collectives, also taking advantage of named entity information); the application of this method to a larger corpus will make it possible to take into account a broader range of text genres and referent types.

Such a description is a complement to more traditional descriptions and may have longer term

applications to coreference resolution, since taking into account variation is crucial to improve the systems (Recasens and Hovy, 2010; Uryupina and Poesio, 2012). Soon et al. (2001) show that entity-type information could be a useful feature for coreference resolution. In the same line, (Khosla and Rose, 2020) demonstrate that neural models using contextualised representations like BERT (Peters et al., 2018) improve coreference resolution performances when entity-type features are explicitly taking into account.

References

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

Nicholas Asher, Philippe Muller, Myriam Bras, Lydia-Mai Ho-Dac, Farah Benamara, Stergos Afantenos, and Laure Vieu. 2017. ANNODIS and related projects: case studies on the annotation of discourse structure. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 1241–1264. Springer Netherlands.

Christian Brzinsky-Fay, Ulrich Kohler, and Magdalena Luniak. 2006. Sequence analysis with Stata. *The Stata Journal*, 6(4):435–460.

Francis Corblin. 1995. *Les formes de reprise dans le discours. Anaphores et chaînes de référence*. Presses Universitaires de Rennes.

Alexis Gabadinho, Gilbert Ritschard, Nicolas Séverin Mueller, and Matthias Studer. 2011. Analyzing and visualizing state sequences in R with TraMineR. *Journal of statistical software*, 40(4):1–37.

Alexis Gabadinho, Gilbert Ritschard, Matthias Studer, and Nicolas S Müller. 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*.

Yulia Grishina and Manfred Stede. 2015. Knowledge-lean projection of coreference chains across languages. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, Beijing, China.

Michael A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman: London.

Sopan Khosla and Carolyn Rose. 2020. Using type information to improve entity coreference resolution. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 20–31. Association for Computational Linguistics.

K. Kunz and Ekaterina Lapshinova-Koltunski. 2015. Cross-linguistic analysis of discourse variation across registers. *Nordic Journal of English Studies*, 14:258–288.

Kerstin Kunz, Ekaterina Lapshinova-Koltunski, and José Manuel Martinez. 2016. Beyond identity coreference: Contrasting indicators of textual coherence in English and German. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 23–31.

Ekaterina Lapshinova-Koltunski and Kerstin Kunz. 2020. Exploring coreference features in heterogeneous data. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 53–64. Association for Computational Linguistics.

Laurent Lesnard and Thibaut de Saint Pol. 2006. Introduction aux méthodes d'appariement optimal (optimal matching analysis). *Bulletin de méthodologie sociologique. Bulletin of sociological methodology*, 90:5–25.

Laurence Longo. 2013. *Vers des moteurs de recherche "intelligents" : un outil de détection automatique de thèmes.* Phd thesis, Université de Strasbourg.

Anna Nedoluzhko and Ekaterina Lapshinova-Koltunski. 2016. Contrasting coreference in Czech and German: from different frameworks to joint results. In *Computational Linguistics and Intellectual Technologies: Proceedings of the 22nd International Conference "Dialogue-21"*, Moscow, Russia.

Vanessa Obry, Julie Glikman, Céline Guillot-Barbance, and Bénédicte Pincemin. 2017. Les chaînes de référence dans les récits brefs en français: étude diachronique (xiiie-xvie s.). *Langue française*, 3:91–110.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Massimo Poesio, Sameer Pradhan, Marta Recasens, Kepa Rodriguez, and Yannick Versley. 2016a. Annotated corpora and annotation tools. In Massimo Poesio, Roland Stuckardt, and Yannick Versley, editors, *Anaphora Resolution: Algorithms, Resources, and Applications*, pages 97–140. Springer Berlin Heidelberg, Berlin, Heidelberg.

Massimo Poesio, Roland Stuckardt, and Yannick Versley, editors. 2016b. *Anaphora Resolution*. Theory and Applications of Natural Language Processing. Springer Berlin Heidelberg.

Richard Popper and Hildegarde Heymann. 1996. Analyzing differences among products and panelists by multidimensional scaling. In Tormod Naes and Einar Risvik, editors, *Multivariate analysis of data in sensory science*, volume 16 of *Data Handling in Science and Technology*, pages 159–184. Elsevier.

Marie-Paule Péry-Woodley, Stergos Afantenos, Lydia-Mai Ho-Dac, and Nicholas Asher. 2011. La ressource ANNODIS, un corpus enrichi d'annotations discursives. *Traitement Automatique des Langues*, 52(3):71–101.

Marta Recasens and Eduard Hovy. 2010. Coreference resolution across corpora: Languages, coding schemes, and preprocessing information. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1423–1432, Uppsala, Sweden. Association for Computational Linguistics.

Marta Recasens, Eduard Hovy, and M. Antònia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Matthias Studer and Gilbert Ritschard. 2014. A comparative review of sequence dissimilarity measures. *LIVES Working Paper No. 33.*, Geneva, Switzerland: Swiss National Centre of Competence in Research.

Olga Uryupina and Massimo Poesio. 2012. Domain-specific vs. uniform modeling for coreference resolution. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Marilyn A. Walker. 2000. Vers un modèle de l'interaction du centrage avec la structure globale du discours. *Verbum*, XXII(1):31–58.

Kalidas Yeturu. 2020. Chapter 3 - machine learning algorithms, applications, and practices in data science. In Arni S.R. Srinivasa Rao and C.R. Rao, editors, *Principles and Methods for Data Science*, volume 43 of *Handbook of Statistics*, pages 81–206. Elsevier.