A linear approach of chain composition

Silvia Federzoni, Lydia-Mai Ho-Dac, Cécile Fabre

A linear approach of chain composition

Silvia Federzoni (1), Lydia-Mai Ho-Dac (1), Cécile Fabre

CLLE (UMR 5263) – CNRS & Université de Toulouse Jean Jaurès

Abstract

This corpus-based approach to coreference chains analyzes recurrences in the patterns of chains, providing new insights into conventions or preferences in the forms of referential expressions. By taking into account the linearity of discourse and the succession of mentions, it goes beyond the more commonly implemented analysis of global characteristics.

We analyze 581 reference chains from the French corpus AnnoDis. Using clustering methods, we first show that the resulting clusters are linguistically interpretable. We then demonstrate that animacy and genre influence chain composition. Finally we identify the main patterns of coreference chains in the corpus. This highlights different types of chains and discourse strategies, which vary across genres, and confirms a major influence of referent type.

Keywords: coreference chains, linear approach, corpus-based analysis

1. Introduction

Coreference chains (CRCs) are discourse structures that group together several clauses around a common referent expressed via a “sequence of

expressions between which the interpretation builds a referential identity" (Corblin 1995). Viewed in a Systemic Functional Linguistics framework (Halliday 1985), CRCs play a crucial role in the organization and the interpretation of discourse, providing a fundamental mechanism that ensures referential continuity by creating referential cohesive ties (Halliday & Hasan 1976). We are interested in the CRCs' ability to create texture and to segment discourse into referential continuation spans (Goutsos 1996). In this view, we consider CRCs as a whole, i.e. as structures signaled by sequences of coreferential expressions, also called mentions, rather than as a set of connected expressions. Our objective is to analyze chain composition and identify different types of chains leading to various series of coreferential expressions, according to linguistic factors. As is already well known, there are multiple factors to consider, such as language, mode or register (Lapshinova-Koltunski & Kunz 2020), text type (e.g. narrative or non-narrative) or genre, the semantic nature of the referent (Longo & Todirascu 2010) and the importance of the CRC in discourse organization. For example, topical chains, which are built around a common referent which denotes the discourse topic, typically cover large portions of text and contain a series of referential expressions, with the first one introducing the topic and the next ones maintaining it active via coreferential expressions with a high degree of accessibility (Ariel 1990) and short distances between them.

CRC composition is usually tackled by studying how their global characteristics vary according to these discourse factors (see for instance Nedoluzhko & Lapshinova-Koltunski 2016, Kunz & Lapshinova-Koltunski 2015, Schnedecker 2021). The features used in the literature as global characteristics are mainly the number of mentions, the average distance between them, the distribution of their categories, the category of the first mention. This global approach is called *paradigmatizing* by Schnedecker (2021) because it considers a CRC as a set of mentions denoting a common entity of discourse (Recasens 2010, Uryupina et al. 2016).

Although the global approach has been useful to highlight some differences between chains (Lapshinova-Koltunski & Kunz 2020, Schnedecker & Landragin 2014), it has limitations: since it does not take the linearity of discourse into consideration, it is not relevant for studying the cohesive ties that are built by the CRCs. Chains that have a similar global composition may in fact behave very differently if considered linearly. An alternative approach is needed, to which we refer as *linear* (or *syntagmatizing* in Schnedecker 2021) to provide a new description of CRCs that will take into account the series of mentions, variations in terms of distance between the mentions, competition among referents and the factors that have an impact on the referential choices at a specific stage of the continuity (e.g. other discourse structures) (Kibrik 2011).

To illustrate the need for a linear analysis, we introduce examples (1) and (2) (with mentions underlined) and their corresponding schematic representation CRC1 and CRC2 in figure 1, which show how the chains differ in the two examples despite a similar global composition (similar number of mentions, mixture of proper names and pronouns).

(1) Le journaliste britannique William Thomas Stead a mené, durant sa carrière, de nombreux combats par le biais d'articles et de nouvelles. L'un d'entre eux concerne le manque de moyens de sauvetage à bord des paquebots. Il publie une première nouvelle en 1886 intitulée *Comment le Paquebot Poste sombra au milieu de l'Atlantique, par un Survivant*, racontant une collision entre deux navires dont les passagers ne sont pas tous sauvés, faute de moyens de sauvetage. Stead conclut : « C'est exactement ce qui se produira si les paquebots sont lancés avec trop peu de canots ». Six ans plus tard, il publie *De l'Ancien Monde au Nouveau*, nouvelle dans laquelle il raconte un voyage fictif qu'il aurait fait à bord du paquebot (bien réel) Majestic de la White Star Line sous le commandement d'Edward Smith. Au cours de la traversée, le navire s'arrête pour repêcher les naufragés d'un paquebot ayant heurté un iceberg. Stead conclut cette fois-ci en disant que « les océans parcourus par de rapides paquebots sont jonchés des os blanchis de ceux qui ont embarqué comme nous et qui ne sont jamais arrivés à bon port ». Le 15 avril 1912, Stead se trouve à bord du Titanic, commandé par Edward Smith, et meurt dans le naufrage. (wik2_titanicCT_coder2_1281096316275)[1]

The British journalist William Thomas Stead has fought many battles in his career through articles and short stories. One of them concerns the lack of rescue means on board liners. He published a first short story in 1886 entitled *How the Mail Steamer Sank in the Mid-Atlantic, by a Survivor*, about a collision between two ships whose passengers are not all saved, due to a lack of rescue facilities. Stead concludes:

[1] For each example, we provide the ID of the chain. This ID can be used to find the complete version of the example on the AnnoDis website: http://redac.univ-tlse2.fr/corpus/annodis/.

"This is exactly what will happen if the liners are launched with too few boats." Six years later, he published *From the Old World to the New*, a short story in which he wrote about a fictional voyage he had taken aboard the (real) White Star Line liner Majestic under the command of Edward Smith. During the voyage, the ship stops to rescue survivors from a liner that has hit an iceberg. On April 15, 1912, Stead was aboard the Titanic, commanded by Edward Smith, and died in the sinking.

(2) Léonard de Vinci a eu beaucoup d'amis qui sont reconnus dans leurs domaines respectifs ou ont eu une influence importante sur l'Histoire. Il s'agit notamment du mathématicien Luca Pacioli avec qui il a collaboré pour un livre, César Borgia au service duquel il a passé deux années, Laurent de Médicis et le médecin Marcantonio della Torre. Il a rencontré Nicolas Machiavel, avec qui il développera plus tard une étroite amitié, et Michel-Ange avec qui il a été rival. Parmi ses amis, se trouvent également Franchini Gaffurio et Isabelle d'Este. Léonard semble ne pas avoir eu d'étroites relations avec les femmes, sauf avec Isabelle. Il a fait un portrait d'elle, au cours d'un voyage qui le mena à Mantoue, qui semble avoir été utilisé pour créer une peinture, aujourd'hui perdue. Il était également ami de l'architecte Jacopo Andrea da Ferrara jusqu'à son assassinat.
(wik2_leonardDeVinciCT_coder2_1257955114587)

Leonardo da Vinci had many friends who are recognized in their respective fields or have had an important historical influence. These include the mathematician Luca Pacioli, with whom he collaborated on a book, Caesar Borgia, in whose service he spent two years, Lorenzo de Medici and the physician Marcantonio della Torre. He met Nicholas Machiavelli, with whom he later developed a close friendship, and Michelangelo of whom he was a rival. Among his friends were also Franchini Gaffurio and Isabella d'Este. Leonardo does not seem to have had close relationships with women, except with Isabella. He made a portrait of her during a trip that took him to Mantua, which seems to have been used to create a painting, now lost. He was also a friend of the architect Jacopo Andrea da Ferrara until his murder.

(CRC 1) 

(CRC 2) 

Figure 1: Schematic representation of examples (1) and (2), adapted from Schnedecker (2021:55)

In Figure 1, the CRCs of (1) and (2) are coded by geometric shapes representing each mention underlined in the text (square for proper names, diamond for pronouns, circle for definite NPs, triangle for possessives) and disposed linearly according to the relative distance between mentions. This visualization highlights the linear differences between CRC1 and CRC2. It shows that the main differences between the two chains concern the way in which the referent is introduced (category of the first mention) and the way in which it is maintained, i.e. the inter-distance between the mentions and the linear distribution of the categories throughout the chains.

First, the two chains begin in very different ways, with a complete definite NP followed by a possessive for CRC1 and a proper name followed by a pronoun for CRC2. Secondly, the two chains show a different linear arrangement of mentions or inter-distance as suggested by Rousier-Vercruyssen and Landragin (2019). In CRC1, the first three mentions are very close, in contrast with CRC2 where the second mention is distant from the first one. These differences may indicate that the referent of CRC1 (William Thomas Stead, a secondary character in the story of the Titanic

that this Wikipedia article relates) is less familiar to the reader than the referent in CRC2 (Leonardo da Vinci, the main topic of the eponymous Wikipedia article). This status of the CRC2 referent allows for the series of pronouns in the subsequent mentions despite the large inter-distance and numerous competing human referents (Luca Pacioli, Caesar Borgia, etc.).

As illustrated from these examples, considering CRCs as sequences of mentions should allow for a better understanding of the strategies that ensure referential continuity in a text. It will facilitate the study of the factors that influence the speakers' referential choice at a particular stage of their discourse (introduction, maintain or shift stage) (Fossard et al. 2018).

In this paper, we present the results of an experiment that implements this linear approach for the analysis of CRCs. This experiment consists in using computational methods for clustering sequences of mentions and identifying recurring patterns of sequences in a diversified French corpus annotated in CRCs. The results produced by this bottom-up approach are linguistically interpreted against the backdrop of the Accessibility Theory (Ariel 2001), which provides potential explanations for the observed patterns. We analyze the impact of two linguistic factors on the resulting classification, and demonstrate that the animacy of the referent and genre have an impact on coreference chains composition. Then, we show how this approach gives rise to recurring patterns of CRCs. As a result, the patterns are used to

characterize different types of chain and to propose new elements for the definition of a typology of CRCs.

2. Application to the analysis of an annotated French corpus: the AnnoDis Corpus

The AnnoDis corpus is a written French corpus annotated with three types of discourse phenomena: rhetorical relations, enumerative structures and topical chains. In this study we use this latter layer of annotation, which concerns CRCs that are built upon a prominent or topical element (Péry-Woodley et al. 2011, Asher et al. 2017). The corpus comprises full long non-narrative texts, unlike other corpora which are usually made up of short texts such as newspaper articles (e.g. OntoNotes corpus, Weischedel et al. 2013) or excerpts (e.g. Democrat corpus, Landragin 2015). The texts are organized in three subcorpora falling into three genres and domains: reports in geopolitics (GEOP) published by the French Institute of International Relations from 2001 to 2003, research papers from the 2008 edition of the International Congress of French Linguistics (LING) and encyclopedic articles from the 2009 version of the French version of Wikipedia (WIK2).

The annotation was carried out following a top-down approach, which first consists in identifying the relevant CRC by delimiting the portion of text

corresponding to the chain. Annotators were guided by potentially coreferential pre-marked expressions, e.g. reiterated nominal phrases, proper names and pronouns (personal and demonstrative), all in subject position. This premarking allowed them to zoom in on segments of text that contained potential topical coreferential expressions. Then, they had to identify and annotate every expression which indicated that the topical chain is still ongoing, possibly across several sentences but always within a heading section.

Annotators received a broad definition of the referential expressions and were asked to consider any expression that allowed them to perceive a referential continuity. A post-processing step converted the annotations into coreference chains annotations and eliminated some biases induced by the protocol (suppression of duplicates, redefinition of mention boundaries when necessary, etc.) (Péry-Woodley et al. 2017, Federzoni et al. 2020). In all the examples given in the next sections from the AnnoDis corpus, the mentions are presented as they have been identified and delimited by the annotators.

This annotation method contrasts with traditional CRC annotation methods, which favor a more bottom-up approach in which the annotators are first asked to identify the mentions, while the structure is reconstructed a posteriori (cf. Biber et al. 2007 and Asher et al. 2017 for more detail about

top-down vs. bottom-up approaches to discourse). Table 1 shows the composition of the AnnoDis corpus.

Subcorpus	Texts	Words	CRCs	Mentions
GEOP	32	266,000	234	1,125
LING	25	169,000	87	478
WIK2	30	231,000	260	1,853
Total AnnoDis	87	666,000	581	3,456

Table 1: Composition of the AnnoDis corpus annotated with CRCs

3. A sequence analysis approach to coreference chains

In order to identify classes of CRCs on the basis of their linear composition, we propose a computational method that analyzes the CRCs as sequences of mentions and provides a classification of the resulting sequences. This method resorts to sequence analysis (Abbott 1995), which is usually used in social sciences to model the chronology of states or events to "identify regularities, similarities or to build typologies of typical sequences" (Robette 2011) for the study of life-course trajectories, family histories or professional career paths (Dietrich et al. 2014, Fasang 2014). We propose to apply these techniques to CRCs by considering each mention as a linguistically characterized state, starting with a single feature, the grammatical category, which has proven to provide accurate information to

build chain typologies and to assess their degree of homogeneity (Obry et al. 2017). Figure 2 shows the representation of CRC1 and CRC2 as sequences, with one state per mention.

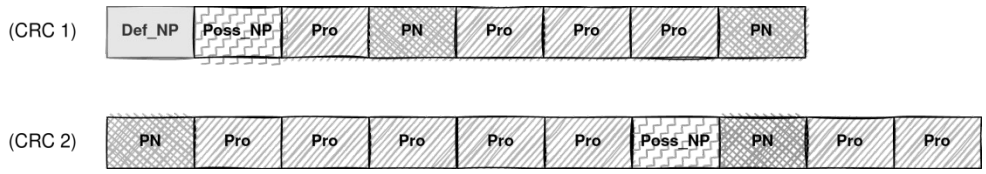

Figure 2: CRC1 and CRC2 as sequences

The sequential method was implemented with the TraMineR toolbox (Gabadinho et al. 2009, 2011), which provides a large range of automatic methods for visualizing and analyzing sequences. Among these methods we chose the clustering method in order to identify classes of CRCs automatically. The next section lists the parameter settings we adopted to apply the clustering method to the specificity of our data.

3.1 Parameters of the sequence analysis

Four parameters must be considered to adapt the sequential method to CRCs: the clustering algorithm, the number of clusters (i.e. the number of classes that must be distinguished), the similarity measure and the maximal sequence length.

In this study, we used a hierarchical clustering algorithm, as we had no hypothesis regarding the optimal number of clusters. This clustering algorithm proceeds in a bottom-up manner, grouping together similar items into clusters according to a measure of similarity. The hierarchy of clusters is visualized as a tree diagram (dendrogram) that can be cut at different levels depending on the number of classes that we want to consider. We chose to examine the classes that emerged due to a larger vertical gap between two successive clusters on the dendrogram, corresponding to three configurations with 2, 3 and 5 classes. In this article, our analyses focus on the two configurations we found to be the most relevant: the 3-classes configuration (3c) and the 5-classes configuration (5c).

More specifically, the classification program requires to choose among different measures to compute the similarity within each pair of sequences. The similarity measure that we used is the optimal matching analysis, based on the Levenshtein distance (Halpin 2010, Lesnard 2014, Levenshtein 1966). The optimal matching is the minimal cost to transform one sequence into the other when considering operations of substitution, insertion or deletion. One example of substitution is the transformation of a definite NP into a pronoun. If one sequence has more mentions than the other, the operations will consist in deleting mentions. The cost is the sum of all the operations needed to transform one sequence into the other. We chose this

method because it is efficient to handle sequences of unequal lengths (Studer & Ritschard 2014, Gabadinho et al. 2011).

Dealing with sequences of different lengths is indeed a major problem which could influence the computation of the similarity measure. We therefore had to reduce length variation in our data. In the AnnoDis corpus, CRCs have a mean length of 5.95 mentions, ranging from 2 mentions (only 6% of the chains) to 40, as well as a standard deviation of 4.87 and a median value of 4. Since 78.66% of the chains contain 7 mentions or less, we chose this value as a cut-off point. As a consequence, the length of sequences can vary from 2 to 7 states and 21.34% of the AnnoDis CRCs were cut off.

3.2 From mentions to states

As stated above, to apply the sequence analysis to CRCs categorization we consider each mention as a state of the sequence. Each state is characterized by its grammatical category. This means that the input sequences are composed of states corresponding to labels assigned to each mention in the AnnoDis corpus. Note that during the development of the corpus, the grammatical category was assigned semi-automatically: the labeling is based on the original manual categorization of each mention and completed by automatic POS-tagging and syntactic parsing with Talismane (Urieli 2013). We chose to label the mentions following the 8 categories that are

frequently used for the description of CRCs, so as to distinguish between different noun phrases, proper names and pronouns. This distinction is important for the analysis of CRC composition, to understand the strategies for introducing, maintaining and shifting the referent, and to study the impact of various factors on the choice of referential expression forms. Table 2 indicates the distribution of grammatical categories in the CRCs (before the cut-off) and in the input sequences (after the cut-off). The last column indicates the amount of excluded information.

	Before cut-off		**After cut-off**			**Information excluded**	
Label	**#**	**%**	#	**%**	**% to the total**	**#**	%
Def_NP	1198	34.7	1019	37.38	29.48	179	5.18
Pro	1026	29.7	794	29.13	22.97	232	6.71
PN	442	12.8	292	10.71	8.45	150	4.34
Dem_NP	272	7.9	247	9.06	7.15	25	0.72
Poss_NP	182	5.3	123	4.51	3.56	59	1.71
Other	146	4.2	99	3.63	2.86	47	1.36
Ind_NP	126	3.7	99	3.63	2.86	27	0.78
NoDet_NP	64	1.9	53	1.94	1.53	11	0.32
Total	3456	100	2726	100	78.88	730	21.12

Table 2: Distribution of the mentions into grammatical categories in the AnnoDis corpus before and after selecting a cut-off point

In the next sections, we discuss the results in three steps: first we show that the resulting classes are linguistically interpretable and are likely to reflect different types of chains. Then we demonstrate that the semantic nature of the referent and text genres influence the composition of coreference chains. Finally, we present the main patterns of CRCs that emerge from the corpus, corresponding to different strategies for introducing and maintaining the referent.

4. Clustering coreference chains according to their sequence of mentions

The 581 sequences were clustered into classes that clearly differ in terms of linear composition. We examine first the 3 classes provided by the 3c configuration, then the 5 classes provided by the 5c configuration, considering, on the one hand, the type of the first mention, and on the other hand, the types of the following mentions.

The first class named C1-3c[2] groups together the largest number of sequences (317). C1-3c sequences usually start with a definite NP (284

[2] We use this notation to indicate the number of the class (C1) and whether it is part of the 3-classes or 5-classes configuration (3c or 5c).

chains, that is, 89.6% of the cluster[3]) typically followed by other definite NPs, as in example (3) in which all mentions are definite NPs.

(3) Juste avant le début de la révolution française, le vignoble champenois s'étendait sur [...]. Dans la seconde moitié du xixe siècle, le vignoble connaît [...]. Après les fléaux du phylloxéra et de la Grande guerre, le vignoble s'est réduit à [...]. Aujourd'hui, en 2007, le vignoble champenois s'étend sur 32 341 hectares.
(WIK2 - wik2_vinDeChampagneCT_coder3_1254135673500)

Just before the start of the French Revolution, the Champagne vineyards covered [...]. In the second half of the 19th century, the vineyard reached [...]. After the scourges of phylloxera and the Great War, the vineyard was reduced to [...]. Today, in 2007, the Champagne vineyards cover 32,341 hectares.

The class C2-3c groups together 63 sequences, among which 47.6% (30) begin with an indefinite NP as first mention and 36.5% (23) with a demonstrative NP, mostly followed by other demonstrative NPs or pronouns. Example (4) illustrates a C2-3c chain which starts with an indefinite NP (*an oppositional connotation*), followed by two demonstratives (*this*, *this dichotomy*).

(4) le "communicatif" présente ainsi fréquemment une connotation oppositionnelle [...]. Ceci est particulièrement crucial [...]. Cette dichotomie pose problème [...].
(LING - ling_barbazanCT_coder2_1320622605359[4])

the "communicative" frequently has an oppositional connotation [...]. This is particularly crucial [...]. This dichotomy is problematic [...].

[3] The remaining C1-3c sequences have a different type of first mention but the same main types of next mentions.

[4]

The last class, C3-3c, groups together 201 sequences beginning with a proper name (50,7%) or a definite NP (28,8%) followed by proper names or pronouns. In contrast to class C1-3c which also starts with definite NPs, there are very few definite NPs as next mentions. Example 5 shows a typical C3-3c chain.

(5) Chez F. de Saussure, l'analogie apparaît comme un principe central [...]. Pour lui, cette tendance [...]. Comme H. Paul, il ramène le concept au [...]. Il lui assigne deux rôles majeurs […]
(LING - ling_dalCT_coder3_1253024655578)

In the works of F. de Saussure, analogy is a fundamental principle of [...]. To him, this tendency [...]. Like H. Paul, he reduces the concept to [...]. He assigns it two major roles: [...]

The three classes mentioned above seem to point at three different strategies for introducing and maintaining the referent. While chains in C3-3c use a proper name to introduce a referent which can then easily be maintained via a sequence of pronouns, chains in C1-3c are mainly composed of definite NPs used both for introducing and for maintaining the referents. A preliminary interpretation of these results suggests that C3-3c sequences may be used for referring to an individual referent that can be named, typically a human referent as in (5). A complementary interpretation suggests that C3-3c sequences are more likely to be used when the referent is highly accessible all along the CRC. In other words, these sequences could be found when the mentions are close enough and/or when there are no competing referents that could cause referential ambiguity.

Sequences that are grouped in the smallest C2-3c class tend to be introduced by indefinite NPs or demonstrative NPs. First observations suggest that these CRCs are typically used for presenting ideas or concepts. These sequences are characterized by either a first mention associated to a low degree of accessibility (i.e. a brand new referent introduced via indefinite NP) or a first mention introduced with a demonstrative NP that is likely to encapsulate arguments under an abstract anaphora which introduces a new referent (e.g. *this question*). Next mentions are mainly associated to a high degree of accessibility (e.g. other demonstrative NPs or pronouns).

If we now consider the 5-classes configuration (5c), two facts emerge from the comparison with the 3c one: a similar clustering of the chains that are composed of indefinite and demonstrative NPs (C2-3c and C2-5c are strictly the same, i.e. both configurations have grouped the same sequences in one class) and a different distribution of the chains grouped in C1-3c and C3-3c. The 5c configuration splits the class C1-3c in two, i.e. the C1-3c sequences have been grouped in two different classes by the 5c configuration: C1-5c gathers 45 chains, most of which are composed exclusively of sequences of definite NPs (as in example (3)), while C3-5c consists of 272 chains with a definite NP as first mention and is much more heterogeneous, with a variety of grammatical categories as next mentions. Example (6) illustrates such C3-5c sequences with a definite NP followed by a mixture of demonstrative NPs, possessives and pronouns.

(6) L’affaire Dreyfus a pour origine une erreur judiciaire [...] Cette affaire a bouleversé la société française [...].
La révélation de ce scandale, dans « J'Accuse...! » [...] À son paroxysme en 1899, elle révèle [...]. Elle divise [...]. Cette affaire est le symbole [...]. Enfin, elle suscite [...].
(WIK2 - wik2_affaireDreyfussCT_coder2_1254316308199)

The Dreyfus Affair began with a miscarriage of justice [...]. This affair disrupted the French society [...] The publication of this scandal, in "J’Accuse...", [...] At its peak in 1899, it marked [...]. It divided [...]. This affair is the symbol [...]. Finally, it aroused [...].

The 5c configuration also splits the C3-3c class into two finer-grained clusters: C4-5c and C5-5c. C4-5c groups together 96 chains, mostly beginning with a definite NP and followed by pronouns. Chains in C5-5c (105) begin with a proper name, and the next mentions mostly alternate between proper names, pronouns and possessives.

The strong presence of pronouns in C4-5c chains suggests that the referent is highly accessible and not in competition with other referents, as shown in example (7).

(7) Le colonel Picquart est lui aussi réhabilité officiellement et réintégré dans l’armée au grade de général de brigade. Il est même ministre de la Guerre de 1906 à 1909 dans le premier gouvernement Clemenceau. Il meurt en 1914 d’un accident de cheval.
(WIK2 - wik2_affaireDreyfussCT_coder2_1254320843928)

Colonel Picquart was also officially rehabilitated and reinstated in the army with the rank of brigadier general. He was even Minister of War from 1906 to 1909 in the first Clemenceau government. He died in 1914 in a horse accident.

In contrast, the alternation between proper names and pronouns or possessives in C5-5c suggests that the referent is not sufficiently accessible, because of competition between referents or due to long-distance coreferential relations. For example, in (8), the referent (*Godfroy Cavaignac*, underlined) competes with other referents (in bold). As a consequence, it is necessary to restore the referent via proper names to avoid the risk of referential ambiguity.

(8) En effet, Godefroy Cavaignac, nouveau ministre de la Guerre et anti- révisionniste farouche, veut démontrer définitivement la culpabilité de **Dreyfus**, en « tordant le cou » au passage à **Esterházy**, qu'il tient pour « un mythomane et un maître chanteur ». Il est absolument convaincu de la culpabilité de **Dreyfus**, renforcé dans cette idée par la légende des aveux, après avoir rencontré **le principal témoin, le capitaine Lebrun-Renault**. Cavaignac a l'honnêteté d'un doctrinaire intransigeant, mais ne connaît absolument pas les dessous de l'Affaire, que l'État-Major s'est gardé de lui enseigner. Il avait eu la surprise d'apprendre que [...]
(WIK2 - wik2_affaireDreyfussCT_coder2_1254320273671)

Indeed, Godefroy Cavaignac, the new Minister of War and a fierce anti-revisionist, wanted to demonstrate **Dreyfus**'s guilt once and for all, and in passing "strike a fatal blow" at **Esterházy**, whom he considered "a mythomaniac and a blackmailer." He was absolutely convinced of **Dreyfus**'s guilt, reinforced in this idea by the legend of the confession, after he met **the main witness, Captain Lebrun-Renault**. Cavaignac had the honesty of an uncompromising doctrinaire, but was completely unaware of the inner workings of the Affair, which General Staff carefully kept from him. He was surprised to learn that . [...]

To sum up, the 3c as well as the 5c configurations allow us to distinguish the composition of chains according to their sequences of mentions. The 3c

configuration draws the outlines of a first classification that the 5c configuration refines and helps to characterize:

- chains composed of many definite NPs including the first mention (C1-3c)
    - almost only definite NPs (C1-5c)
    - mixture of demonstrative NPs, possessives and pronouns as next mentions (C3-5c)
- chains composed of indefinite and demonstrative NPs (C2-3c or C2-5c)
- chains composed mainly of proper names and pronouns (C3-3c)
    - beginning with a definite NP followed by pronouns (C4-5c)
    - beginning with a proper name followed by both pronouns and proper names (C5-5c)

This first classification, and the observation of some corresponding chains, confirm some semantic and discursive distinctions made in the literature (Schnedecker 2005, Longo & Todirascu 2014, among others). From these results, we make the hypothesis that definite NPs are used for introducing as well as for maintaining non-human referents, and that proper names are used for introducing and maintaining human ones. Indefinite NPs are used for introducing conceptual referents, pronouns to maintain unambiguous human referents, and demonstrative NPs to maintain or encapsulate conceptual referents.

The next section goes further in the evaluation of these tendencies by testing the impact of two variables in the clusters, namely the animacy of the referent and text genre.

5. Impact of animacy and text genre on chain composition

As seen in section 2, the AnnoDis corpus was designed to ensure some diversification in a small range of text genres: the three subcorpora correspond to three sub-genres of expository texts (encyclopedic articles, scientific papers, reports). This allows for a preliminary study, on a small scale, of the impact of genre on chain composition. On the other hand, there is no information in the corpus about the other variable that we want to manipulate, which is the semantic type of the referent. The hypothesis that emerged from the clustering results is that the classes may depend on the human vs non-human distinction. To be able to test this hypothesis, we added this label to each CRC of the AnnoDis corpus via a semi-automatic procedure: if the chain contained a mention that is a proper name, it was tagged *human*; otherwise, it was tagged *non-human*. This automatic tagging was then reviewed manually. 169 chains (29.9%) needed manual correction. This concerned cases where a mention was tagged as a proper name but did not denote a human (e.g. *Linux*) or, conversely, where the referent was

human but was not designated by a proper name (e.g. *Le cycliste - il - sa position - les cyclistes / Cyclists - he - his position - cyclists).*

As shown in Table 3, this binary labeling results in a balanced distribution of the chains in the overall corpus, with discrepancies between sub-corpora: in the linguistic research papers (LING), human referents are predictably underrepresented.

	Human	**Non-human**	**Chains**
GEOP	136 (58%)	98 (42%)	234
LING	22 (25%)	65 (75%)	87
WIK2	138 (53%)	122 (47%)	260
AnnoDis	296 (51%)	285 (49%)	581

Table 3: Distribution of chains referring to human or non-human entities in the AnnoDis corpus

We conducted a chi-squared test which reveals a significant relationship between the classes obtained by the 3c configuration and the human or non-human nature of the referent (df = 2, p-value < 2.2e-16). Specifically, there is a positive relationship between the C3-3c and chains referring to humans. There is also a positive relationship between C1-3c and C2-3c and chains referring to non-humans.

If we look at the 5c configuration, the chi-squared test is still conclusive (df = 4, p-value < 2.2e-16), with a positive relationship between C4-5c and C5-5c and chains referring to humans. There is also a positive relationship

between C2-5c and C3-5c and chains referring to non-humans which validates our hypothesis (section 4).

Considering text genres, the chi-squared test shows a significant dependence between the classes obtained by the 3c configuration and the text genre (df = 4, p-value = 2.575e-06). In particular, for encyclopedic texts (the WIK2 sub-corpus), we found a negative relationship with cluster C2-3c and a positive relationship with C3-3c. We also found a negative relationship between cluster C3-3c and the geopolitical texts (the GEOP sub-corpus).

Interestingly, if we now consider the two variables together, we note that the C3-3c is both positively related to the feature *human* and to the WIK2 subcorpus, while negatively related to the GEOP corpus. These results indicate that CRCs referring to humans are composed differently depending on the subcorpus. Chains consisting of proper names and pronouns dominate in the encyclopedic texts only (e.g. *Leonardo da Vinci... Leonardo da Vinci... Leonardo... he / Osama bin Laden... the latter... Osama bin Laden*). These results suggest that encyclopedic texts favor references to people as individuals, while geopolitical texts focus rather on collective referents, in line with the observations that Longo & Todirascu (2014) have made on non-narrative legal and administrative texts. This is illustrated by examples (9) and (10).

(9) Mathieu Dreyfus, le frère aîné d'Alfred Dreyfus, est convaincu de l'innocence du condamné. Il est le premier

artisan de la réhabilitation de son frère, [...] Mathieu essaie toutes les pistes […]
(WIK2- wik2_affaireDreyfussCT_coder2_1254319039686)

Mathieu Dreyfus, Alfred Dreyfus's elder brother, was convinced of the innocence of the condemned man. He was the first person to work for his brother's rehabilitation, [...] Mathieu tried all the tracks [...]

(10) En 1873, la Standard Oil détenait déjà […] Au tournant du siècle, la S.O. exportait 50 % de sa production; [...]
L'accession de la Standard Oil à une situation de quasi-monopole sur l'aval pétrolier [...]
(GEOP - geop_11CT_coder1_1254301984671)

By 1873, Standard Oil already owned [...]. By the turn of the century, S.O. was exporting 50% of its production; [...]
The accession of Standard Oil to a virtual monopoly on downstream oil - [...]

Note that significant relationships between sub-corpora and classes obtained by the 5c configuration are fairly similar, with C5-5c positively related to WIK2 and negatively related to GEOP, and C2-5c negatively related to WIK2.

6. Patterns of coreference chains

We have shown that the application of sequence analysis provides a first classification of CRCs that is linguistically coherent and interpretable according to their linear composition. We now adopt a more fine-grained

method to identify recurring patterns of sequences that represent the linear composition of prototypical chains.

For this purpose, we used the TraMineR toolbox for extracting the list of sequences of mentions and their frequency in the AnnoDis corpus. As expected, the number of unique sequences was high, with 393 types out of 581 sequences. To establish more general patterns, we grouped together mentions that belonged to the same grammatical category and indicated the number in the following form: *CAT{n}*. For example, a sequence of 3, 4 or 5 definite NPs corresponds to the pattern *def_NP{3:5}*. Because the distinction between the first mention and the next ones is important to interpret chain composition, we kept the first mention apart from the next ones in the patterns. As a consequence, a sequence starting with 1 definite NP followed by several NPs, themselves followed by several pronouns, will correspond to the pattern *def_NP > def_NP{n1:n2} > Pro{n1:n2}*. The symbol > indicates the transition to another category. For some patterns, we considered a transition as optional, as in d*ef_NP > Pro {1:4} > def_NP{1:3} (> Pro {1:2})*, for which the last transition is not necessary. The 6 most frequent patterns are given in Table 4, with their frequency according to the animacy of the referent (human vs. non-human).

	Pattern	Human	Non human
P1	**def_NP (> def_NP{1:3}) > Pro {1:6}**	42	24
P2	**def_NP > def_NP{1:6}**	22	24
P3	**def_NP (> def_NP{1:2}) > Pro {1:4} > def_NP{1:4} (> Pro {1:2})**	18	32
P4	**def NP (> def_NP{1:2}) > dem_NP{1:4} (>Pro|def_NP{1:5}})**	3	26
P5	**PN (> PN{1:2}) > Pro {1:4} > PN{1:2} > Pro{1:4} (> PN{1})**	15	0
P6	**PN (> PN) > Pro {1:4}**	15	0
Subtotal		115	106
Total		296	285

Table 4: The most frequent patterns of reference chains in the AnnoDis corpus

The three most frequent patterns characterize CRCs beginning with a definite NP and followed by other definite NPs or pronouns, in different configurations: one or more pronouns may follow a first sequence of NPs (P1), NPs may be used exclusively throughout the sequence (P2), or both categories may alternate to maintain the referent (P3). These three patterns alternate expressions associated to high and low degrees of Accessibility.

Regarding the animacy of the referent, P1 is rather used to refer to humans, P3 to non-humans and there is no clear tendency at this stage for P2. Lexical repetition (P2) is an option to maintain both human referents (typically in our corpus, in the form of a job title followed by the name of the referent) and non human referents, as shown also by Longo & Todirascu (2014) in legal and administrative texts.

The fourth pattern is characterized by the inclusion of a demonstrative NP in the next mentions, and is almost only found with non-human referents, as illustrated in example (11).

(11) La distinction des registres énonciatifs *histoire* et *discours* illustre [...] Cette dissociation entre perspective énonciative et perspective textuelle, [...] Il est permis de penser qu'elle était une étape nécessaire [...]
(LING - ling_barbazanCT_coder2_1320622234494)[5]

The distinction between the enunciative registers *history* and *discourse* illustrates [...] This dissociation between the enunciative perspective and the textual perspective [...] It is possible to think that it was a necessary step [...]

The fifth and sixth patterns are dedicated to human referents introduced with a proper name and maintained with pronouns, as in (5), or with alternating proper names and pronouns, as in (12).

(12) En mars 1499, Léonard de Vinci est alors employé [...] Il élabore des méthodes pour défendre la ville [...] Il étudie les cours d'eau du Frioul [...] En avril 1500, il revient à Venise pour deux mois, après avoir séjourné à Mantoue, [...] Ainsi,

[5] This example shows an error of annotation on the mention *qu'elle* ('*that it*') that should normally be delimited as *'elle'* ('*it*'). Such error is probably due to the Annodis protocol that does not include specific instructions on the mention delimitation or to difficulties to delimit tokens with an apostrophe via the annotation tool.

> Léonard de Vinci poursuivait bien des recherches plus larges. Il séjourne dans le couvent de la Santissima Annunziata [...] Il fait un bref séjour à Rome à la villa d'Hadrien à Tivoli. [...]
> (WIK2 - wik2_leonardDeVinciCT_coder2_1257953464805)
>
> In March 1499, Leonardo da Vinci was employed [...] He developed methods to defend the city [...]. He studied the waterways of Friuli [...] In April 1500, he returned to Venice for two months, after a stay in Mantua, [...] Thus, Leonardo da Vinci was carrying out much broader research. He stayed in the convent of the Santissima Annunziata [...] He stayed briefly in Rome at Hadrian's Villa in Tivoli.

These six patterns cover 38% of the input chains. They do not capture the full diversity of the strategies for introducing and maintaining a referent, but shape the outline of a first typology of CRCs. First, CRCs that start with a definite NP differ from CRCs that start with a proper name. Secondly, CRCs that show an alternation between expressions with a high degree of Accessibility and expressions with a medium or low degree of Accessibility differ from CRCs that are mostly made up of high Accessibility expressions.

7. Discussion

In this study we presented an analysis of CRCs following a linear approach. We proposed a method that applies sequence analysis to CRCs and allows for a fine-grained analysis of the chain composition. We then demonstrated

that text genres and the semantic nature of the referent influenced chain composition.

This method allowed us to obtain classes that are linguistically coherent and that reflect different strategies for introducing and maintaining the referent. We showed that the description of chain composition using a linear approach facilitates the identification of patterns of CRCs which correspond to different types of chain.

For future research, the study of these patterns in a different corpus annotated for CRCs will evaluate both the method and the outline of our first typology of CRCs. A first glimpse can be given from the Democrat corpus (Landragin 2015, Quignard et al. 2021), which is composed of narrative and non narrative French texts from the 11th century to the present. It is annotated following a bottom-up approach which consists in identifying all the referential expressions and then associating to each a referential identifier. CRCs are automatically reconstructed in a second step. The six patterns made out in the present study cover 17.71% of the CRCs in 19th-21st century texts (750 out of 4,236). This coverage is lower than in Annodis (38%), possibly because the CRCs annotated in the Democrat corpus are not solely topical CRCs. This first comparison may suggest that the six patterns are likely to identify topical CRCs. A qualitative analysis of Democrat CRCs is necessary to assess such a hypothesis.

As for classes, the 3-classes and the 5-classes configurations in the Democrat corpus shows the same correlations with textual genre, with a conclusive chi-squared test (p-value < 2.2e-16). These first results confirm the tendencies found in the AnnoDis corpus and suggest that the method is reliable and reproducible. A full application of our sequence analysis to the Democrat corpus will refine the description of CRCs by providing more of them, with greater diversity in terms of composition, text type and genre.

This will also allow us to incorporate other features in the analysis, in addition to the grammatical category, such as the syntactic function of the mention, and to combine different features to represent the sequences of mentions. Finally, a more detailed analysis following the linear approach will enable us to provide a description of referential strategies, by considering parameters such as discourse structure (presence of headings, paragraph breaks) and the presence of competing referents, in order to study their impact on the choice of referential expressions. A promising direction for future work consists in taking into account the factors proposed by the cognitive theories on referential form and memory states, such as Accessibility Theory (Ariel 2001), Centering Theory (Walker et al. 1998) and the Giveness Hierarchy (Gundel et al. 1993).

REFERENCES

Abbott, Andrew. 1995. Sequence analysis: New methods for old ideas. *Annual review of sociology, 21*(1): 93-113.

Ariel, Mira. 2001. Accessibility Theory: An Overview. In *Text Representation: Linguistic and psycholinguistic aspcts* [Human Cognitive Processing], Ted J.M Sanders, Joost Schilperoord & Wilbert Spooren (eds), 29-87. Amsterdam: John Benjamins Publishing Company.

Ariel, Mira. 1990, *Accessing Noun-Phrase antecedents*. London/New York: Routledge.

Asher, Nicholas, Philippe Muller, Myriam Bras, Lydia Mai Ho-Dac, Farah Benamara, Stergos Afantenos, & Laure Vieu. 2017. ANNODIS and Related Projects: Case Studies on the Annotation of Discourse Structure. In *Handbook of Linguistic Annotation*. Nancy Ide & James Pustejovsky (eds), 1241-1264. Dordrecht: Springer Netherlands.

Biber, Douglas, Connor, Ulla & Upton, Thomas A. 2007. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure.* Amsterdam: John Benjamins Publishing.

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

Dietrich, Julia, Andersson, Hakan & Salmera-Aro, Katariina. 2014. Develipmental Psychologists' Perspective on Pathways Through School and Beyon. In *Advances in Sequence Analysis: Theory, Method, Applications (Vol. 2)*. Philippe Blanchard, Felix Bühlmann, & Jacques-Antoine Gauthier (eds). Cham: Springer International Publishing.

Fasang, Annette Eva. 2014. New perspectives on Family Formation: What Can We Learn from Sequence Analysis?. In *Advances in Sequence Analysis: Theory, Method, Applications (Vol. 2)*. Philippe Blanchard, Felix Bühlmann, & Jacques-Antoine Gauthier (eds). Cham: Springer International Publishing.

Federzoni, Silvia, Ho-Dac, Lydia-Mai & Rebeyrolle, Josette. 2020. Les chaînes topicales dans la ressource ANNODIS. In *7^{e} Congrès Mondial de Linguistique Française (Montpellier, France)* [SHS Web of Conferences 78]. Franck Neveu, Bernard Harmegnies, Linda Hriba, Sophie Prévost & Agnes Steuckardt & (eds). Les Ulis, EDP Sciences.

Fossard, Marion, Achim, Amélie M., Rousier-Vercruyssen, Lucie, Gonzalez, Sylvia, Bureau, Alexandre, & Champagne-Lavau, Maud. 2018. Referential Choices in a Collaborative Storytelling Task: Discourse Stages and Referential Complexity Matter. *Frontiers in Psychology* 9 (février): 176.

Gabadinho, Alexis, Ritschard, Gilbert, Müller, Nicolas S. & Studer, Matthias. 2011. Analyzing and Visualizing State Sequences in *R* with TraMineR. *Journal of Statistical Software* 40 (4).

Gabadinho, A., G. Ritschard, M. Studer and N. S. Müller (2009). Mining Sequence Data in R with the TraMineR package: A user's guide. Department of Econometrics and Laboratory of Demography, University of Geneva

Gundel, Jeanette K., Hedberg, Nancy & Zacharski, Ron. 1993. Cognitive status and the form of referring expressions in discourse. *Languages* 69(2): 274-307.

Goutsos, Dyonisos. 1996. A model of sequential relations in expository test. *Text*, 16(4): 501–533.

Halliday, Michael AK. 1985. *An introduction to Functional Grammar*. London: Edward Arnold.

Halliday, Michael AK. & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Halpin, Brendan. 2010. Optimal matching analysis and life course data: The importance of duration. *Sociological Methods & Research*, 38(3): 365–388.

Kibrik, Andrej A. 2011. *Reference in Discourse*. New York: Oxford University Press.

Kunz, Kerstin & Lapshinova-Koltunski, Ekaterina. 2015. Cross-linguistic analysis of discourse variation across registers. *Nordic Journal of English Studies* 14 (1): 258-288.

Landragin, Frédéric. 2015. Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT), Bulletin de l'Association Française pour l'Intelligence Artificielle, AFIA, 11-15.

Lapshinova-Koltunski, Ekaterina & Kunz, Kerstin. 2020. Exploring Coreference Features in Heterogeneous Data. In *Proceedings of the First Workshop on Computational Approaches to Discourse*: 53-64. Online: Association for Computational Linguistics.

Lesnard, Laurent. 2014. Using Optimal Matching Analysis in Sociology: Cost Setting and Sociology of Time. In *Advances in Sequence Analysis: Theory, Method, Applications (Vol. 2)*. Philippe Blanchard, Felix Bühlmann, & Jacques-Antoine Gauthier (eds). Cham: Springer International Publishing. Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8): 707–710.

Levenshtein, V.I. (1966) Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10, 707-710.

Longo, Laurence & Amalia Todirascu. 2014. Vers une typologie des chaînes de référence dans des textes administratifs et juridiques: *Langages* N° 195 (3): 79-98.

Longo, Laurence & Todirascu, Amalia. Genre-based Reference Chains Identification for French. *Investigationes Linguisticae*, Adam Mickiewicz University, 2010, XXI: 57-75.

Nedoluzko, Anna & Lapshinova-Koltunski, Ekaterina. 2016. Contrasting Coreference in Czech and German: From Different Frameworks to Joint Results. In *Computational Lingistics and Intellectual Technologies: Proceedings of the 22nd International Conference Dialogue-21*. Moscow, Russia.

Obry, Vanessa, Julie Glikman, Céline Guillot-Barbance & Bénédicte Pincemin. 2017. Les chaînes de référence dans les récits brefs en français : étude diachronique (XIIIe-XVIe s.). *Langue française* 195 (3): 91-110.

Péry-Woodley, Marie-Paule, Afantenos, Stergos, Ho-Dac, Lydia-Mai & Asher, Nicholas. 2011. La ressource ANNODIS, un corpus enrichi d'annotations discursives, *Revue TAL*, 52 (3): 71-101.

Péry-Woodley, Marie-Paule, Ho-Dac, Lydia-Mai, Rebeyrolle, Josette, Tanguy, Ludovic & Fabre, Cécile. 2017. A corpus-driven approach to discourse organisation: from cues to complex markers. *Dialogue & Discourse* 8 (1): 66-105.

Quignard, Matthieu, Le Mené, Marine & Landragin, Frédéric. (2021). Élaboration du corpus DEMOCRAT : procédures d'annotation et d'évaluation. *Langages*, 224, 25-46.

Recasens, Potau, Marta. 2010. Coreference: Theory, Annotation, Resolution and Evaluation. Universitat de Barcelona.

Robette, Nicolas. 2011. *Explorer et décrire les parcours de vie: les typologies de trajectoires*. Les collections du CEPED. Paris: CEPED.

Rousier-Vercruyssen, Lucie & Landragin, Frédéric. 2019. Interdistance et instabilité au sein des chaînes de référence : indices textuels ? *Discours*, n° 25 (décembre).

Schnedecker, Catherine. 2021. *Les chaînes de référence en français*. Collection L'essentiel français. Paris: Éditions Ophrys.

Schnedecker, Catherine. 2005. Les chaînes de référence dans les portraits journalistiques : éléments de description: *Travaux de linguistique* no 51 (2): 85-133.

Schnedecker, Catherine & Landragin, Frédéric. 2014. Les chaînes de référence : présentation: *Langages* N° 195 (3): 3-22.

Studer, Matthias & Ritschard, Gilbert. 2014. *A comparative review of sequence dissimilarity measures*. Geneva, Switzerland.

Urieli, Assaf. 2013. Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit. Université Toulouse 2 Le Mirail.

Uryupina, Olga, Kabadjov, Mijail & Poesio, Massimo. 2016. Detecting Non-Reference and Non-Anaphoricity. In *Anaphora Resolution*, Massimo Poesio, Roland Stuckardt & Yannick Versley (eds), 369-392. Theory and Applications of Natural Language Processing. Berlin, Heidelberg: Springer Berlin Heidelberg.

Walker, Marilyn A., Joshi, Aravind, K. & Prince, Ellen F. 1998. *Centering Theory in Discourse*. Clarendon Press.

Weischedel, Ralph, Palmer, Martha, Marcus, Mitchell, Hovy, Eduard Pradhan, Sameer, Ramshaw, Lance, Xue, Nianwen, Taylor, Ann, Kaufman, Jeff, Franchini, Michelle, El-Bachouti, Mohammed, Belvin, Robert & Houston, Ann. 2013 OntoNotes Release 5.0 LDC2013T19. Web Download. Philadelphia: Linguistic Data Consortium.