

Referring expressions and coreference chains in French:

annotation strategies, annotating tools,
and annotated resources

Frédéric Landragin

Text Structure and Corpus Linguistics

Prague, November, 12th 2018



CC Attribution 4.0 International





Content

- Issues and objectives
- Reference, referring expressions, and coreference chains
- Computer-aided corpus linguistics for the analysis of referring expressions and coreference chains
- A framework: the ANR “Democrat” project
- The Democrat corpus and its annotation
- Natural language processing: the challenge of the automatic identification of coreference chains
- Future works



Issues and objectives

Three objects of study

Comme tout avait brûlé – **la mère**, les meubles et les photographies de **la mère** -, pour **Fabre** et **le fils Paul** c'était tout de suite beaucoup d'ouvrage : toute cette cendre et ce deuil, **déménager**, **courir** **se refaire** dans les grandes surfaces. **Fabre** trouva trop vite quelque chose de moins vaste, deux pièces aux fonctions permutables sous une cheminée de brique dont l'ombre donnait l'heure, et qui avaient ceci de bien d'être assez proches du quai de Valmy.

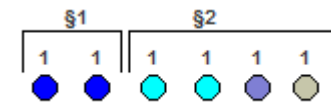
Le soir après le dîner, **Fabre** parlait à **Paul** de **sa mère**, **sa mère à lui Paul**, parfois dès le dîner. Comme **on** ne possédait plus de représentation de **Sylvie Fabre**, **il** s'épuisait à vouloir **la** **décrire** toujours plus exactement : au milieu de la cuisine naquirent des hologrammes que dégonflait la moindre imprécision. Ça ne se rend pas, soupirait **Fabre** en **posant** une main sur **sa** tête, sur **ses** yeux, et le découragement **l'**endormait. Souvent ce fut à **Paul** de **déplier** le canapé convertible, **transformant** les choses en chambre à coucher.

The succession of referring expressions



1 = Sylvie Fabre 2 = Mr. Fabre 3 = Paul Fabre
 4 = {Mr. Fabre, Paul Fabre}

Coreference chain that concerns Sylvie Fabre :



la mère – la mère – sa mère – sa mère – Sylvie Fabre – la



Definitions and objectives

- References: access vs. evocation of a referent
 - linguistic **expression** that refers to a referent, and makes this referent a discourse entity that is involved in the syntactic, semantic, and informational structures of the sentence
 - linguistic **clue** (e.g. morpheme, zero subject) that evokes a referent, without really referring to it, but that contributes to its salience
- References succession
 - succession of referring expressions in the text
 - study of the transitions from one referent to another one, of associative anaphora...
 - towards a typology of referential transitions: continuation on the same referent, bifurcation, confrontation of two referents...
- Coreference chains
 - succession of expressions and clues that concern the same referent
 - study of the typologies of coreference chains



Underlying issues

- Nature of the referring expressions
 - what refers in a text?
 - what evokes a referent without referring?
 - if several degrees of reference are distinguished, how can they be taken into account in a corpus annotation methodology?
- Nature of the coreference chains, and links with the suite of references
 - how does a chain start? end?
 - what are the archetypal chains? the typologies of chains?
 - how do the chains intersect each others in the text?
 - is it possible to predict “templates” for the suite of references?
 - are there correlations between chains typologies and syntactic, semantic, and pragmatic characteristics?
 - is it possible to deduce an operational definition of salience?



First steps of the work

- Identify and categorize the referents (world objects)
- Connect referents to each other (groups, individuals)
- Identify and categorize the referring expressions
- Connect referring expressions that refer to the same referent, i.e. build on the coreference chains
- Characterize the coreference chains



Scientific issues

- To propose an “integrated” model of reference
 - that takes into account reference and coreference from the point of view of the discourse and not only locally
 - that is enriched by comparisons with other languages (contrastive approach) and with several language states (diachronic approach)
 - that takes into account the text genre
- To bridge the gap between linguistic theories and natural language processing techniques
 - we annotate a corpus to provide data for machine learning techniques
 - we highlight referring phenomena that have often been neglected by natural language processing works
- To propose the first *end-to-end* system for the identification of coreference chains in French



Expected contributions and benefits

- To provide enriched data and new knowledge about the French language, that will be available for the whole community
- To provide new tools and new visualization processes for the manipulation of these data and knowledge
- To provide new methods for the linguistic and statistical analysis of coreference chains
- To represent the French language in NLP international evaluation challenges dealing with the identification of coreference chains
- to contribute to Digital Humanities
 - perpetuation of linguistic data, standardization of linguistic data,
 - representation of the French language in the current DH efforts,
 - contribution to didactics, and teaching French as a foreign language



Reference,
referring expressions,
and coreference chains

Definition problems

- « Le village était désert. Il semblait abandonné. La place principale était vide. Elle en paraissait triste. Tout reprendrait vie le lendemain matin, Ø repartirait de zéro : le village s’animerait, la place se remplirait de monde »
(“the village” – “the main square” – etc.)
- What are the referents? Does « en » have a referent?
- What are the referring expressions?
On what criteria should the zero subjects be taken into account?
- If we make distinctions, what are the referring chains?
the coreference chains? the anaphoric chains?
- What are the antecedents? do they correspond to the last mentioned referring expression, or to the first mention?

Problems related to the identification of referring expressions

- Reference is a linguistic question, which has consequences on the annotation procedure
- A short analysis of another constructed example:
 - « Pierre et Paul ont chacun eu un fils cette année. Il se trouve qu'ils ont la même nourrice. »

“Peter and Paul each had a son this year. It turns out that they have the same nanny”
 - characters: Peter, son of Peter, Paul, son of Paul, the nanny
 - “Peter and Paul”: because of the coordination, should we consider that there is here a reference to a group of two characters?
 - “a son”: is it a reference?
 - “they”: apparently refers to the group of the two sons, but this group has not been mentioned before. Is it a first mention?
 - “each”?

“Solid” expressions and “attenuated” expressions

- In addition to referring expressions, some words or morphemes may participate to the coreference chains
 - the marks of agreement in gender and/or number (which, even if not referring, recall the referent existence and thus participate to the coreference chains)
 - in “John lies down and sleeps”, the “-s” recall that the referent is singular
 - is it a phenomenon to annotate? using a specific category?
 - zero subjects (in particular for infinitive and participle forms)
 - the advantage of annotating them is that they can be salient and thus contribute strongly to the study – if not to the coreference chains themselves
 - we can then compare examples like “he came in and took his hat”
and “he came in and he took his hat”
 - if we annotate zero forms, it is necessary to choose a technical solution such as annotating the verb itself (since it is not reasonable to annotate a space)
 - pronominal constructions, etc.

The case of attributes and labels

Je suis sursitaire, âgé de 24 ans, et je suis marié à une veuve de 44 ans, laquelle a une fille qui en a 25. Mon père a épousé cette fille. A cette heure, mon père est donc devenu mon gendre, puisqu'il a épousé ma fille^[1]. De ce fait, ma belle-fille^[2] est devenue ma belle-mère, puisqu'elle est la femme de mon père.

Ma femme et moi avons eu en Janvier dernier un fils. Cet enfant est donc devenu le frère de la femme de mon père, donc le beau-frère de mon père. En conséquence, mon oncle, puisqu'il est le frère de ma belle-mère. Mon fils est donc mon oncle.

[1] A step is missing: “the daughter of my wife” becomes “my daughter”...

[2] The indirect referent is ignored, as well as in “a parricide”

- Some expressions have a reference, others work like labels and are not really referential

The case of attributes and labels

I am a baker, 24 years old, and I am married to a 44 years old widow, who has a daughter who is 25. My father married this girl. At that time, my father became my son-in-law, since he married my daughter^[1]. By consequence, my daughter-in-law^[2] became my mother-in-law, since she is my father's wife.

My wife and I had a son last January. So this child became the brother of my father's wife, and therefore my father's brother-in-law; consequently, my uncle, since he is my mother-in-law's brother. So my son is my uncle.

[1] A step is missing: "the daughter of my wife" becomes "my daughter"...

[2] The indirect referent is ignored, as well as in "a parricide"

- Some expressions have a reference, others work like labels and are not really referential



Problems to delimitate a referring expression

Some examples of first mentions:

1. **President Emmanuel Macron** said...
2. **The President of the Republic**, Emmanuel Macron, said...
3. **Emmanuel Macron, President of the Republic**, said...
4. **Emmanuel Macron** – yes, yes! – **President of the Republic**, said...
5. **The President of the Republic, who** is Emmanuel Macron, said...
6. **Emmanuel Macron** is the first President to say...

Several possibilities depending on the example:

- a single referring expression (that sometimes groups several phrases)
- several referring expressions
- several expressions, the first one being the only one that is referential
- several expressions, the most “direct” (proper name) being considered as referential

Problems with the annotation:

- it is sometimes difficult to determine precise limits
- the example with a discontinuous text span poses technical problems

Assigning a referent may be impossible

- Some pronouns may remain ambiguous, even when taking into account the encyclopaedic knowledge of the reader
- Example: abstract of the film *The Counterfeiters of Paris*

Eric Masson, un "demi-sel", est devenu l'amant de la belle **Solange Mideau**, femme d'un graveur raté. Eric veut se servir de **Robert Mideau** pour monter, à **son** insu, un trafic de fausse monnaie. Il s'associe à **Charles Lepicard**, tenancier d'une ancienne maison close, et à **Lucas Malvoisin**, l'homme d'affaires de celui-ci. Charles et Lucas n'ont pas grande confiance en Eric, mais Solange **leur** promet **son** concours. Elle souhaite en effet mener la grande vie. Avec l'accord de **ses complices**, Charles contacte **Ferdinand Maréchal**, dit le Dabe, vieux truand célèbre qui s'est retiré dans une île des Tropiques. Il le décide à venir à Paris.

 - « à son insu » ("without his knowledge): Robert or Solange Mideau ?
 - « leur » ("them") : Charles (sure) + Lucas (sure) + Eric (possible)
 - « son concours » ("his/her help"): ambiguous between Solange and Eric
 - « ses complices » ("his accomplices"): Lucas (sure) + Solange (probable) + Eric (?)

Assigning a referent may be impossible

- Some pronouns may remain ambiguous, even when taking into account the encyclopaedic knowledge of the reader
- Example: abstract of the film *The Counterfeiters of Paris*

Eric Masson, a hoodlum, became the lover of the beautiful **Solange Mideau**, who is married to a failed engraver. Eric wants to use **Robert Mideau** to set up a counterfeit currency trade without **his** knowledge. He joins **Charles Lepicard**, owner of a former brothel, and **Lucas Malvoisin**, his businessman. Charles and Lucas do not have much confidence in Eric, but Solange promises **them his/her** help. She wants to lead a high life. With the agreement of **his accomplices**, Charles contacted **Ferdinand Maréchal**, known as “the boss”, a famous old gangster who had retired to a tropical island. He decides to come to Paris.

- « à son insu » (“without his knowledge): Robert or Solange Mideau ?
- « leur » (“them”) : Charles (sure) + Lucas (sure) + Eric (possible)
- « son concours » (“his/her help”): ambiguous between Solange and Eric
- « ses complices » (“his accomplices”): Lucas (sure) + Solange (probable) + Eric (?)

Assigning a referent can evolve during the reading process

- The reference of a referring expression may change...

L'ancien président de la République de Côte d'Ivoire, **Henri Konan Bédié** et **son épouse** ont reçu à dîner l'ancien Premier ministre **Alassane Dramane Ouattara** et **son épouse**, le 23 septembre. La rencontre très médiatisée avait un objectif, celui de montrer que **les héritiers du premier président de Côte d'Ivoire** peuvent se retrouver pour reconquérir le pouvoir. **Les deux leaders** ont l'habitude de se voir et de s'appeler depuis le déclenchement, le 19 septembre 2002 de la rébellion en Côte d'Ivoire. A Paris, à Abidjan, à Accra, **les deux hommes** se côtoient, mais dans des cadres formels. **Leur** rencontre en soi n'est donc pas un événement, sauf qu'**ils** ont voulu donner à cette entrevue un cachet particulier. Les retrouvailles autour d'un même idéal politique que commande la mémoire du "Vieux" dont **ils** se réclament. [...] Mais après que **tout le monde** ait perdu le pouvoir, en faveur d'**un autre héritier, le général Robert Guéi**, par un coup d'Etat en décembre 1999, la gestion du pays semble échapper aux "**enfants**".

- at the beginning: « les héritiers » (“the heirs”) = H.K.B. + A.D.O.
- but it is a fuzzy group: « les héritiers » = H.K.B. + A.D.O. + their wives
- then: « un autre héritier » (“another heir”) = R.G.,
so « les héritiers » (“the heirs”) = a group with fuzzy boundaries,
which includes at least the three men that are mentioned

Assigning a referent can evolve during the reading process

- The reference of a referring expression may change...

The former President of the Republic of Côte d'Ivoire, **Henri Konan Bédié** and **his wife**, hosted former Prime Minister **Alassane Dramane Ouattara** and **his wife** for dinner on September, 23. The highly mediatized meeting had one objective, that of showing that **he heirs of the first President of Côte d'Ivoire** can meet to regain power. **The two leaders** have been in the habit of seeing and calling each other since the outbreak of the rebellion in Côte d'Ivoire on September, 19, 2002. In Paris, in Abidjan, in Accra, **the two men** rubbed shoulders, but in formal settings. **Their** meeting in itself is therefore not an event, except that **they** wanted to give this meeting a special touch. The reunion around the same political ideal that the memory of the "Old man" of whom **they** claim to be part demands.

[...] But after **everyone** had lost power to **another heir, General Robert Guéi**, in a coup d'état in December 1999, the country's management seemed to escape the "**children**".

- at the beginning: « les héritiers » ("the heirs") = H.K.B. + A.D.O.
- but it is a fuzzy group: « les héritiers » = H.K.B. + A.D.O. + their wives
- then: « un autre héritier » ("another heir") = R.G.,
so « les héritiers » ("the heirs") = a group with fuzzy boundaries,
which includes at least the three men that are mentioned




Consequences on the annotation: several strategies are possible

- 1. We focus on linguistic forms**, without taking into account the possible subsequent reinterpretations (linear strategy)
 - advantage: theoretically, interpretative biases are reduced and the steps of interpretation are reported in the annotations
 - disadvantages: assigning a referent with the linguistic form as only basis is illusory, because our encyclopaedic knowledge is constantly involved; annotating something we know wrong is not very relevant...
- 2. We focus on the concepts**, and we only annotate after having understood all the text and having calculated all the references
 - advantage: we get closer to the reality behind the text
 - disadvantage: we ignore the stylistic effects intended by the writer
- 3. We start from the concepts and we extend to the possible interpretations**, using a dedicated attribute (immediate vs. delayed)
 - advantage: we model reference in a satisfying and complete manner
 - disadvantage: writing an annotation manual can therefore be complex

Consequences on the annotation: it is a fuzzy process...

- We do not therefore try to assign a referent at any price, but we take into account the possibilities of ambiguity, imprecision, vagueness
- The notion of fuzziness is taken into account, on the one hand for the determination of groups (strict groups *versus* fuzzy groups), on the other hand for the relationship “belongs to” (strict *versus* fuzzy)
- These aspects can be modelled with the theory of Fuzzy Sets (Zadeh, but also Kaufmann, Prade...)
 - « Solange Mideau » (individual reference without any problem): A_{strict}
 - « son concours » (ambiguous: Solange or Eric): $A_{\text{strict}} \text{ ou } B_{\text{strict}}$
 - « le cave » : A_{fuzzy}
 - « Charles et Lucas » (constructed group): $\text{group}_{\text{strict}} \{ A_{\text{strict}} ; B_{\text{strict}} \}$
 - « ses complices »: $\text{group}_{\text{strict}} \{ A_{\text{strict}} ; B_{\text{strict}} ; C_{\text{fuzzy}} \}$
 - « les héritiers »: $\text{group}_{\text{fuzzy}} \{ A_{\text{strict}} ; B_{\text{strict}} ; C_{\text{fuzzy}} ; D_{\text{fuzzy}} \}$



**Computer-aided linguistics
for the analysis of references
and coreference chains**

Visualization of chains

Comme tout avait brûlé **la mère**, les meubles et les photographies de **la mère**, pour Fabre et le fils Paul c'était tout de suite beaucoup d'ouvrage : toute cette cendre et ce deuil, déménager, courir se refaire dans les grandes surfaces. Fabre trouva trop vite quelque chose de moins vaste, deux pièces aux fonctions permutable sous une cheminée de brique dont l'ombre donnait l'heure, et qui avaient ceci de bien d'être assez proches du quai de Valmy.

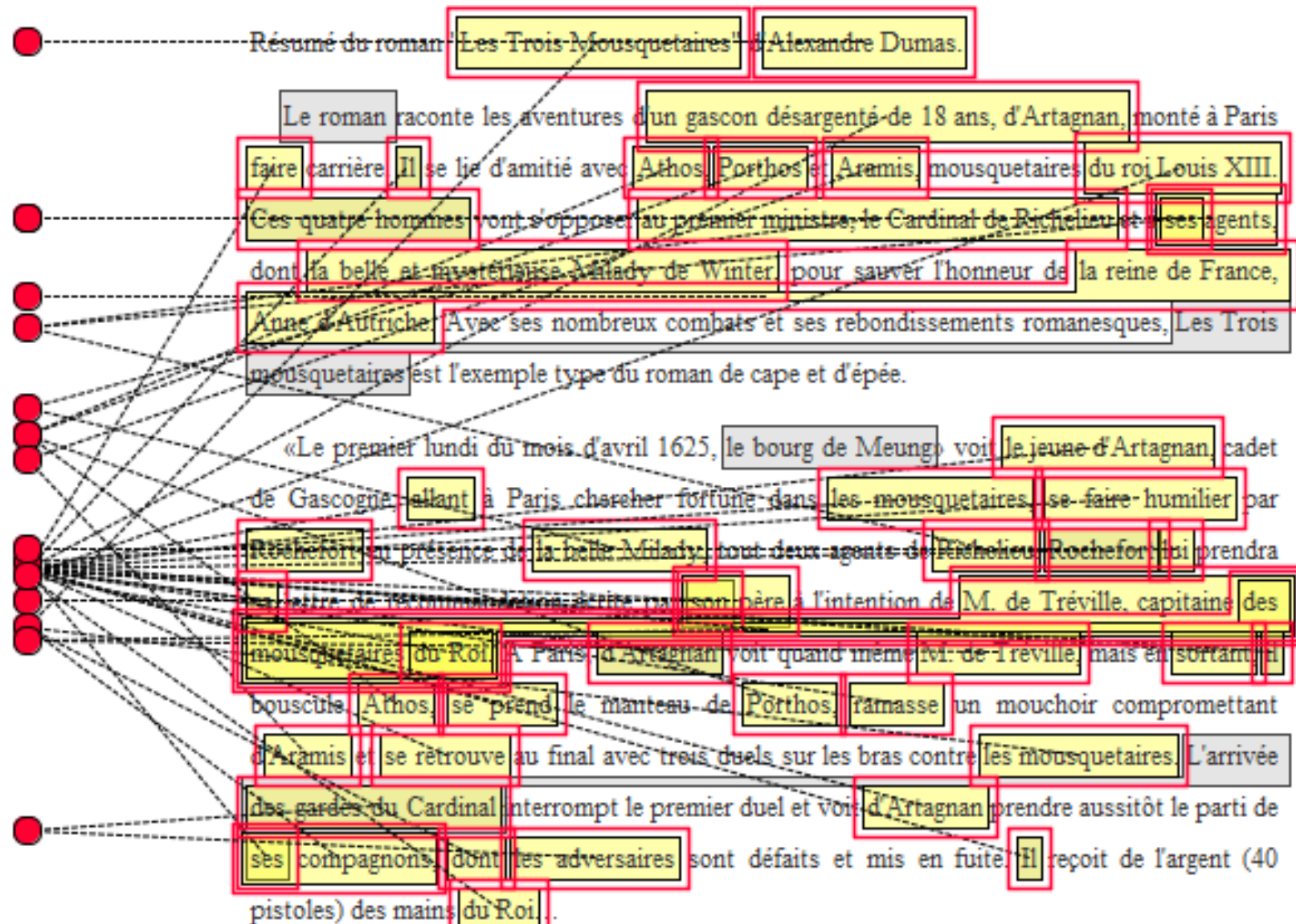
Le soir après le dîner, Fabre parlait à Paul de **sa mère**, **sa mère** à lui Paul, parfois dès le dîner. Comme on ne possédait plus de représentation de **Sylvie Fabre**, ils s'épuisaient à vouloir **la** décrire toujours plus exactement : au milieu de la cuisine naquirent des hologrammes que dégonflait la moindre imprécision. Ça ne se rend pas, soupirait Fabre en posant une main sur sa tête, sur ses yeux, et le découragement l'endormait. Souvent ce fut à Paul de déplier le canapé convertible, transformant les choses en chambre à coucher.

Le dimanche et certains jeudis, ils parlaient sur le quai de Valmy vers la rue Marseille, la rue Dieu, ils allaient voir **Sylvie Fabre**. **Elle** les regardait de haut, tendait vers eux le flacon de parfum Piver, Forvil, **elle** soupirait dans quinze mètres de robe bleue. Le gilet d'un soupirait trouait **la** manche. Il n'y avait pas d'autre image d'**elle**.

L'artiste Flers **l'**avait représentée sur le flanc d'un immeuble, juste avant le coin de la rue. L'immeuble était plus maigre et plus solide, mieux tenu que les vieilles constructions qui se collaient en grinçant contre lui, terrifiées par le plan d'occupation des sols. En manque de marquise, son porche saturé de moulures portait le nom (Wagner) de l'architecte-sculpteur gravé dans un cartouche en haut à droite. Et le mur sur lequel, avec toute son équipe, l'artiste Flers avait peiné pour figurer **Sylvie Fabre**, en pied, surplombait un petit espace vert rudimentaire, sorte de square sans accessoires qui ne consistait qu'à former le coin de la rue.

- In the short story *L'occupation des sols (Plan of Occupancy)* by Jean Echenoz, two referents are strongly linked:
- Sylvie Fabre
- a painting on a wall, representing Sylvie Fabre

All chains at once...



Manual study of the chains

Character	Coreference chain	Proper names proportion
Fabre, the father	Fabre – Fabre – Fabre – il – s' – Fabre – sa – ses – l' – Fabre – s' – Fabre – que – il – ses – le père – Fabre – son père – Fabre – le veuf – Fabre – se – il – Fabre – s' – le – ses – il – ses – son – il – Fabre – le père – Fabre – s' – il – se – lui-même – il – le père de Paul – Fabre	32% (13 of 41)
Paul Fabre, the son	le fils Paul – Paul – sa – sa – lui – Paul – Paul – Paul – qui – ta – Paul – Paul – sa – il – son – Paul – Paul – sa – se – Paul – il – Paul – lui – Paul – se – Paul – son fils – du fils – il – Paul – Paul – Paul	50% (16 of 32)
Group with the father and his son	se – on – ils – les – eux – on – se – on – leur – on – on – on – on – s' – on – s' – ils – s' – leur – ils – leurs – on – s' – se – on – on – on – on – on – on	0%
The mother	(tout) – la mère – la mère – sa mère – sa mère à lui – Sylvie Fabre – la – elle – l' – Sylvie Fabre – Sylvie – elle – Sylvie	31% (4 of 13)
The painting	Sylvie Fabre – elle – elle – sa – ta mère – l'effigie – sa mère – Sylvie Fabre – Sylvie – son – Sylvie Fabre – son – ses – Sylvie – sa mère – l' – Sylvie	35% (6 of 17)
Flers	l'artiste Flers – son – l'artiste Flers – Flers	75%
The user	l'usager – l'usager – s' – l'usager – sa – il – se – son – sa – soi	0%
Jacqueline	une femme – qui – s' – celle – qui – j' – tu – Jacqueline – la femme – s' – qui – c'	8%

Computer-aided study of the chains

Analyse d'une chaîne de corréférence





Paramètres d'affichage Exporter les données Fusionner des types d'unités

Chaînes :
Corréférence

Unités :
Expression référentielle

Afficher les histogrammes de répartition

Champ à filtrer :
Position

	Initiale	49,711
	Médiane	42,486
	Finale	7,803
	<aucune valeur>	

Champ de la chaîne à analyser :
Nom du référent















Oui	M. Lantin	58,382
Oui	Mme Lantin	22,254
Oui	Bijoutier n°2	10,116
Oui	Bijoutier n°1	4,335
Oui	Groupe : M. et Mme Lantin 1	1,445
Oui	Mère de Mme Lantin	0,578
Oui	Les flâneurs	0,578
Oui	Commis du bijoutier n°1	0,578
Oui	2e épouse de M. Lantin	0,578
Oui	Sous-chef de M. Lantin	0,578
Oui	Groupe : Mme Lantin et	0,578

Paragraphe à filtrer :

Oui	Paragraphe3 " M. Lantin, ayant
Oui	Paragraphe4 " C'était la fille"

M. Lantin :

§3 M. L son l' §7 M. L ll il l' §8 ll §9 son le sa il la il §10

exporter en SVG

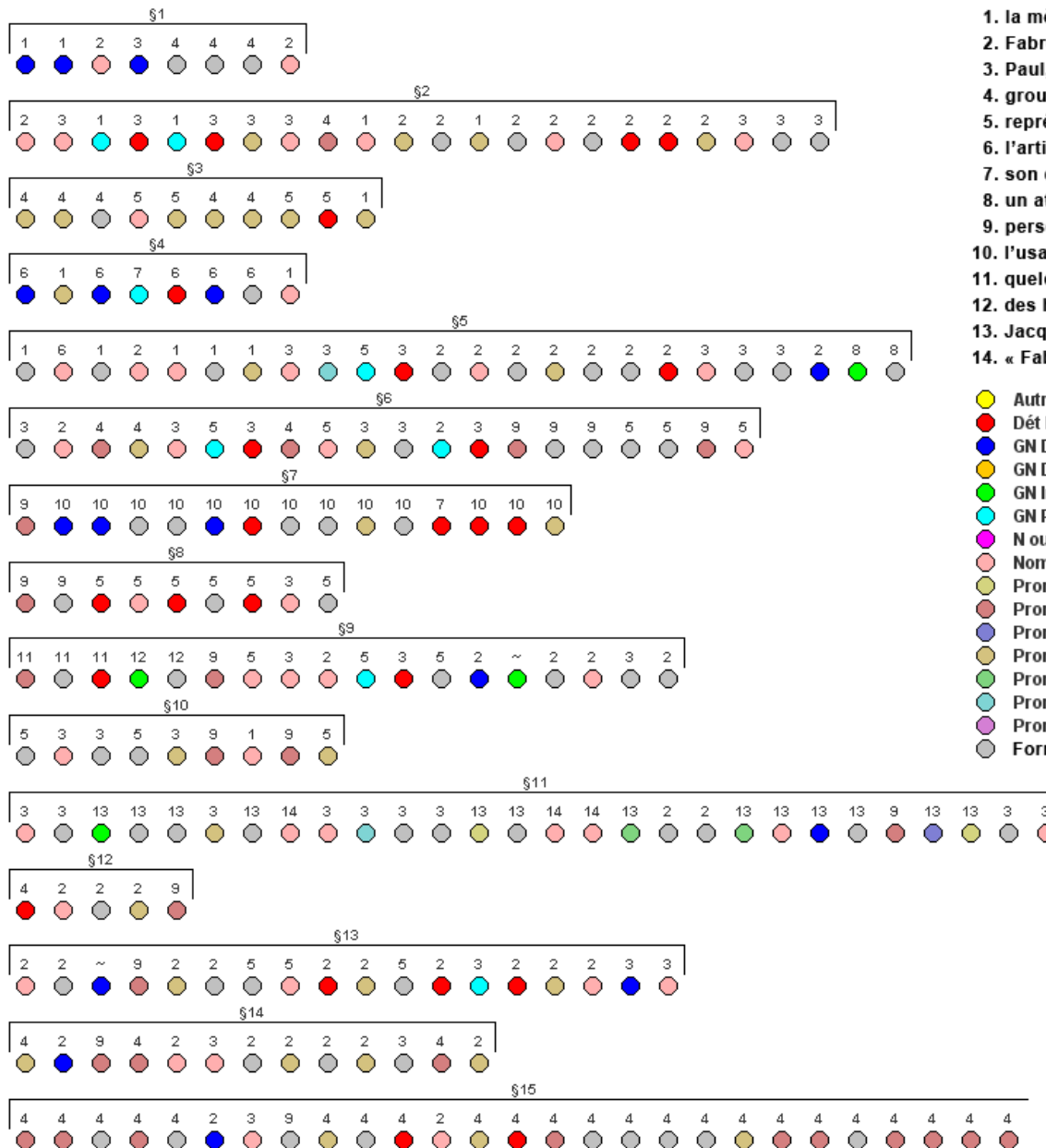
Statistique :

valeur de la chaîne ^	Finale	Initiale	aucune val...	Médiane
2e épouse de M. Lantin	0	100	0	0
Bijoutier n°1	6,667	53,333	0	40
Bijoutier n°2	17,143	57,143	0	25,714
Commis du bijoutier n°1	50	0	0	50
Groupe : M. et Mme Lantin 1	20	60	0	20
Groupe : Mme Lantin et sa mère	0	100	0	0
Les flâneurs	0	0	0	100
M. Lantin	5,941	50,99	0	43,069
Mère de Mme Lantin	0	0	0	100
Mme Lantin	7,792	45,455	0	46,753
Sous-chef de M. Lantin	0	0	0	100

Masquer cette valeur de la chaîne

Ne montrer que cette valeur de la chaîne

Study of the references succession

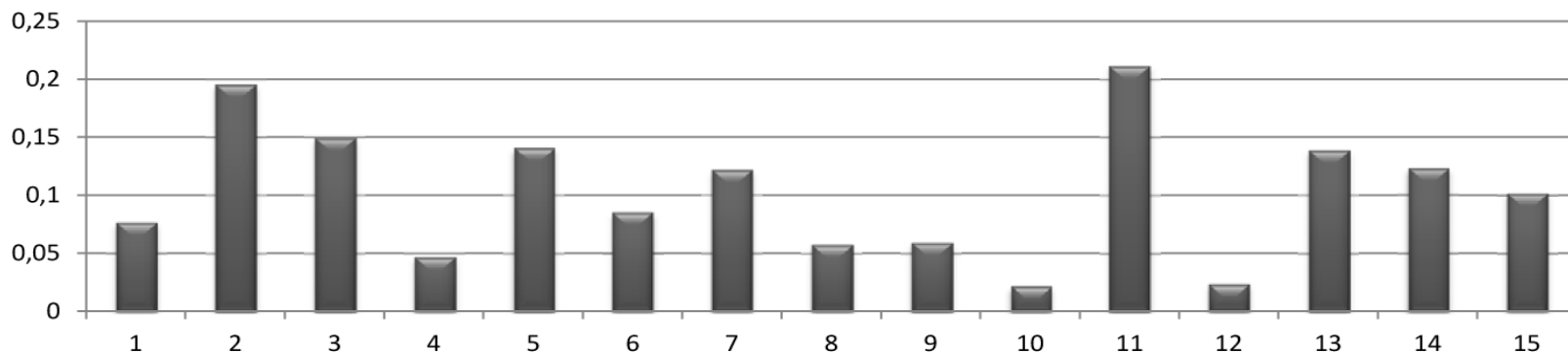


1. la mère
2. Fabre, le père
3. Paul, le fils
4. groupe formé par le père et le fils
5. représentation de la mère
6. l'artiste Fiers
7. son équipe
8. un attroupement
9. personne indéfinie
10. l'usager
11. quelqu'un
12. des hommes casqués de jaune
13. Jacqueline
14. « Fabre »

- Autre
- Dét Possessif
- GN Défini
- GN Démonstratif
- GN Indéfini
- GN Possessif
- N ou GN sans dét
- Nom Propre
- Pron Démonstratif
- Pron Indéfini
- Pron Interrogatif
- Pron Pers Anaphorique
- Pron Pers Déictique
- Pron Relatif
- Pron possessif
- Formes atténuées

Study of referential densities

Paragraph	Narrative content of the paragraph	Main characters
§1	Fire and relocation of the father and his son	father, son, mother
§2	New life (inside) for the father and the son	father, son, mother
§3	New life (outside) for the father and the son	father, son, painting
§4	The Wagner building and the painting (flashbacks)	Flers
§5	The painting (flashbacks), back to father and son	father, son, mother, painting
§6	End of common life + demolition	father, son, painting
§7	Declining of the nature space	user
§8	Damage to the place and to the painting	painting
§9	Construction of a new building	son
§10	End of the son's visits	son
§11	The son meets his father again, installed...	Jacqueline
§12	...in a new apartment	father
§13	Flashback on the father's move in	father
§14	Return of the son for the week-end	father, son
§15	Lunch, then scratching...	father and son as "on"



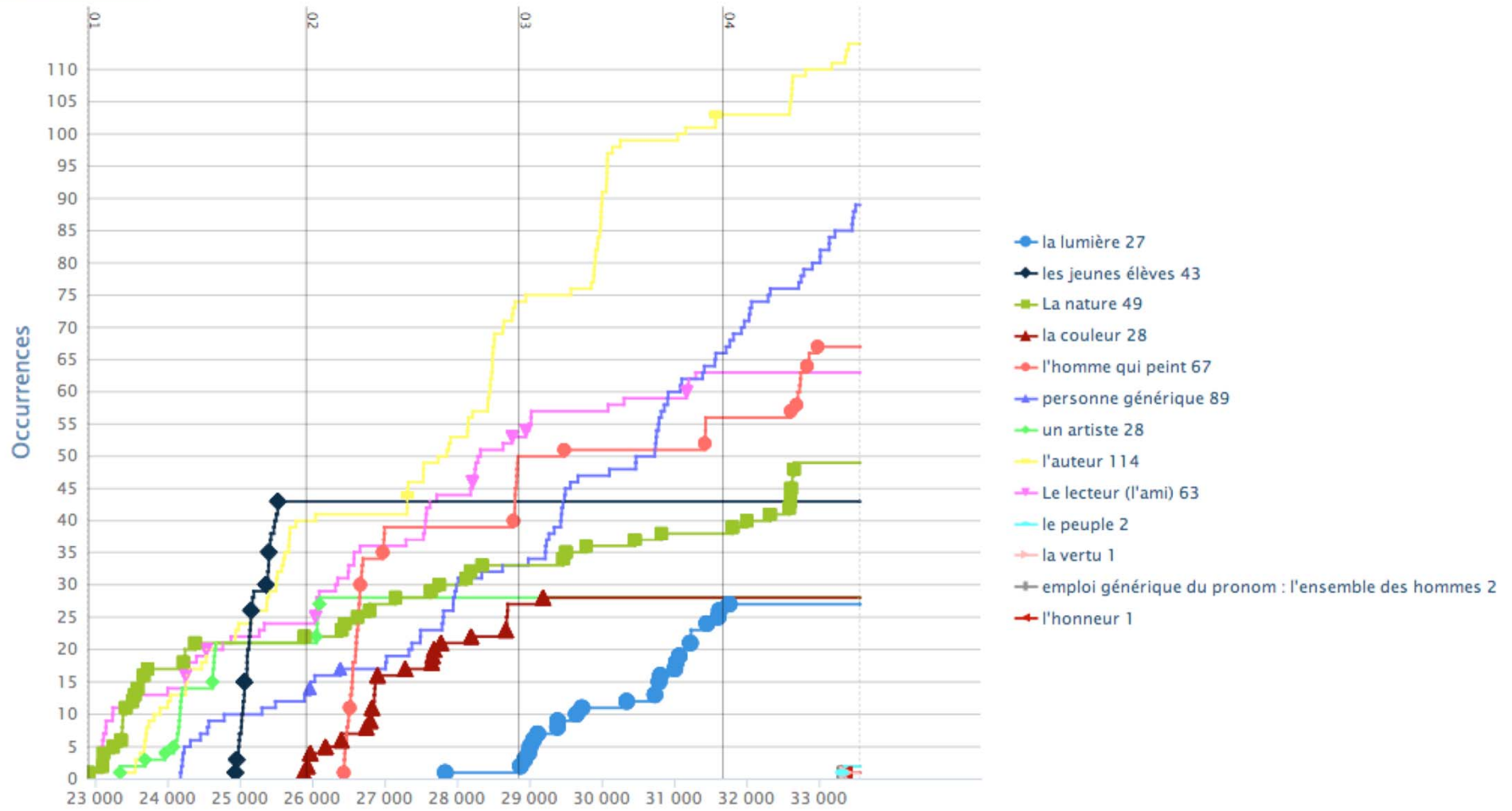
A concordancer applied to chains

Requête : Pivot: word

Clés de tri : #1 #2 #3 #4

text_id	Contexte gauche	Pivot	Contexte droit
Desperiers	et Polite. LES pages avoyent attaché l'oreille	à Caillette	avec un clou contre un posteau, et le povre Caillette c
Desperiers	avec un clou contre un posteau, et	le povre Caillette	demeuroit là, et ne disoit mot: Car il n'avoit point
Desperiers	le povre Caillette demeuroit là, et ne	disoit	mot: Car il n'avoit point d'autre apprehension, sinon
Desperiers	là, et ne disoit mot: Car	il	n'avoit point d'autre apprehension, sinon qu'il penso
Desperiers	Car il n'avoit point d'autre apprehension, sinon	qu'il	pensoit estre confiné là pour toute sa vie. Il passe un
Desperiers	sinon qu'il pensoit estre confiné là pour toute	sa	vie. Il passe un des Seigneurs de court, qui le
Desperiers	passé un des Seigneurs de court, qui	le	voit ainsi en conseil avec ce pillier, qui le fait incontine
Desperiers	ainsi en conseil avec ce pillier, qui	le	fait incontinent desgager de là: s'enquerant bien exp
Desperiers	expressement qui avoit fait celà, et qui	l'ha	mis là? Que voulez vous, un sot l'ha mis là
Desperiers	là? Que voulez vous, un sot	l'ha	mis là, un sot l'ha là mis. Quand on disoit
Desperiers	un sot l'ha mis là, un sot	l'ha	là mis. Quand on disoit, Ce ont esté les pages
Desperiers	disoit, Ce ont esté les pages,	Caillette	respondoit bien en son idiotisme, ouy ouy, ce ont est
Desperiers	esté les pages, Caillette respondoit bien en	son	idiotisme, ouy ouy, ce ont esté les pages. Sauras
Desperiers	, ce ont esté les pages. Sauras	tu	cognoistre lequel ce ha esté? ouy ouy, disoit Caillette
Desperiers	ce ha esté? ouy ouy, disoit	Caillette	, je say bien qui c'ha esté. L'escuyer par commandeme

Chains progression diagram





**A framework:
The ANR Democrat project**

At the beginnings...

November 2008	definition of the objectives of the “COREF” working group
December 2008	referring expressions and ambiguities
January 2009	evolving referents; strict and fuzzy groups
March 2009	methodology for the annotation of references
March 2009	relations between referents and Fuzzy Sets Theory
April 2009	types of referential transitions ; cinematographic metaphor
June 2009	types of referent introduction ; MMAX versus GLOZZ
November 2009	templates to determine referential transitions
December 2009	annotation methodology , annotation structure
January 2010	interactions between individuals and other entities
<i>January 2010</i>	<i>special session: TEI, ANANAS...</i>
<i>January 2010</i>	<i>Lattice seminar = first public presentation</i>

...there was a Lattice working group called “COREF”

February 2010	Centering Theory; annotation of salience
March 2010	coreference chains as theme markers; GLOZZ
April 2010	plurality, group, collective, collection; evocation of referents
May 2010	French-Hungarian contrastive study; diachronic preoccupations
June 2010	a single discourse centre <i>versus</i> several scales for salience
September 2010	NLP; pronouns, predications, attributions; ANALEC
October 2010	COREF project; ambiguities and under-determinations
December 2010	scope of a chain; referring to non-human entities
January 2011	functions names and attributive expressions
January 2011	<i>NLP special session: methods, algorithms, evaluation, projects</i>
February 2011	basic (“level 0”) annotation schema; labels and coreferences
March 2011	definite vs demonstrative; solid vs attenuated elements of a chain
March 2011	Lattice seminar = second public presentation

Then a first funded project: MC4



Modélisation Contrastive et Computationnelle des Chaînes de Coréférence

Produit le 15 juin 2015 par :

Langues, textes, traitements informatiques, cognition - UMR 8094 (LaTTiCe, Paris FR)

Description

Le corpus MC4 a été constitué par les membres participants du projet MC4. Le projet a pour but d'annoter les phénomènes référentiels, à savoir un ensemble défini d'indices présents dans le texte. Chacun de ces indices est nommé « maillon » et entre dans la constitution d'une « chaîne de référence ».

Le corpus écrit du projet MC4 comprend 8 textes, soit environ 18 000 mots et 3800 maillons. L'ensemble des textes réunis n'est pas homogène puisque constitué de textes en vers ou en prose, d'époques différentes, de longueur variable, correspondant ou non à l'ensemble de l'œuvre, à savoir : 6 récits du Gracial d'Adgar (12e s, vers), le premier livre des Quatre Livres des Rois (12e s, prose), La vie de Saint Thomas de Becket (12e s, vers), Li Estoires de Chiaus qui conquissent Coustantinoble de Robert de Clari (12e-13e s, prose), la Queste del saint Graal (13e s, prose), Les Bijoux et La

Téléchargement

**Licence Creative Commons
Attribution - Pas
d'Utilisation Commerciale -
Partage dans les Mêmes
Conditions 3.0 France**

Cette licence permet aux autres de remixer, arranger, et adapter votre œuvre à des fins non commerciales tant qu'on vous crédite en citant votre nom et que les nouvelles œuvres sont diffusées selon les mêmes conditions [conditions](#)

And now: ANR Democrat project

4-year project
funded by the
ANR (2016-2020)

Website:

<http://www.lattice.cnrs.fr/democrat/>

4 partners,
around 40
participants

ANR-15-CE38-0008

Projet ANR DEMOCRAT



MOTIVATIONS MODÈLE ET CORPUS LINGUISTIQUE OUTILLÉE SYSTÈME DE TAL PUBLICATIONS DU PROJET

LABORATOIRES PARTENAIRES



ORGANISMES TUTELLES



Présentation

DEMOCRAT est un projet financé par l'ANR pour 4 ans, entre 2016 et 2020. Il réunit des chercheurs issus de plusieurs laboratoires français, notamment Lattice (Paris), LILPa (Strasbourg), ICAR et IHRIM (Lyon). C'est un projet qui vise à développer les recherches sur la langue et la structuration textuelle du français via l'analyse détaillée et contrastive des chaînes de référence (instanciations successives d'une même entité) dans un corpus diachronique de textes écrits entre le 9ème et le 21ème siècle, avec des genres textuels variés. Le sigle DEMOCRAT signifie : DESCRIPTION et MODÉLISATION des Chaînes de Référence : outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement automatique.



Photo prise à TALN 2018 lors de la présentation d'un poster DEMOCRAT par Marine Delaborde, Loïc Grobol et Yoann Dupont (de gauche à droite).



Project participants

- **Partner 1: ENS Paris, Lattice laboratory**
 - Responsible: Frédéric Landragin, project leader
 - Initially 9 participants, currently 15 participants: 10 members of Lattice, 5 associate participants, members of other laboratories, a few Ph.D. students, including one funded by the project
- **Partner 2: University of Strasbourg, LiLPa laboratory**
 - Responsible: Catherine Schnedecker
 - Initially 10 participants, currently 13 participants: 10 members of LiLPa, 3 members of other laboratories, a few Ph.D. students and one post-doc funded by the project
- **Partner 3: ENS Lyon, ICAR and IHRIM laboratories**
 - Responsible : Céline Guillot-Barbance
 - Initially 7 participants, currently 9 participants: 3 members of ICAR, 4 members of IHRIM, including a engineer funded by the project



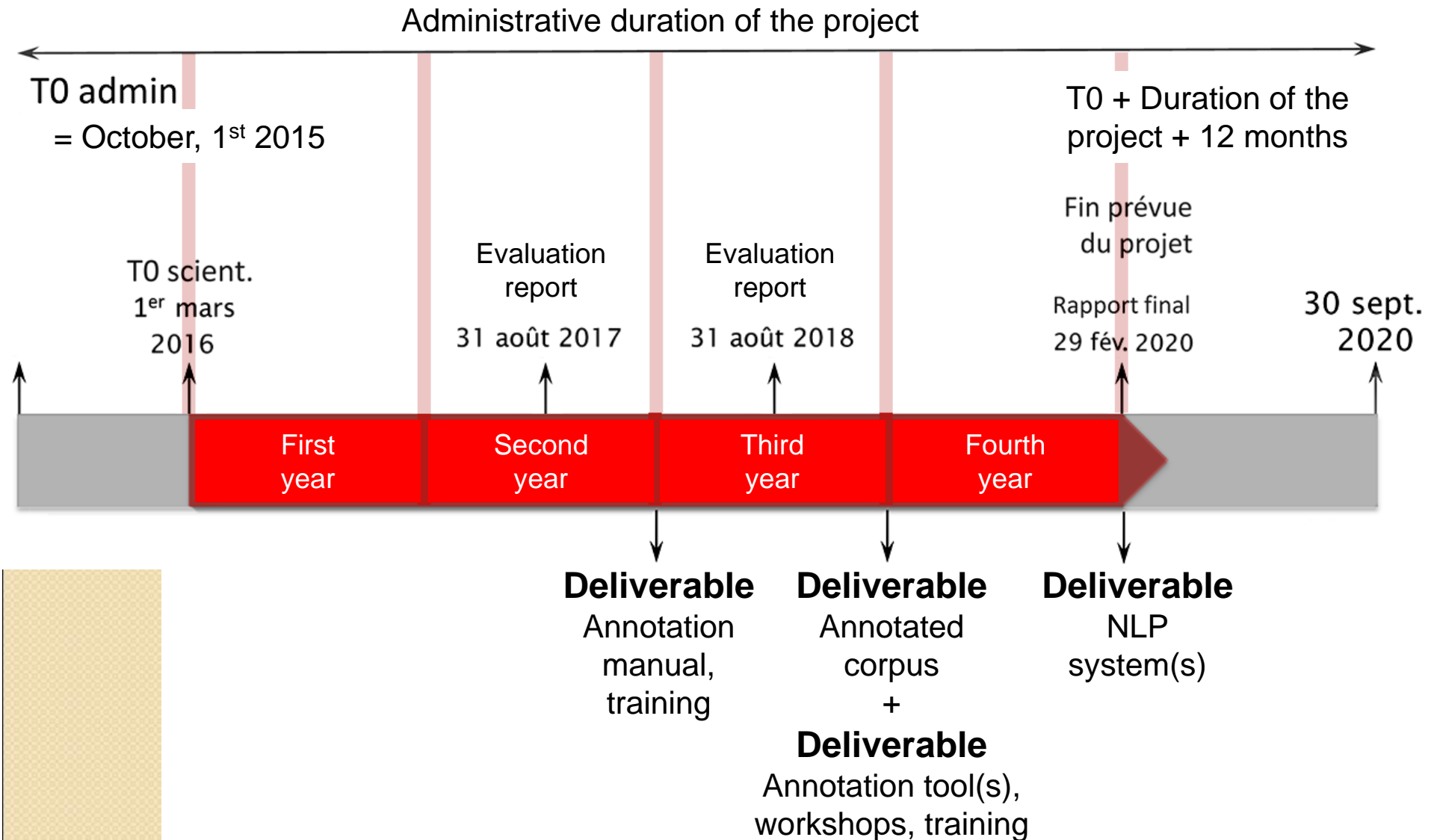
4 years

4 objectives

4 deliverables

1. Linguistic modelling (discursive, contrastive...)
no specific deliverable except a pseudo-deliverable “publications and formations”, which is common to the 4 objectives and spread over the 4 years of the project
2. Constitution of an annotated corpus
deliverable “annotation methodology” delivered in March 2018, which will lead to the deliverable “corpus” in March 2019
3. Design of an annotation and query tool
deliverable “TXM”
4. Design of an automatic detection system
deliverable “NLP”, with potentially several systems

Planning





Scientific highlights

- June 2015: workshop on “corpus approaches for the study of coreference chains” at the LiLPa
- March 2016: “kick-off” plenary meeting at the Lattice
- February 2017: plenary meeting at the Lattice
- November 2017: workshop on “coreference chains and text structure(s)” at the ENS of Lyon
- March 2018: plenary meeting at the Lattice
- March 2018: workshop on “contrastive approaches for the study of coreference chains” at the Lattice

Main publications





Work in progress

1. Discursive linguistic modelling

- Links between theory and corpus
- Links between coreference chains and text structures
- Coreference chains in contrast
- Fuzzy (co)reference

2. Constitution of an annotated corpus

- Constitution: database for corpus parts, selection criteria, metadata
- Finalization of the annotation manual for the annotation of chains
- Organisation of new experiments to evaluate the quality of the annotations, inter-annotators agreements, internal consistency
- Setting up internships to provide additional annotators
- Collective discussion on the last phases of the annotation: the case of non-coreferential anaphora, the case of text structure
- Design of the XML-TEI format to represent the corpus



Work in progress

3. Design of an annotation and query tool

- TXM: interface for the annotation of complex structures (“schemas”)
- TXM: interface for the annotation of relations
- Identification of measures to quantify the analyses of chains, and to adapt the corpus query possibilities to the project
- Collective discussion on how to merge TXM to other annotation tools
- Collaboration with the designers of GLOZZ

4. Design of a NLP system

- Ongoing developments based on the ANCOR corpus, which is available... pending the Democrat corpus
- Exploration in parallel of several machine learning techniques, with different concerns (hybrid systems, for instance)
- Evaluation of the use of syntactic data
- Development for the French language of the “end-to-end” neural coreference resolution approach from (Allen *et al.*, EMNLP 2017)

Tasks for the 2019-2020 year

- Continuation of research on the linguistic modelling
- Finalisation of the annotated corpus (March 2019)
- Finalisation of TXM-Democrat (March 2019)
- Design of NLP systems
 - Implementation of experiments on the Democrat corpus
 - Comparison of the experiments conducted on the ANCOR corpus with those conducted on the Democrat corpus
- Linguistic analysis of system errors
 - Questioning potential feedbacks from linguistics to NLP
 - Diagnosis of the systems according to the technologies
- Organization of new workshops, with new topics



The Democrat corpus and its annotation



Constitution of the corpus

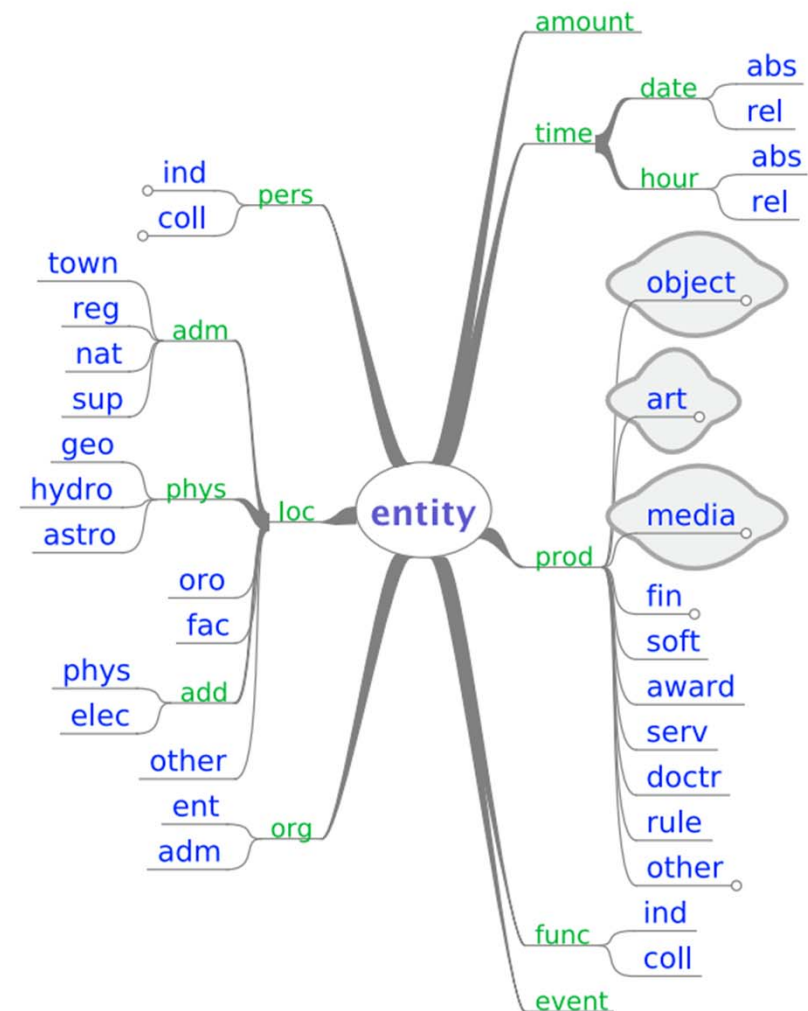
- 50% texts in contemporary French – 50% others
- 50% narrative texts – 50% others
 - Narrative texts: short stories, novels beginnings
 - Others: journalistic, scientific, and juridical texts
- Diachronic distribution as homogeneous as possible
- Some numbers
 - 52 texts currently identified (among 100 firstly planned)
 - Each text contains 10,000 words and is as coherent as possible (e.g. a complete chapter of a novel, or articles from one author)
 - Each text contains about 3,500 referring expressions

Types of referents: ESTER vs. QUAERO

- **Amount.** This category includes quantifiable data (age, duration, temperature, high, weight, width, distance, area, volume, speed, currencies).
- **Facility.** Facilities include buildings such as hospitals, factories, houses, museums, stadiums, ...
- **GPE.** Geo Political Entities refer to politically geographical regions. These entities don't distinguish between a geographical region, its people or its government.
- **Localisation.** This category includes geographical areas, circulation axes, postal and electronic addresses and telephone numbers.
- **Organization.** Expressions, names, acronyms that refer to an organisation that can be of political, religious, cultural nature are annotated as organisation entities.
- **Person.** Real persons as well as imaginary persons are considered in this category.
- **Product.** This category includes awards, vehicles, artistic work, and printed work.
- **Time.** Both date and time expressions are annotated as Time entities.

ESTER2

QUAERO





The Glozz URS metamodel

- Glozz, Analec and now TXM share the same metamodel for the representation of annotations: URS
 - U = units: they correspond to the markables
 - R = relations: they are (oriented) links between two markables
 - S = schemas: they are (heterogeneous) sets of units, relations, and schemas, which make it possible to model complex objects such as argumentative structures or... coreference chains
- Democrat's choices for the supports of annotations
 - The referring expressions are modelled using one "unit" type
 - Coreference chains are modelled using one "schema" type
 - Eventually, anaphoric relations may be modelled using one of several "relation" type(s)
 - Other objects (with their annotations) are possible, but will not belong to the public corpus

Materialization of Democrat's choices

A unit of the
“referring
expression”
type

A schema of the
“coreference
chain” type

● Comme tout avait brûlé la mère, les meubles et les photographies de la mère, pour Fabre et le fils Paul c'était tout de suite beaucoup d'ouvrage : toute cette cendre et ce deuil, déménager, courir se refaire dans les grandes surfaces. Fabre trouva trop vite quelque chose de moins vaste, deux pièces aux fonctions permutables sous une cheminée de brique dont l'ombre donnait l'heure, et qui avaient ceci de bien d'être assez proches du quai de Valmy.

Le soir après le dîner, Fabre parlait à Paul de sa mère, sa mère à lui Paul, parfois dès le dîner. Comme on ne possédait plus de représentation de Sylvie Fabre, ils s'épuisaient à vouloir la décrire toujours plus exactement : au milieu de la cuisine naquirent des hologrammes que dégonflait la moindre imprécision. Ça ne se rend pas, soupirait Fabre en posant une main sur sa tête, sur ses yeux, et le découragement l'endormait. Souvent ce fut à Paul de déplier le canapé convertible, transformant les choses en chambre à coucher.

Le dimanche et certains jeudis, ils partaient sur le quai de Valmy vers la rue Marseille, la rue Dieu, ils allaient voir Sylvie Fabre. Elle les regardait de haut, tendait vers eux le flacon de parfum Piver, Forvil, elle soupirait dans quinze mètres de robe bleue. Le gil d'un soupirait trouait sa blanche. Il n'y avait pas d'autre image d'elle.

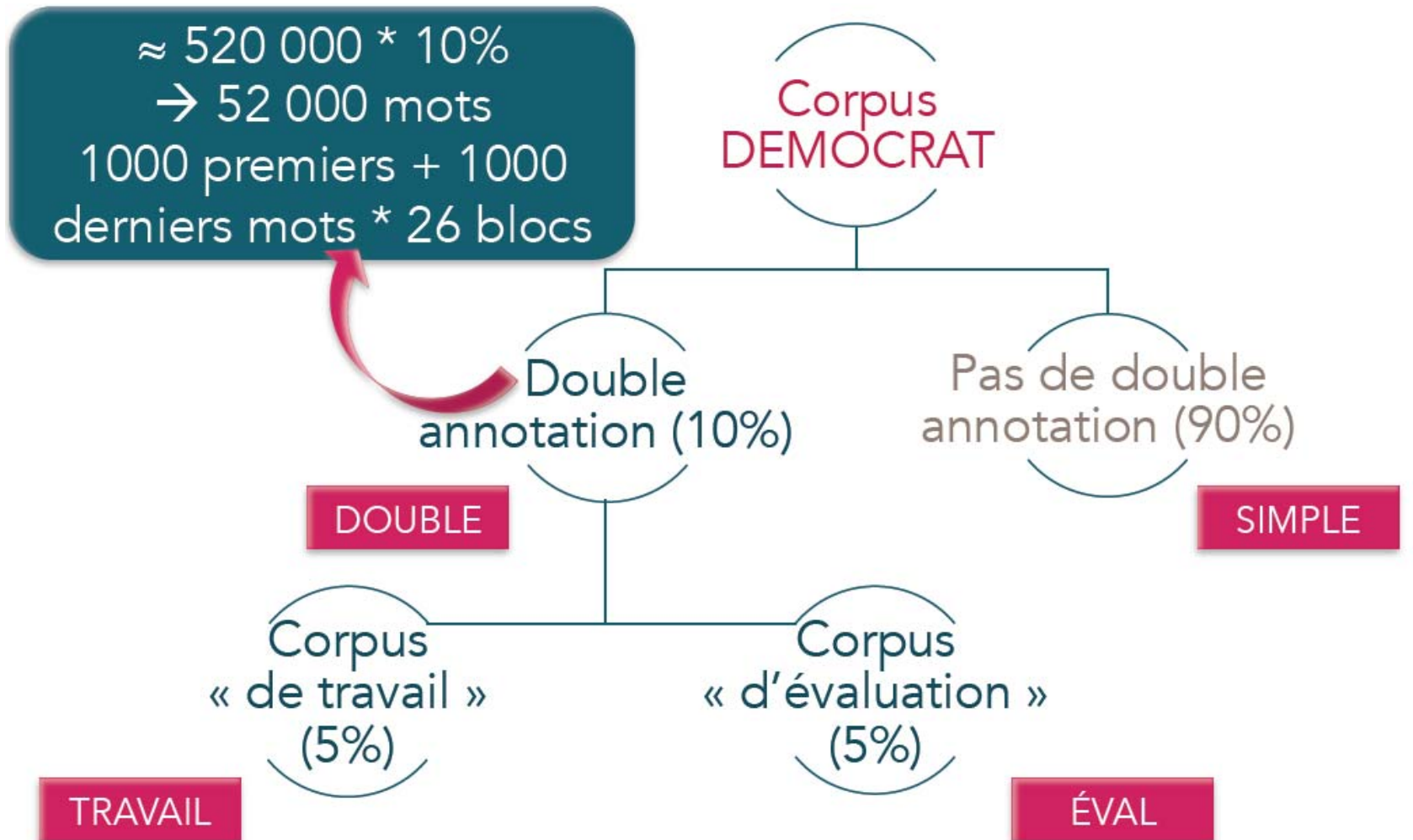
L'artiste Flers l'avait représentée sur le flanc d'un immeuble, juste avant le coin de la rue. L'immeuble était plus maigre et plus solide, mieux tenu que les vieilles constructions qui se collaient en grinçant contre lui, terrifiées par le plan d'occupation des sols. En manque de marquise, son porche saturé de moulures portait le nom (Wagner) de l'architecte-sculpteur gravé dans un cartouche en haut à droite. Et le mur sur lequel, avec toute son équipe, l'artiste Flers avait peiné pour figurer Sylvie Fabre en pied, surplombait un petit espace vert rudimentaire, sorte de square sans accessoires qui ne consistait qu'à former le coin de la rue.



Annotating referring expressions, annotating chains

- The most complex and most time-consuming task is the identification of referring expressions
 - All referring expressions! Not just the ones that refer to human beings
 - Hence a large number of “singletons” (e.g. spatial or temporal referents)
 - Many difficulties to delimit expressions : problems with relative subordinate clauses, with appositions, etc.
 - The annotation manual contains more than 30 pages that describe a number of cases and present a lot of examples of annotations
- The second important task is the assignment of a referent to each referring expression
 - Faced with an ambiguity, the annotator must choose...
 - There is no room for vagueness, nor for a “good-enough” approach...
 - This task leads to the automatic construction of the chains

Quality assessment and splitting of the corpus



The corpus annotation structure

The screenshot shows a software window titled "Structure des annotations" with three main panels: "Unités", "Relations", and "Schémas".

- Unités:** A tree view showing a hierarchy of folders and files. The root is "TYPES :". Under "TYPES :", there is a folder "MENTION". Under "MENTION", there is a folder "REF". Under "REF", there are several files: "SI", "duel générique", "Meung", "Paris", "armure de Porthos", "cheval de Porthos", "Aramis", "Porthos", and "Athos".
- Relations:** A panel titled "Relations" containing a folder "TYPES :".
- Schémas:** A panel titled "Schémas" containing a folder "TYPES :".

Phase 1 =
Manual
annotation
of referring
expressions :
- delimitation
- REF feature

The corpus annotation structure

The screenshot shows a software window titled "Structure des annotations" with three main panes: "Unités", "Relations", and "Schémas".

- Unités:** A tree view showing a hierarchy of annotation units. The root is "TYPES :". Under "TYPES :", there are three main categories: "MENTION", "DETERMINATION", and "REF".
 - MENTION:** Includes sub-units like "GENRE", "NOMBRE", "LONGUEUR", and "CATEGORIE". Under "CATEGORIE", there are sub-units: "NONE", "pronom clitique", "pronom relatif", "pronom", "zéro", "possessif", "groupe nominal", and "adv".
 - DETERMINATION:** Includes sub-units: "NONE", "ambigu", "démonstratif", "défini", and "indéfini".
 - REF:** Includes sub-units: "SI", "duel générique", "Meung", and "Paris".
- Relations:** Currently empty, with only "TYPES :" visible.
- Schémas:** Currently empty, with only "TYPES :" visible.

A text box is overlaid on the "Unités" pane, containing the following text:

Phase 2 =
Automatic annotation of referring expressions :
- morphosyntactic properties
- eventually structural properties

The corpus annotation structure

The screenshot shows a window titled "Structure des annotations" with three main panels:

- Unités:** A tree structure starting with "TYPES :". It contains three main categories: "MENTION", "DETERMINATION", and "REF".
 - MENTION:** Includes "GENRE", "NOMBRE", "LONGUEUR", and "CATEGORIE". "CATEGORIE" has sub-items: "NONE", "pronom clitique", "pronom relatif", "pronom", "zéro", "possessif", "groupe nominal", and "adv".
 - DETERMINATION:** Includes "NONE", "ambigu", "démonstratif", "défini", and "indéfini".
 - REF:** Includes "SI", "duel générique", "Meung", and "Paris".
- Relations:** A tree structure starting with "TYPES :". It is currently empty.
- Schémas:** A tree structure starting with "TYPES :". It contains one main category: "CHAINE".
 - CHAINE:** Includes "REF".
 - REF:** Includes "duel générique", "Paris", "armure de Porthos", "cheval de Porthos", "Aramis", "Porthos", and "Athos".

**Continuation
of phase 2 =
Automatic**
construction
of chains
thanks to the
REF values
(one chain per
REF value)

The corpus annotation structure

The screenshot shows a window titled "Structure des annotations" with three main panes: "Unités", "Relations", and "Schémas".

- Unités:** A tree structure under "TYPES :".
 - MENTION
 - GENRE
 - NOMBRE
 - LONGUEUR
 - CATEGORIE
 - NONE
 - pronom clitique
 - pronom relatif
 - pronom
 - zéro
 - possessif
 - groupe nominal
 - adv
 - DETERMINATION
 - NONE
 - ambigu
 - démonstratif
 - défini
 - indéfini

- Relations:** A tree structure under "TYPES :".
- Schémas:** A tree structure under "TYPES :".
- CHAINE
 - REF
 - duel générique
 - Paris
 - armure de Porthos
 - cheval de Porthos
 - Aramis
 - Porthos
 - Athos

End of phase 2 = Automatic deletion of the REF feature here

The corpus annotation structure

The screenshot displays the 'Structure des annotations' window, which is divided into three main panels: 'Unités', 'Relations', and 'Schémas'. Each panel shows a hierarchical tree structure of annotation types and their sub-properties.

- Unités:** Contains a 'TYPES' folder with two main categories: 'MENTION' and 'DETERMINATION'. 'MENTION' includes sub-properties like 'GENRE', 'NOMBRE', 'LONGUEUR', and 'CATEGORIE' (with sub-items: NONE, pronom clitique, pronom relatif, pronom, zéro, possessif, groupe nominal, adv). 'DETERMINATION' includes sub-properties: NONE, ambigu, démonstratif, défini, indéfini.
- Relations:** Contains a 'TYPES' folder, which is currently empty.
- Schémas:** Contains a 'TYPES' folder with three main categories: 'CHAINE', 'CARDINAL', 'SEXE', and 'REF'. 'CHAINE' includes 'TYPE DE REFERENT' (with sub-items: NONE, humain, animal, objet concret, objet abstrait, date, lieu, organisation, produit) and 'CARDINAL' (with sub-items: groupe strict, groupe flou, singulier). 'SEXE' includes sub-items: confondu, féminin, masculin, indéterminable. 'REF' includes sub-items: duel générique, Paris.

A central text box with a black border and white background contains the following text:

Phase 3 =
Manual
annotation of
chains with
properties of
the referents

The corpus annotation structure

Phase 4 = Automatic
annotation
of additional
properties
of referring
expressions
(for compatible
annotations with
ANCOR corpus,
for instance)

Unités

- TYPES :
 - MENTION
 - SINGLETON
 - oui
 - non
 - NEW
 - oui
 - non
 - GENRE
 - NOMBRE
 - LONGUEUR
 - CATEGORIE
 - NONE
 - pronom clitique
 - pronom relatif
 - pronom
 - zéro
 - possessif
 - groupe nominal
 - adv
 - DETERMINATION
 - NONE
 - ambigu
 - démonstratif
 - défini

Schémas

- TYPES :
 - CHAINE
 - TYPE DE REFERENT
 - NONE
 - humain
 - animal
 - objet concret
 - objet abstrait
 - date
 - lieu
 - organisation
 - produit
 - CARDINAL
 - groupe strict
 - groupe flou
 - singulier
 - SEXE
 - confondu
 - féminin
 - masculin
 - indéterminable
 - REF
 - duel générique
 - Paris

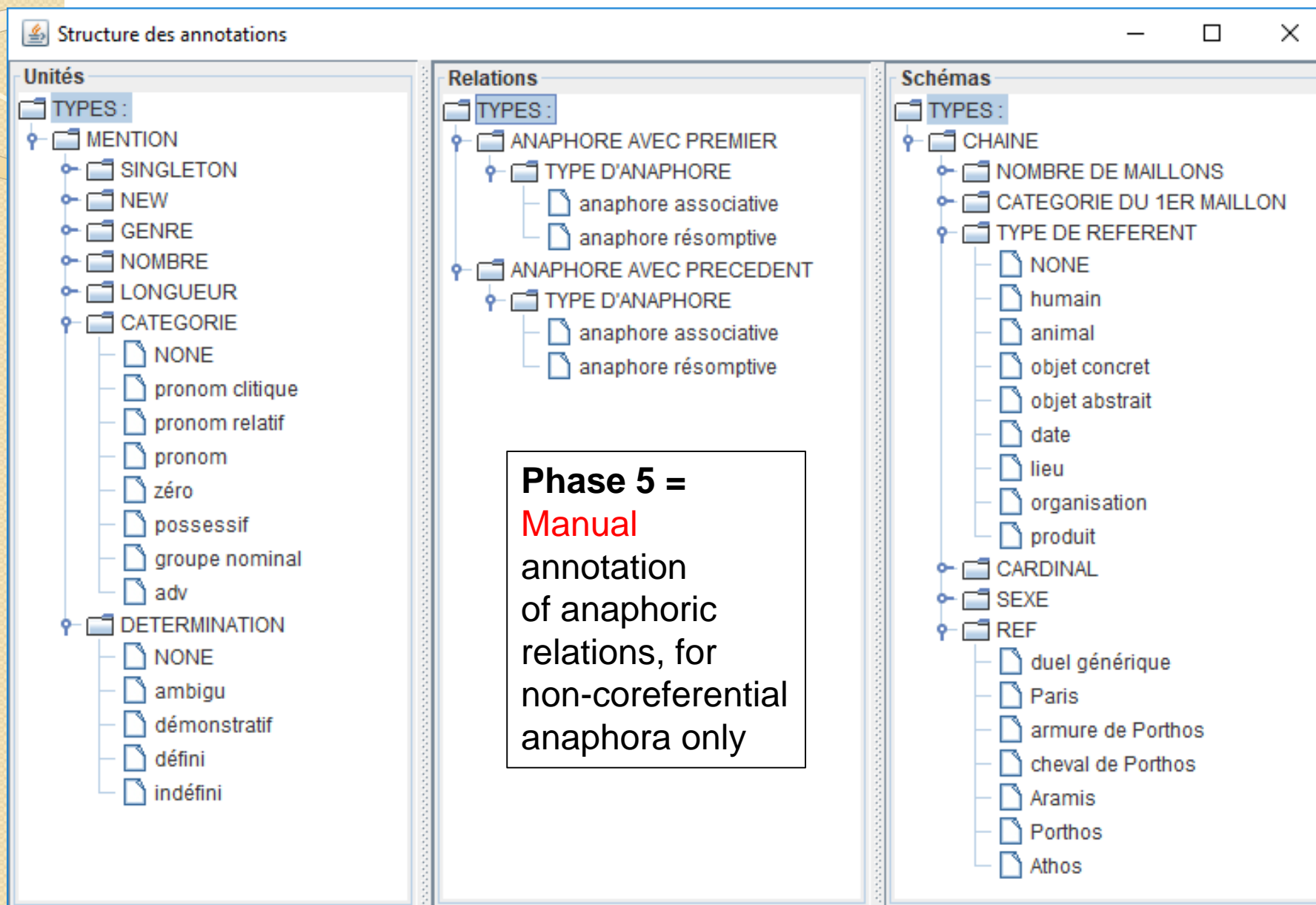
The corpus annotation structure

The screenshot displays the 'Structure des annotations' window, which is divided into three main sections: 'Unités', 'Relations', and 'Schémas'. Each section contains a hierarchical tree structure of folders and files representing the annotation schema.

- Unités:** Contains a 'TYPES' folder with sub-folders: MENTION (SINGLETON: oui, non; NEW: oui, non), GENRE, NOMBRE, LONGUEUR, CATEGORIE (NONE, pronom clitique, pronom relatif, pronom, zéro, possessif, groupe nominal, adv), and DETERMINATION (NONE, ambigu, démonstratif, défini).
- Relations:** Contains a 'TYPES' folder, which is highlighted by a callout box.
- Schémas:** Contains a 'TYPES' folder with sub-folders: CHAINE (NOMBRE DE MAILLONS, CATEGORIE DU 1ER MAILLON, TYPE DE REFERENT: NONE, humain, animal, objet concret, objet abstrait, date, lieu, organisation, produit), CARDINAL (groupe strict, groupe flou, singulier), SEXE (confondu, féminin, masculin, indéterminable), and REF.

Phase 4 = Automatic annotation of additional properties of chains

The corpus annotation structure



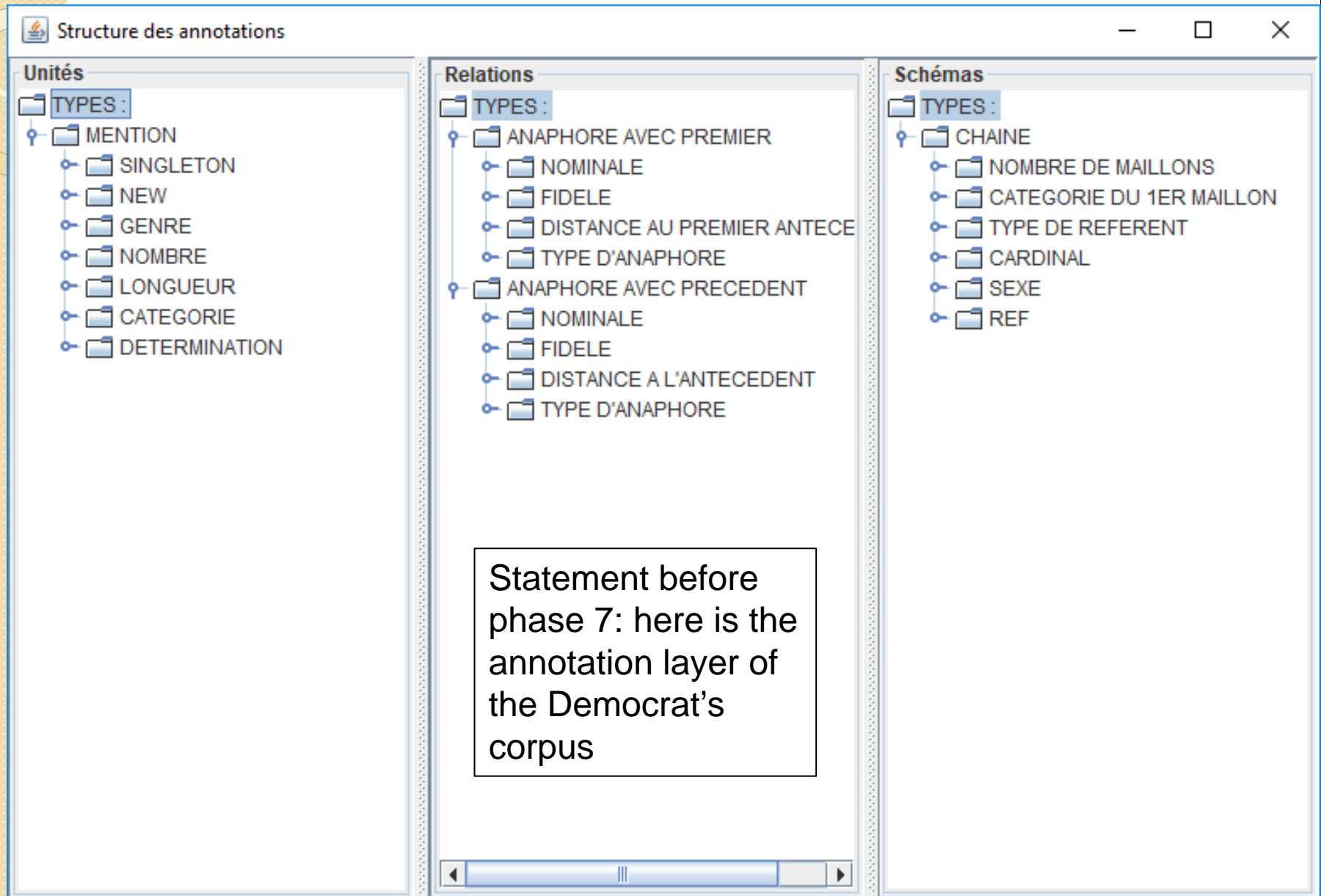
The corpus annotation structure

The screenshot displays the 'Structure des annotations' window, which is organized into three main panels:

- Unités:** A tree structure starting with 'TYPES :'. It includes folders for 'MENTION', 'SINGLETON', 'NEW', 'GENRE', 'NOMBRE', 'LONGUEUR', and 'CATEGORIE'. Under 'CATEGORIE', there are items like 'NONE', 'pronom clitique', 'pronom relatif', 'pronom', 'zéro', 'pos', 'grou', 'adv', and 'DETER'. Below this is a folder for 'DETER' with items like 'NON', 'am', 'dém', 'défi', and 'indé'.
- Relations:** A tree structure starting with 'TYPES :'. It includes folders for 'ANAPHORE AVEC PREMIER', 'NOMINALE', 'FIDELE', 'DISTANCE AU PREMIER ANTECE', 'TYPE D'ANAPHORE', 'ANAPHORE AVEC PRECEDENT', 'NOMINALE', 'FIDELE', 'DISTANCE A L'ANTECEDENT', and 'TYPE D'ANAPHORE'. Under 'TYPE D'ANAPHORE' and the second 'TYPE D'ANAPHORE', there are items like 'anaphore coréférente', 'anaphore associative', and 'anaphore résomptive'.
- Schémas:** A tree structure starting with 'TYPES :'. It includes folders for 'CHAINE', 'NOMBRE DE MAILLONS', 'CATEGORIE DU 1ER MAILLON', 'TYPE DE REFERENT', 'CARDINAL', 'SEXE', and 'REF'. Under 'TYPE DE REFERENT', there are items like 'NONE', 'humain', 'animal', 'objet concret', 'objet abstrait', 'date', 'lieu', 'organisation', and 'produit'. Under 'REF', there are items like 'duel générique', 'Paris', 'armure de Porthos', 'cheval de Porthos', 'Aramis', 'Porthos', and 'Athos'.

Phase 6 = Automatic annotation of coreferential anaphora, and of the properties of all anaphora

The corpus annotation structure



Statement before phase 7: here is the annotation layer of the Democrat's corpus

The corpus annotation structure

Structure des annotations

Unités

- TYPES :
- PARAGRAPHE
 - ALINEA VERTICAL
 - oui
 - non
 - ALINEA HORIZONTAL
 - oui
 - non
 - MENTION
 - SINGLETON
 - NEW
 - GENRE
 - NOMBRE
 - LONGUEUR
 - CATEGORIE
 - DETERMINATION

Relations

- TYPES :
- ANAPHORE AVEC PREMIER
 - NOMINALE
 - FIDELE
 - DISTANCE AU PREMIER ANTECE
 - TYPE D'ANAPHORE
- ANAPHORE AVEC PRECEDENT
 - MINALE
 - LE
 - DISTANCE A L'ANTECEDENT
 - PE D'ANAPHORE

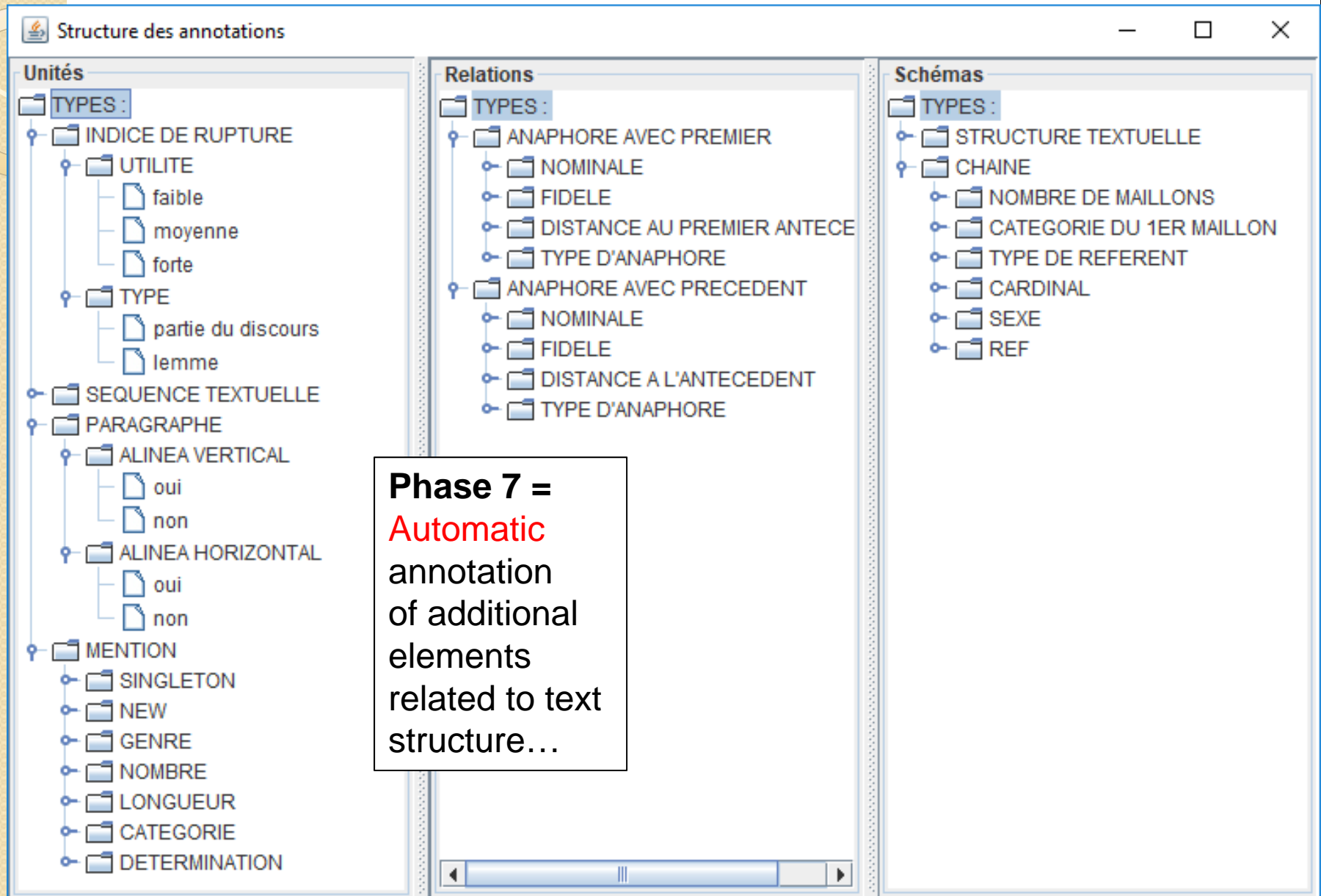
Schémas

- TYPES :
- STRUCTURE TEXTUELLE
- CHAINE

Phase 7 = Automatic annotation of paragraphs

Phase 7 = Automatic annotation of a certain type of text structure

The corpus annotation structure



Phase 7 =
Automatic
annotation
of additional
elements
related to text
structure...




Annotation phases: assessment

- All this procedure has one purpose: minimizing manual annotation and encouraging automatic annotation as soon as it can be considered
- Rational alternation of manual and automatic phases, with the launching of a lot of scripts – rational, but not very easy to understand at first glance
- For the moment, only phase 1 is mandatory
- For the public corpus, we will stop at phase 4 (but not before, otherwise no comparison with ANCOR nor NLP application is possible)



By the way, what are annotations for?

- To constitute a reference corpus on reference and coreference
- To provide linguists with a rich and diversified “pool” of examples
- To provide data for statistical or even textometric computations on coreference chains
- To provide data for the learning phase of NLP systems that are dedicated to the automatic detection of referring expressions and/or coreference chains



**Natural language processing:
automatic detection of
coreference chains**



State of the art: rule-based systems

- Rule-based systems
 - Principle: a set of rules is written by hand:
 - **If definite article then...**
 - **If distance between two expressions is less than 8 words then...**
 - Advantage: the rules are readable (understandable) and are the result of a collaboration between linguists and computer scientists
 - Disadvantages:
 - Lack of flexibility: any rule correction can have collateral effects and degrade overall performance
 - Lack of performance, especially for complex tasks involving many parameters
- Note in passing (about phase 2 of the annotation procedure)
 - It is a rule-based system that is used to automatically annotate the determination of referring expressions, as well as other properties

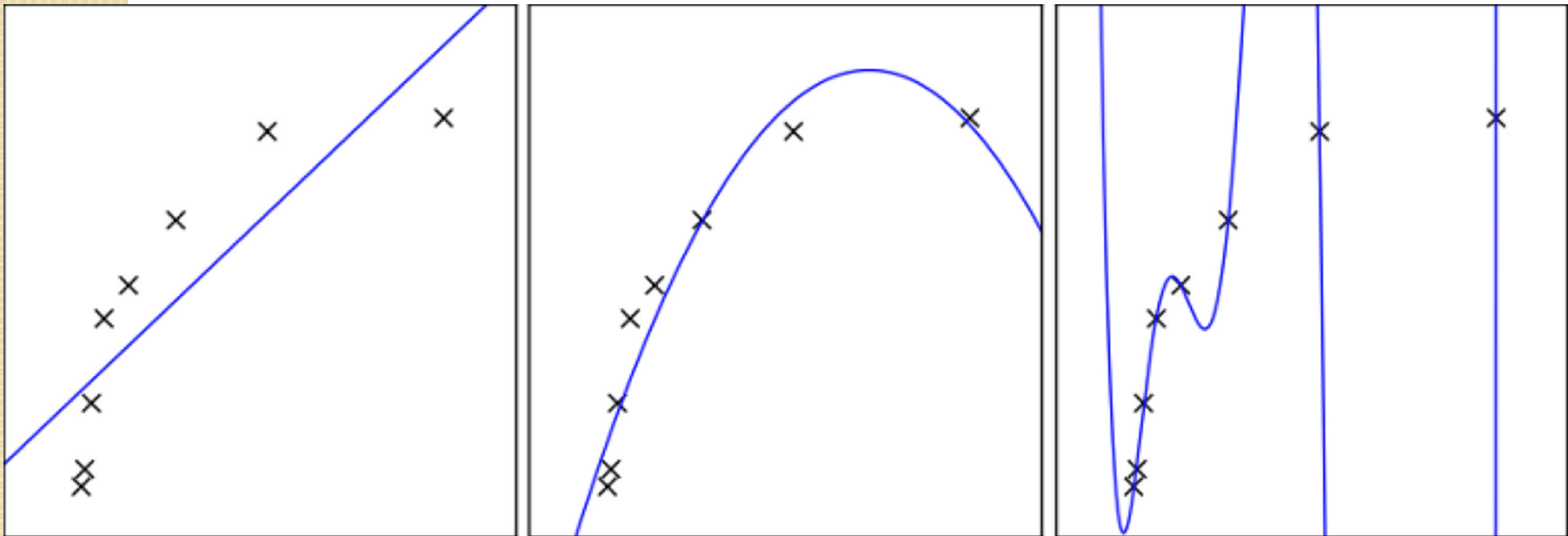


State of the art: machine learning

- We entrust a system:
 - The determination of its own rules
 - The determination of its own thresholds (e.g. distance between 2 expressions)
 - Advantages: great flexibility, little intervention of intuition
 - Disadvantages:
 - The solutions found by the system are sometimes difficult to read and cannot be modified a posteriori: they have to be accepted...
 - Hybrid approaches (rules + machine learning) are difficult to implement: it is often better to restart a new learning phase...
 - Above all: the system learns from a basis, that is an annotated corpus (it is impossible to learn without annotated examples)
- On coreference chains for the French language
 - ANCOR corpus available → CROC system CROC + ongoing works
 - DEMOCRAT corpus → new systems


Machine learning principles

- Transforming separate examples (x, y) into a rule: function $f(x) = y$
 - This requires generalization (learning by heart is useless), but not too strong...
 - This allows to predict the value of $f(x)$ for a new x
 - Of course, there are traps...



2 phases: learning and application

- Learning phase: annotated corpus → **model**
 - We take an annotated corpus
 - We split it into several parts: one dedicated to the machine learning, the others for testing and validating
 - From the annotations, examples with their characteristics are extracted
 - The **machine learning system** learns from these examples...
 - ...and determines a learning model, that is the function $f(x) = y$
 - To avoid traps, the model is forced not to be “too close” to the data, in order to encourage generalization (regularization technique)
- Application phase: raw text ^{model} → annotated text
 - We take a raw text
 - The learning model is applied to it
 - We directly obtain an annotated text
 - It is the **end-to-end system**



Machine learning applied to a specific task

- There are a multitude of machine learning algorithms
 - Various performances depending on the nature of x
 - Various performances depending on the nature of y
 - Various performances depending on the types of function f (choice of search space)
 - Various performances depending on the main evaluation criterion:
 - best ability to predict
 - best interpretability of results
 - robustness when tested in a new domain
 - shortest computation time...
- **No Free Lunch!**
 - There is no algorithm that does better than all others on all problems



Machine learning applied to coreference chains

- Nature of x

- An (annotated) example is a referring expression (delimited and annotated)
- It is a group of consecutive words (and not a single word as for morphosyntactic analysis, i.e. POS tagging)
- We know that gender and number help to identify coreferences, so we add the gender and the number to the characteristics of the examples
- And so on...

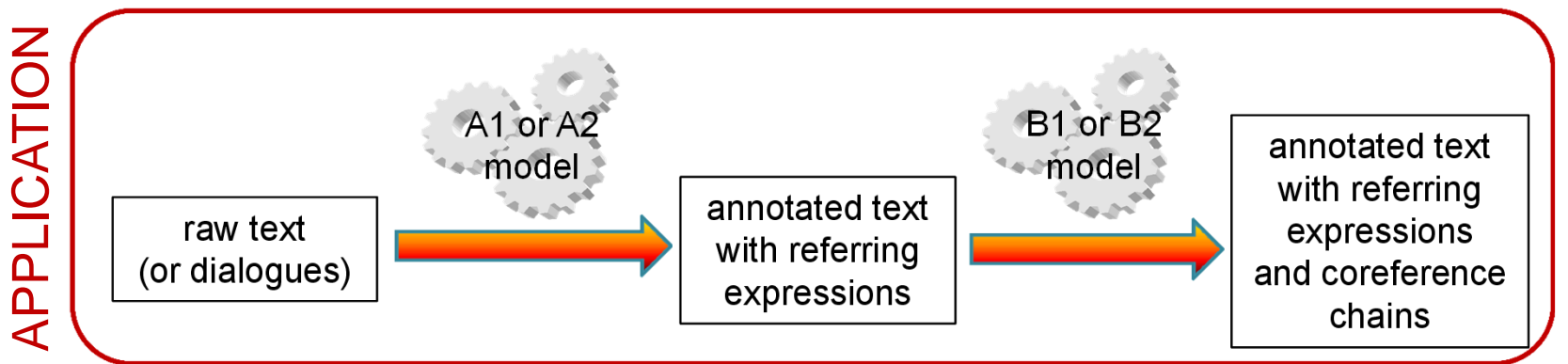
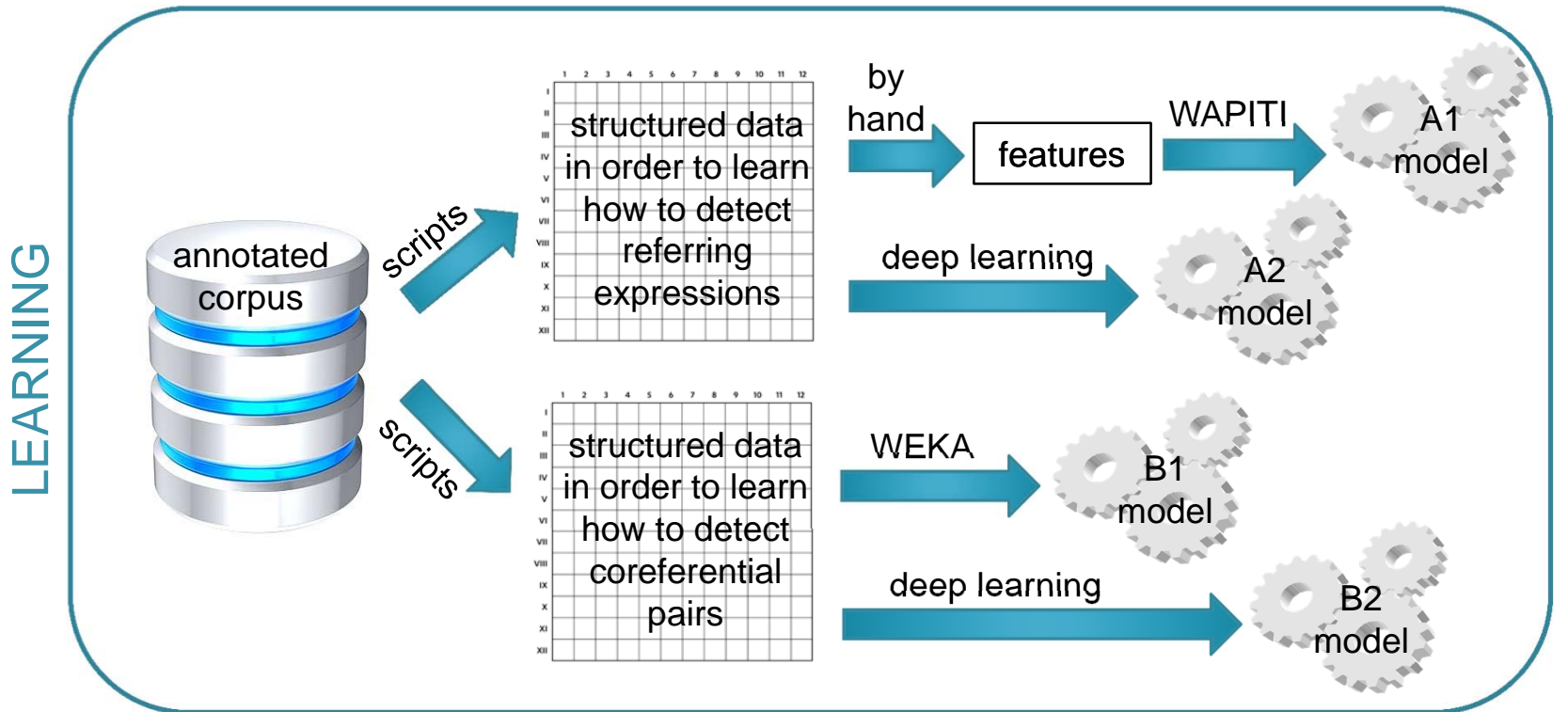
- Nature of y

- Detecting referring expressions is not the same task as deciding whether two referring expressions are coreferential or not
- So we distinguish two phases of machine learning

Two distinct problems

- A first phase: detecting referring expressions
 - x = sequence ; y = BIO format on the referring expressions
 - A task that is close to but not identical = the problem of detecting the named entities in a text, a very famous task in NLP community
 - A task that is close to but not identical = detection of proper names, nominal phrases and pronouns (= nominal chunks)
- A second phase: detecting coreferences
 - x = pair of referring expressions ; y = yes or no
 - First prototype for the French language made par Adèle Désoyer, using methods like SVM (support vector machine)
- Note in passing
 - *Deep learning* brings some additional elements to this presentation...

NLP processes





“Feeding” the machine learning system

- General principle
 - We identify parameters that could help machine learning
 - We compute features
 - We provide a (potentially very huge) file to the learning system
- Everything is done using features
 - We can imagine as many as we want, but we still have to be able to compute them, because of the *end-to-end* system...
 - As a future work, we could consider a hybrid system with both machine learning and rules, that can be applied before and/or after the learning phase (but doing both at the same time is complex)



Future works

Future works

- From the corpus and the analysis of its annotations to the design of a linguistic model
- Democrat has three variations:
 - Text genre – a variation that is materialized in the corpus
 - Time: diachronic approach – materialized in the corpus
 - Language: contrastive approach – not materialized in the corpus
- Other variations are possible
 - Productions of pathological subjects
 - Writing vs. speaking, and also new forms of communication (SMS...)
- In the longer term, Democrat's work could be a first step towards research on the cognitive aspects of reference, including the notion of salience

Bibliography

- Frédéric Landragin, Marine Delaborde, Yoann Dupont, Loïc Grobol (2018) « Description et modélisation des chaînes de référence. Le projet ANR Democrat (2016-2020) et ses avancées à mi-parcours », *TALN 2018, salon de l'innovation en TAL et RI*, Orléans
- Loïc Grobol, Frédéric Landragin, Serge Heiden (2018) “XML-TEI-URS: using a TEI format for annotated linguistic resources”, *CLARIN Annual Conference*, Pisa, Italy
- Matthieu Quignard, Serge Heiden, Frédéric Landragin, Matthieu Decorde (2018) “Textometric Exploitation of Coreference-annotated Corpora with TXM”, *JADT 2018*, Rome, Italy
- Bruno Oberlé (2018) “SACR: A Drag-and-Drop Based Tool for Coreference Annotation”, *LREC 2018*, Miyazaki, Japan
- Frédéric Landragin (2018) « Étude de la référence et de la coréférence : rôle des petits corpus et observations à partir du corpus MC4 », *Corpus*, numéro 18
- Céline Poudat et Frédéric Landragin (2017) *Explorer un corpus textuel*, De Boeck Supérieur
- Frédéric Landragin (2017) « Analyse, visualisation et identification automatique des chaînes de coréférences : des questions interdépendantes ? », *Langue Française*, numéro 195
- Catherine Schnedecker, Julie Glikman, Frédéric Landragin (2017) « Les chaînes de référence : annotation, application et questions théoriques », *Langue Française*, numéro 195
- Frédéric Landragin, Juliette Potier, Meryl Bothua (2017) « Annotation manuelle d'expressions référentielles : expérimentations pour simplifier les prises de décisions », *JLC 2017*, Grenoble
- Frédéric Landragin (2016) « Conception d'un outil de visualisation et d'exploration de chaînes de coréférences », *JADT 2016*, Nice
- Frédéric Landragin, Noalig Tanguy, Michel Charolles (2015) « Référence aux personnages dans *L'Occupation des sols* : apport de la linguistique outillée », *Revue Sciences/Lettres*, numéro 3
- Frédéric Landragin (2011) « Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits », *Corpus*, numéro 10