HAL open science

“ Reference Chains across Languages : a Cross-Linguistic study ”.

Emmanuel Baumer

▶ To cite this version:

Emmanuel Baumer. “ Reference Chains across Languages : a Cross-Linguistic study ”.. Premier Colloque International de linguistique contrastive, Université des Langues et Cultures de Beijing, Sep 2019, Beijing, China. hal-02486510

HAL Id: hal-02486510
https://hal.science/hal-02486510

Submitted on 21 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Reference Chains across Languages : a Cross-Linguistic study

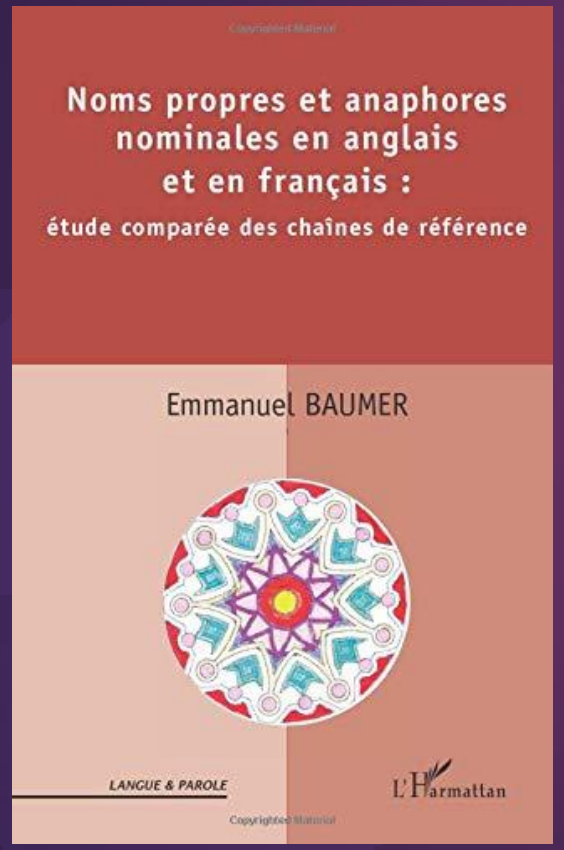
EMMANUEL BAUMER

(UNIVERSITÉ CÔTE D'AZUR / BCL CNRS / ANR DEMOCRAT)

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR

Introduction

Introduction



C. Schnedecher (2019) : "The virtues of the notion of referential chain when compared to the notions of anaphora and coreference".

+ **Baumer (2015) :** "Proper nouns and Nominal Anaphors : a Comparative Study of Reference Chains".

Introduction

- 1/ Theoretical description of Reference Chains
- 2/ Presentation of ANR “Democrat” Project and Methodology of Reference Chain studies
- 3/ A few aspects of some recent contrastive studies of Reference Chains

1/ Theoretical Description of RC

Theoretical Description of RC

- Concept of « **Reference Chain** » (RC) = quite recent, comes from philosophy of language.

- Has aroused **a lot of interest in linguistics since the 1980s** + inspired a large number of studies in different areas of research (**discourse reference**, **textual cohesion**, or **automatic language processing**).

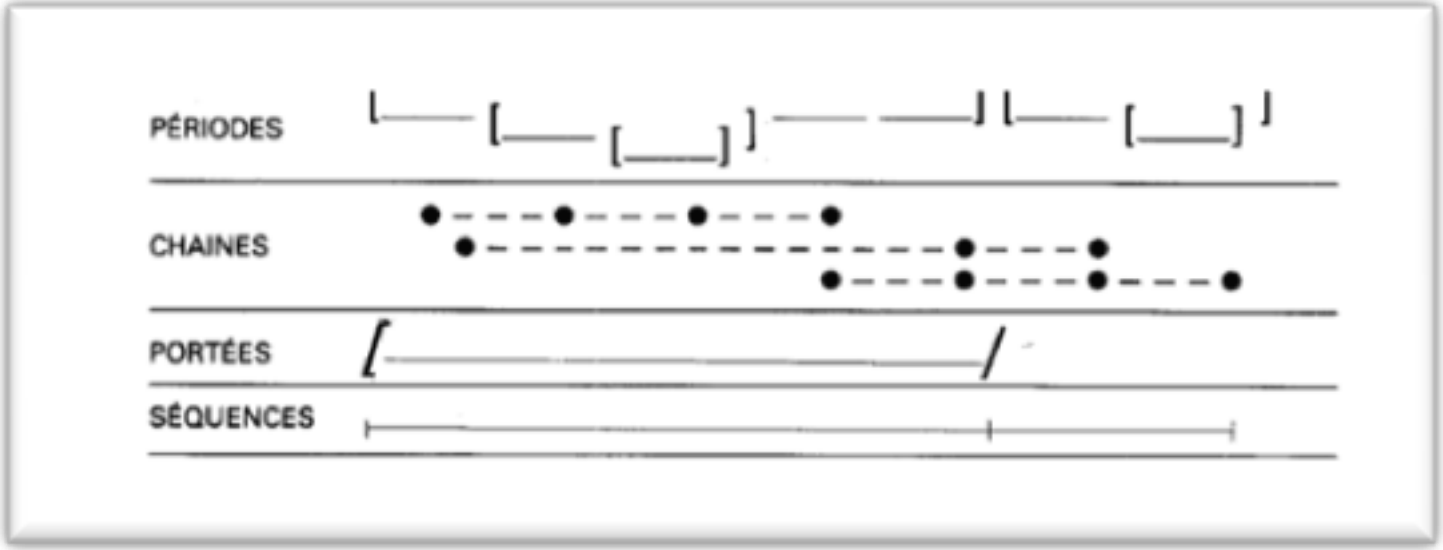
Theoretical Description of RC

- « **RC** » = coined by **C. Chastain in 1975**, + reappeared in various studies in the 80s, i.e : Corblin (1985), Charolles (1988),…

- **Definition** = the **sequence of referential expressions** referring to the same entity (person / object) **in a given text**. Those expressions are called **links** or **mentions**.

Theoretical Description of RC

- Ex : Charolles (1988) : RC = one of the **4 important units of textual structure** (i.e : chains, periods, ranges, sequences)



Theoretical Description of RC

- **For instance :** **Blue RC = Ling**, **Red RC = Wang-Fô**

Une nuit, dans une taverne, [**Ling**] eut **Wang-Fô** pour compagnon de table. **Le vieil homme** avait bu pour se mettre en état de mieux peindre un ivrogne[...]. **Wang-Fô** se pencha pour faire admirer à **Ling** la zébrure livide de l'éclair, et **Ling**, émerveillé, cessa d'avoir peur de l'orage. **Ling** paya l'écot **du vieux peintre** : comme **Wang-Fô** était sans argent et sans hôte, **il lui** offrit humblement un gîte. Ils firent route ensemble ; **Ling** tenait une lanterne ; sa lueur projetait dans les flaques des feux inattendus. Ce soir là, **Ling** apprit avec surprise que les murs de **sa** maison n'étaient pas rouges, comme **il** l'avait cru, mais qu'ils avaient la couleur d'une orange prête à pourrir. (M. Yourcenar)

Theoretical Description of RC

Green RC = Mr. Martin

- **Mr. Martin** bought the pack of Camels on Monday night in the most crowded cigar store on Broadway. It was theatre time and seven or eight men were buying cigarettes. The clerk didn't even glance at **Mr. Martin**, **who** put the pack in **his** overcoat pocket and ø went out. If any of the staff at F & S had seen **him** buy the cigarettes, they would have been astonished, for it was generally known that **Mr. Martin** did not smoke, and ø never had. No one saw **him**. It was just a week to the day since **Mr. Martin** had decided to rub out Mrs. Ulgine Barrows. The term "rub out" pleased **him** because it suggested nothing more than the correction of an error – in this case an error of Mr. Fitweiler. **Mr. Martin** had spent each night of the past week working out **his** plan and examining it. (J. Thurber)

Theoretical Description of RC

Red RC = Xiaolan, **Blue RC = Xiaopeng**

Bànyè liǎng diǎn le, **Xiǎolán** háishi xǐng zhe de. [It was 2 o'clock in the morning and Xiao Lan was still awake].

Fángjiān lǐ hěn hēi, hěn ānjìng, ràng **tā** gǎndào fēicháng kùn. [The room was dark and quiet, making her sleepy].

Kěshì **tā** bù gǎn shuì, yīnwèi **tā** hàipà rúguǒ **tā** shuìzháo dehuà, **tā** yòu huì zuò nàge kěpà de mèng. [but she didn't dare to go to sleep, because she was afraid that if she slept, she would have the same horrible nightmare].

Kě zuìzhōng **tā** háishi shuìzháo, bìng ∅ zuò qǐ mèng lái le... [But eventually, she fell asleep and ∅ started to dream...]

Mèngzhōng, **Xiǎolán** de nánpéngyou **Xiǎopéng**, qí zhe mótuōchē láidào **tā** jiā. [In the dream, Lan's boyfriend Peng had come to Lan's home with a motorcycle].

Tā shuō ∅ yào dài **Xiǎolán** qù yī gè dìfang. [He said that he was going to take Lan to a place.] (mandarincorner.com)

Theoretical Description of RC

Material of RC = belongs to **3 main categories of referential expressions** :

- **Proper nouns** = Mr Martin, Wang-Fo, Ling, Xiao Lan, Xiao Peng, …
- **Referential Pronouns** = il, lui, he, him, his, who, ta（她）, ta（他）, ∅, …
- **Noun Phrases** (NP) = le vieil homme, le vieux peintre, …

The distribution of these categories = can vary according to **genres** and **languages**.

Theoretical Description of RC

Difference between RC and anaphora (and coreference) = C. Schnedecker (2019):

- **Notion of RC =** « Far from being superimposed on that of **anaphora** or of **coreference**, is endowed with a **consistency that makes it truly specific in theoretical and empirical terms** » :
- « It is a well-circumscribed object, which can be grasped with **measures** capable of conferring on it even greater strength, **specific theoretical and empirical questions**, and a **wide field of application**, ranging from **referential semantics** to **textual linguistics**. »

Theoretical Description of RC

Difference between RC and anaphora =

First of all : **anaphora, unlike RC = asymmetrical by nature**

- An anaphoric relationship implies the linking of an expression that is **not autonomous**, as far as reference is concerned, and a referential expression that is likely to "**saturate**" it (Corblin 1995).
- There is an anaphoric relationship between two units A and B when **the interpretation of B crucially depends on the existence of A**. In fact, B « is only interpretable to the extent that it includes all or part of A » (Milner 1982).

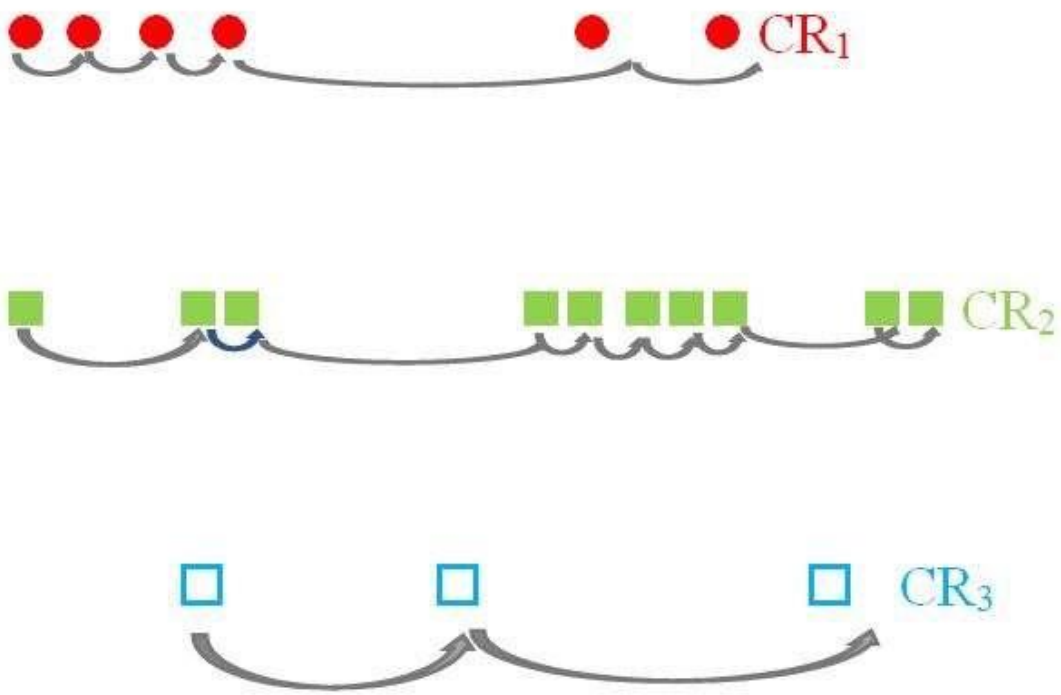
Theoretical Description of RC

- Therefore, RC = different from **anaphora**, because of the minimum number of referential expressions involved in the relationship, which, for RC, implies **more than 2 mentions** or **links**.

- RC is used to understand **the process or follow-up of the referential expression in its textual or discursive form**.

Theoretical Description of RC

Several RC can be **tracked simultaneously** in the linear order of a text :



2/ ANR « Democrat » Project and Methodology

Democrat Project and Methodology

- **ANR « DEMOCRAT » Project** = Describing and Modelling Reference Chains : Tools for Corpus Annotation (including diachronic and **comparative language studies**) and Automatic Processing.
- 5 year project (2015-2020), founded by the **French National Agency for Research**.

Democrat Project and Methodology

« Democrat » Involves 3 French partner labs :

- **LILPA** : University of Strasbourg
- **LATTICE** : Paris Sorbonne, CNRS, Ecole Normale Supérieure Paris
- **IHRIM / ICAR** : Ecole Normale Supérieure Lyon

- Project coordinator = Frédéric Landragin, Lattice

Democrat Project and Methodology

Democrat = 4 steps to develop innovative studies of referring expressions and RC :

- (1) A discursive, diachronic and cross-linguistic **model** of RC
- (2) A large manually annotated **corpus** (now available online / Ortolang site: https://www.ortolang.fr/market/corpora/democrat)
- (3) Different **tools** to annotate and explore the annotations (TXM, SACR, CR Viewer)
- (4) A automatic language processing **system** to pave the way for automated annotation of RC (based on deep learning algorithms)

ortolang.fr

Tenue d'un colloque sur l'éducation de la langu... | Downloads • Jazz Advice | DeepL Traducteur | DEscription et MOdélisation des Chaînes de... | Democrat | ORTOLANG

ORTOLANG — Informations — Langue — Connexion

- Accueil
- Corpus
- Lexiques
- Terminologies
- Outils
- Projets Intégrés
- Actualités
- Informations
- Producteurs

ANR Democrat

Democrat

Laboratoires producteurs :

Lattice - Langues, textes, traitements informatiques, cognition - UMR 8094 (Paris, FR)
LiLPa - Linguistique, Langues, Parole - EA 1339 (Strasbourg, FR)
ICAR - Interactions, corpus, apprentissages et représentations - UMR 5191 (Lyon, FR)
IHRIM - Institut d'histoire des représentations et des idées dans les modernités - UMR 5317 (Lyon, FR)

Soutien institutionnel :

ANR - Agence Nationale de la Recherche (Paris, FR)

Référence à citer

Texte | BibTeX

Frédéric Landragin. Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). Bulletin de l'Association Française pour l'Intelligence Artificielle, AFIA, 2016, pp.11-15. ⟨afia.asso.fr⟩. ⟨hal-01347949⟩

Langues, textes, traitements informatiques, cognition - UMR 8094 (Lattice), Linguistique, Langues, Parole - EA 1339 (LiLPa), Interactions, corpus, apprentissages et représentations - UMR 5191 (ICAR), Institut d'histoire des représentations et des idées dans les modernités - UMR 5317 (IHRIM) (2019). *Democrat* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - **www.ortolang.fr**, **https://hdl.handle.net/11403/democrat/v1**.

Description

Le corpus Democrat est un corpus textuel annoté en références. Les expressions référentielles sont repérées et annotées avec l'identifiant du référent, ce qui permet de construire les chaînes de référence.
La composition du corpus est établie dans l'objectif d'étudier la variation des chaînes de référence en fonction des genres discursifs et des époques. La composition est établie selon trois critères : époque, type de texte (narratif ou non-narratif), genre textuel. La taille du corpus a été envisagée pour permettre des applications de traitement automatique des langues.

Ce corpus est l'un des livrables du projet ANR Democrat (ANR-15-CE38-0008).

Contacter le producteur

Envoyer un mail

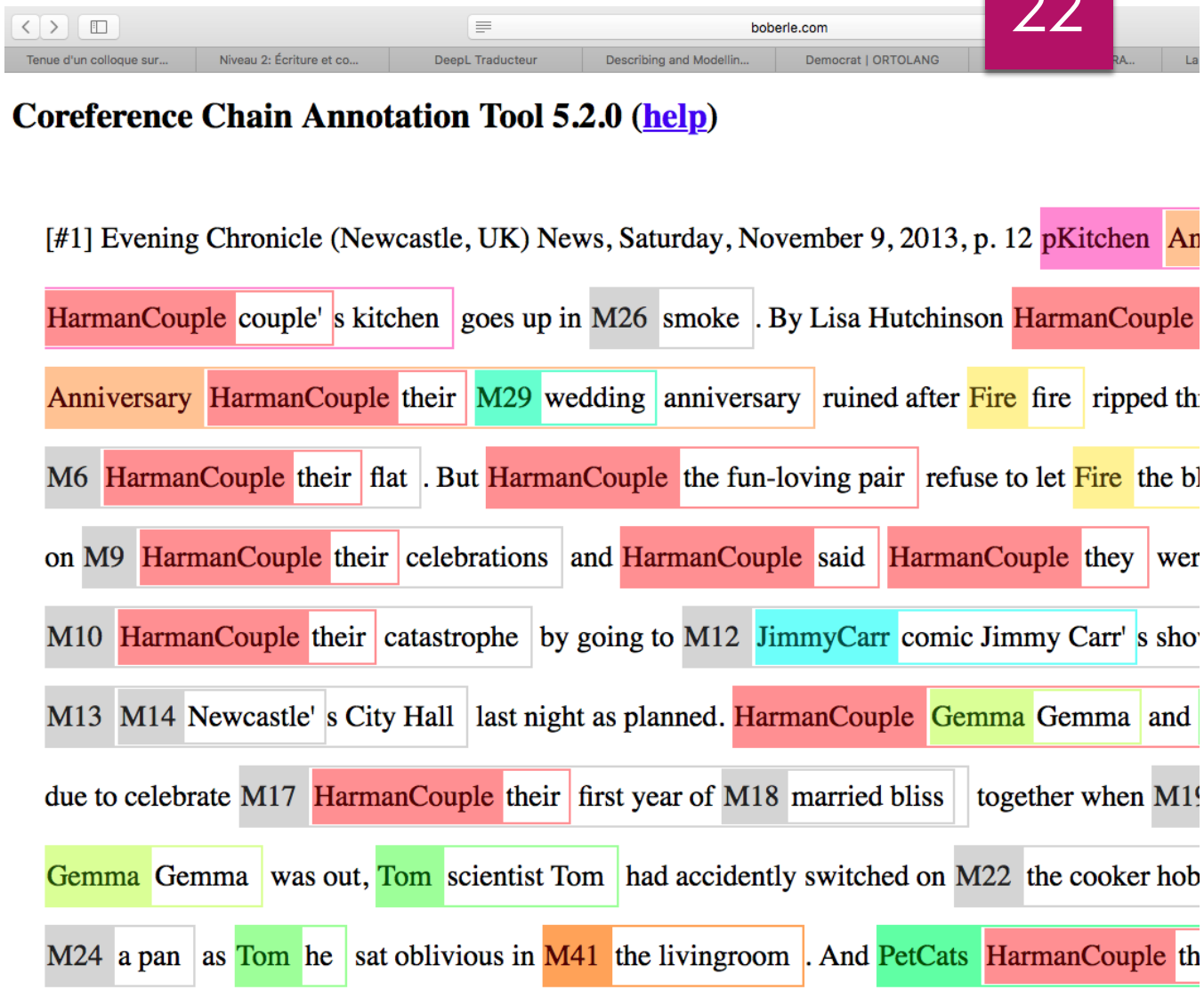
Téléchargement

Licence Creative Commons - Attribution - Partage dans les Mêmes Conditions 4.0 International
Cette licence permet aux autres de remixer, arranger, et adapter votre œuvre, même à des fins commerciales, tant quu2019on vous accorde le mérite en citant votre nom et quu2019on diffuse les nouvelles créations selon des conditions identiques.

Télécharger | Parcourir

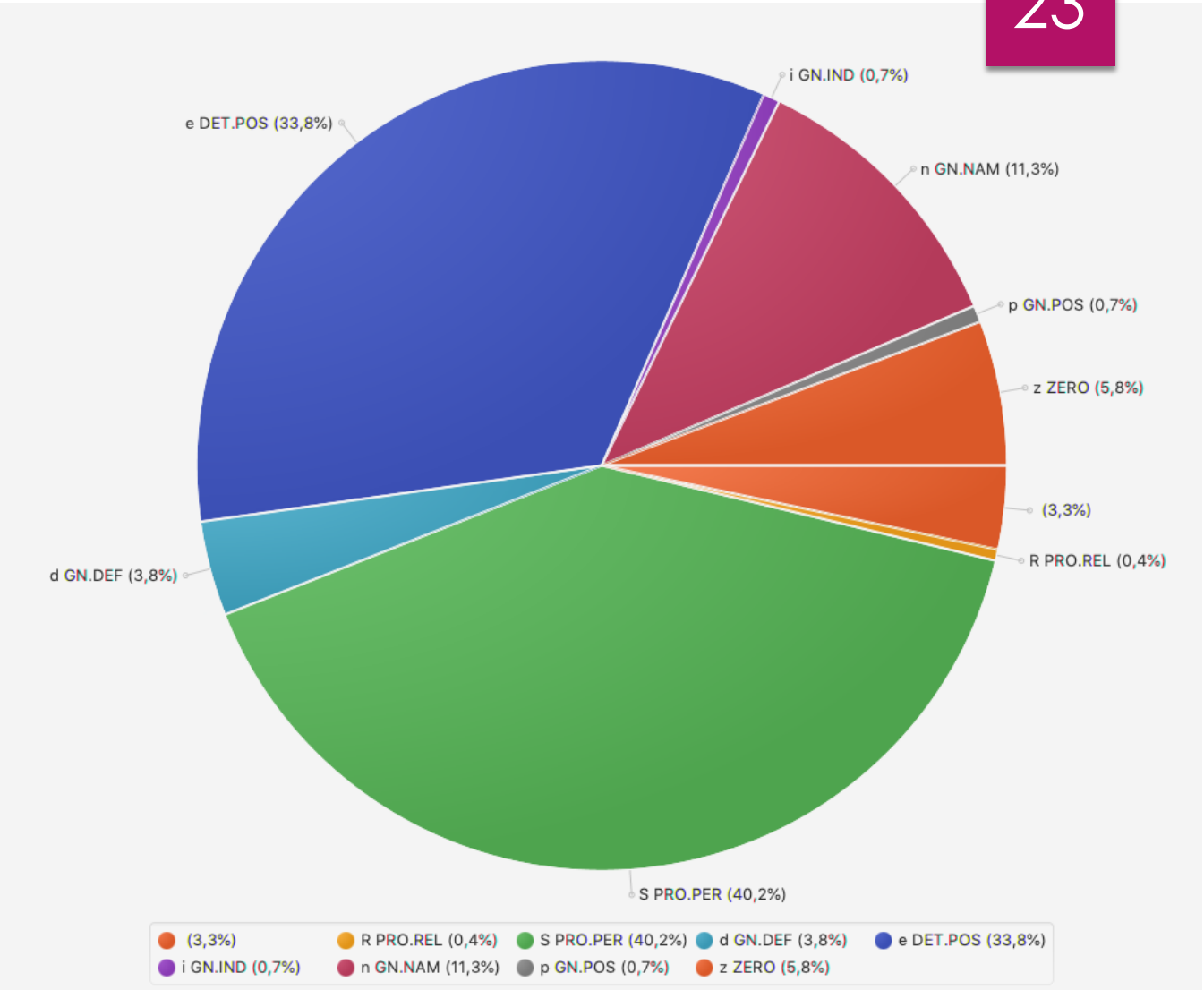
Democrat Project and Methodology

- Democrat Alsace = 2 tools developed by B. Oberlé SACR (2017) : SACR and CR Viewer
- SACR : annotation of RC with a colourful, user-friendly interface



Democrat Project and Methodology

- CR Viewer = calculate indicators and display pie charts
- i.e = grammatical categories of the links



Democrat Project and Methodology

- **Methodology of RC studies in Democrat =** C. Schnedecker (2017) has recently put forward a new, « configurational » approach that aims at unifying the various and heterogenous indicators taken into account in RC studies.
- **"Configurational" approach =** configuration of recurrent features, which combines the quantitative dimension of the **paradigmatic approach** (enumeration of certain grammatical and lexical categories) and the **syntagmatic approach** (linearity of RC : length and range, distances between the links...).

Democrat Project and Methodology

Some of the parameters =

- **Number of links.**
- **Length of links.**
- **Stability coefficient** (Perret 2000) = lexical variety
- **Number of RC in the text.**
- **Length of RC.**

Democrat Project and Methodology

- **Distance between links** (cf. Ariel 1990 = as a factor correlated to referential salience and widely used in many works).
- **Scope of RC** = understood by Boudreau & Kittredge (2005) as the percentage of text covered by each RC , ie concretely the number of paragraphs covered by the same RC.
- **Composition of RC** (grammatical categories and functions of the links).
- **Order** in which RC appear in the text.
- **Way in which RC coexist** and **interact** with each other (succession, interweaving, merging, dissociation, parallelism…)

Democrat Project and Methodology

- One of the results of Democrat so far = need to take the notion of « **genre** » into account in RC studies.
- There is a **strong influence** of text / discourse genres on RC
- I.E = C. Schnedecker (2014) = RC in recipes vs. encyclopedic texts

ER/CR	Recettes	Introductions de textes encyclopédiques
Densité des ER/texte	34 %	29 %
Nombre de CR	68	31
Longueur moyenne des CR	4.85	6.41
Portée des CR	locale	globale
Distance inter-maillonnaire	< 10 caractères	>150 caractères
Persistance des CR	-	-
Coefficient de stabilité	max	max
Coïncidence CR/thème texte	-	+
Fonction dominante des maillons	COD	Sujet > CDN
Mode de cohabitation	Sucession/fusion	Succession/dérivation
Modélisation de CR	SNQ... SNdéf.$_{(COD)}$... SNdéf.$_{(COD)}$	§1 Npr1 ... §2 Npr2 ... §N Nprn ...

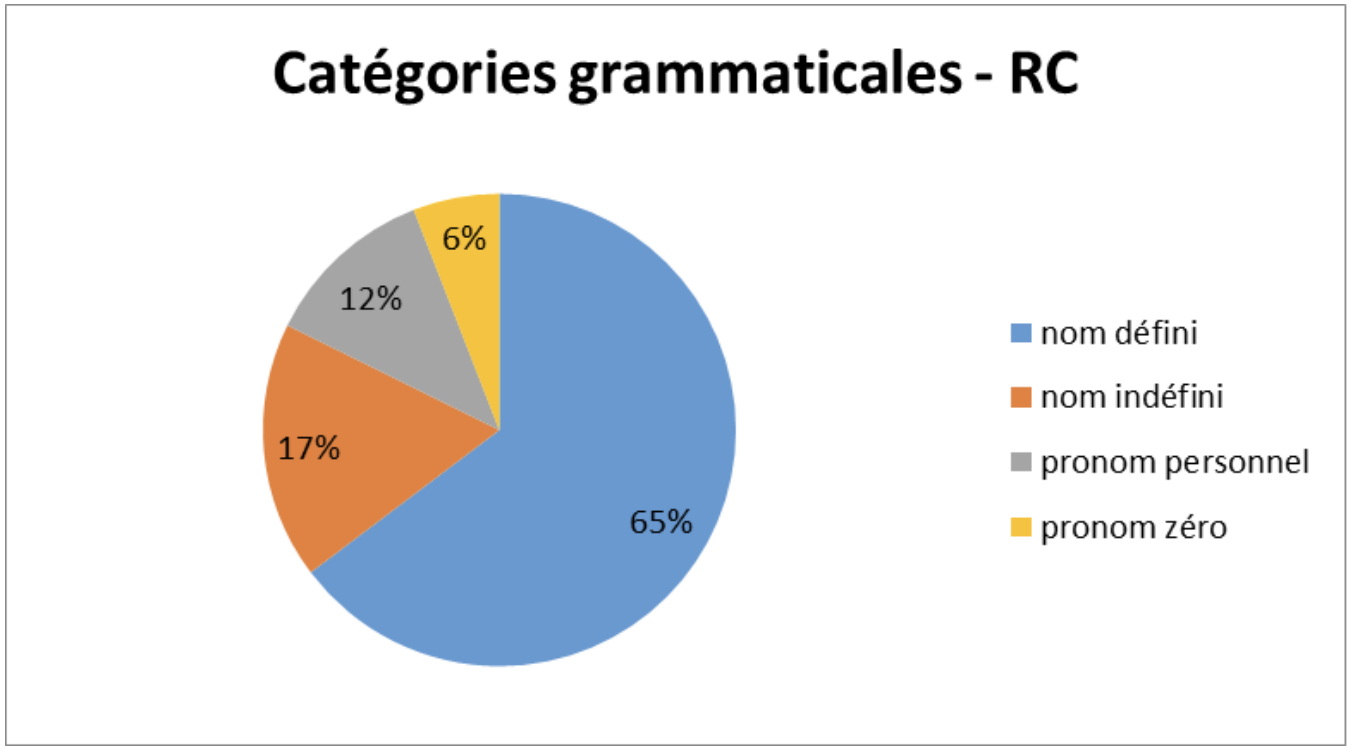
Democrat Project and Methodology

Main differences =

- Density of referential expressions in recipes is slightly higher than that of encyclopedic texts.
- Number of RC twice as high in the corpus of recipes as in that of encyclopedic texts.
- Length and scope of RC in recipes is shorter than that of the encyclopedic texts.
- In recipes, RC (which follow one another or even merge) are generally shorter, but remain limited to 1 (brief) paragraph. In contrast, RC of encyclopedic texts are longer, with a scope that extends over the entire text.

Democrat Project and Methodology

- Main grammatical functions = **Object** for recipes, vs. **Subject** for encyclopedic texts
- Main grammatical categories = **undefinite NPs** + ∅ **anaphors** for recipes, vs. **proper nouns** and **definite NPs** for encyclopedic texts



Democrat Project and Methodology

⇒ **<u>Genre = important factor, but not the only one !</u>**

- Small cross-linguistic group of fellow researchers, inside Democrat (Dominique Dias and Emmanuelle Prak-Derrington = German, Laure Gardelle = English, Catherine Schnedecker = French).
- In addition to my own work on RC in English vs. French, some recent articles with my colleagues => **a few aspects of recent cross-linguistic studies of RC.**

3/ Contrastive Studies of RC

Contrastive studies of RC

- Baumer (2015) : cross-lingusitic analysis of journalistic portraits in French and English
- Same sub-genre (of journalistic texts), yet **some significant differences** in RC!

I.E = In English texts, only **3.5 %** of nominal anaphors

Vs., in French texts **19.2%** of nominal anaphors.

ENG

Edward Behr, who died on Saturday aged 81, enjoyed a long career covering wars from Algeria to Angola and Vietnam for American magazines before turning to writing books and television documentaries. Notable among his books were The Algeria Problem; Hirohito: Behind the Myth; The Last Emperor, a study of the boy emperor Pu Yi; and Kiss the Hand You Cannot Bite, an account of the rise and fall of the Ceausescus in Romania.

In Hirohito, published in 1989, **Behr** concluded that, far from being the innocent tool of the Japanese military in their rampages through Asia and the South Seas, as he came to be portrayed in postwar mythology, the emperor had actually been aware since the early 1930s of exactly what was going on. One of the most hated men in history, **Behr** said, became the ultimate survivor.

His book on the Ceausescus offered some telling footnotes on their state visit to London in 1978. The Queen, **Behr** wrote, never forgave the Prime Minister James Callaghan and David Owen, then foreign secretary, for persuading her to have the Romanians to stay at the Palace. It was a condition the Ceausescus imposed in return for their visiting Britain and for placing orders for aircraft and other arms. Their staff stole items from their quarters, **Behr** said, and the Queen referred privately to Ceausescu as "that frightful little man".

Edward Behr was born on May 7 1926 in Paris. His parents were of Russian descent, and he had a bilingual education at the Lycée Janson de Sailly and St Paul's School, London. He enlisted in the Indian Army on leaving school, serving in Intelligence in the North-West Frontier from 1944 to 1948 and rising to acting brigade major in the Royal Garhwal Rifles at the age of 22. (telegraph)

FR

(Vincent Boloré) A l'extérieur du groupe, les propos sont moins amènes. Pour se tailler une place dans le PAF, **le nouveau venu** s'est fait de solides inimitiés. Quelques-unes de plus! Hier, Martin[Champ] Bouygues. Aujourd'hui, Maurice Lévy. Si le président de Publicis salue «l'une des plus belles histoires du capitalisme français», il se montre cinglant dès qu'il s'agit d'évoquer les méthodes de **l'industriel**. Voilà un mois, trois pointures de Publicis - Frédéric Raillard, Farid Mokart et, surtout, Christophe Lambert (PDG de Publicis Conseil) - ont démissionné sans crier gare pour créer leur propre agence avec l'aide de Bolloré. Lévy n'a pas digéré cet épisode. De quoi hypothéquer, un jour ou l'autre, toute idée d'accord entre Havas et Publicis? «Peut-on raisonnablement s'asseoir à la même table que **ce monsieur**? Cela pose à mon sens quelques problèmes d'éthique...» Au-delà des querelles, **le trouble-fête** continue de soulever des questions. En particulier sur la pérennité de ses engagements. Que restera-t-il de ce bel ensemble médiatique dans cinq ou dix ans? Depuis qu'il a rejoint l'entreprise familiale, en 1981, **ce patron marginal, dans et hors de l'establishment** - conseiller de la Banque de France et redoutable raider - a constamment semblé chercher sa voie, n'hésitant pas à chambouler les activités du groupe. (lexpress)

Contrastive studies of RC

- **Qualification element** => nominal head of the NPs = **Point of view** indirectly conveyed.
- English and French function differently to convey **indirect, masked subjectivity** on the part of the enunciator-journalist.
- In French = **syntaxically integrated nominal anaphors** (subject, object) that play this role, owing to their **implicit predicative value**.
- I.e = if you say "this terrible singer", you **IMPLY** the person sings badly, but without actually saying [he **IS** a terrible singer], or [**I think** he is a bad singer].

Contrastive studies of RC

In English, the situation is quite different, since nominal anaphors are not commonly used: the enunciator either **directly asserts the value judgement** (thus taking responsibility for the content of his sentences):

- I.e : De Kooning **is** one of our most famous artists.
- Ciara **is** arguably the greatest dancer in pop music right now, someone who approaches dance as an art form rather than merely background choreography.

Contrastive studies of RC

Otherwise, the enunciator **doesn't take responsibility for the judgement value** by using (most of the time) **the passive form** (with or without an agent): he / she's said / rumoured to be, …

Or : he / she's been called / branded / hailed / billed/ described / dubbed / tipped / admired (as), …

- Ex : **Despised in New York as a lame-duck mayor** through much of his second term, Giuliani today is leading in opinion polls among contenders for the Republican presidential nomination.
- Nicolas Sarkozy **is often described as an Atlanticist**.
- **Described by Gore Vidal as the worst American writer ever**, Kurt Vonnegut was none the less at one point said to be the most widely taught in American universities.

Contrastive studies of RC

2 other recent contrastive studies of RC =

- Baumer E., Dias D., Gardelle L. et Prak-Derrington E. (à paraître), « A quel point les chaînes de référence peuvent-elles être homogènes au sein d'un même genre discursif? Exploration contrastive allemand / anglais / français », Revue de Sémantique et de Pragmatique.
- Baumer E., Schnedecker C., Dias D. (à paraître), « Les chaînes de référence dans un corpus contrastif (allemand-anglais-français) de romans réalistes du 19ème siècle : analyse quantitative et qualitative ».

Contrastive studies of RC

The 2 studies show similar results :

- 1/ RC appear to be quite similar if you take into account only general indicators and parameters. But a more qualitative, subtle (« configurational ») analysis reveal differences that must be accounted for.
- 2/ in both cases, the usual categorisation **Germanic languages** vs. **Romance languages** (i.e L. Lundquist 2005) is challenged.

Contrastive studies of RC

[**German and French**] Vs. [**English**]

- For instance : In our study of 19th century novel incipits, French and German = rather similar way of **delaying the main character's appearance**
- + Difference in the **composition of RC (more proper nouns in English, vs. more NP in French and German)**
- In both studies, same strategies in **German and French** to handle **lexical variety**, etc…

Contrastive studies of RC

- Chinese = interesting to find out how RC work in Chinese and compae with other languages =>
- Chang Guo's Phd research subject = RC in French Vs. Chinese
- Anaphoric and referential forms = seem fewer than in French, for instance.
- But A lot of ⌀ **pronouns.**
- Anaphoric and referential forms = less numerous **if we do not count ⌀ forms**, but interesting to study how reference is maintained in some ambiguous contexts

Contrastive studies of RC

- For instance : 1. subject ø back-reference. 2. object ø back-reference. 3. possessive determiner subject ø back-reference.

1. 他喝了杯咖啡, 走了。(Tā hēle bēi kāfēi, zǒule) He drank a coffee, then ø left.

2. 他喝了杯咖啡, 很烫。(Tā hēle bēi kāfēi, hěn tàng) He drank a coffee, ø was warm.

3. 他的咖啡很烫, 不能喝。(Tā de kāfēi hěn tàng, bùnéng hē) His coffee was very hot and ø couldn't be drunk.

Conclusion

Conclusion

- Just a few aspects of cross-linguistic studies of RC.
- Some results are fairly recent and thus need to be confirmed by other studies (based on the Democrat online corpus, for instance).
- I.e Interaction of genres and languages in RC = further specific analysis.
- Some new cross-linguistic research based on other languages (Chang Guo = Chinese, Seo Yeoung Kim = Korean) will provide new interesting data and perspectives.

Bibliographie Indicative

- Adam J.-M. et Petitjean A. (1989), *Le texte descriptif*. Paris : Nathan.
- Baumer E., Dias D., Gardelle L. et Prak-Derrington E. (à paraître), « La globalisation communicationnelle à l'épreuve des chaînes de référence : étude exploratoire d'un corpus de presse trilingue » (français / anglais / allemand).
- Baumer, E., à paraître, « A quel point les chaînes de référence peuvent-elles être homogènes au sein d'un même genre discursif? Exploration contrastive allemand / anglais / français », avec L. Gardelle, D. Dias et E. Prak Derrington, Revue de Sémantique et de Pragmatique.
- Baumer, E., (2018), « Chaînes de référence et structuration des textes. Le cas des textes encyclopédiques de l'ancien français à nos jours. Étude pilote », avec B. Oberlé, C. Schnedecker, D. Capin, J. Glikman, C. Guo, T. Revol, A. Todirascu, J. Tushkova, Revue Romane, 2018.
- Baumer, E. (2017),
- Baumer, E. (2015) *Noms propres et anaphores nominales en anglais et en français : étude comparée des chaînes de référence*. Paris : L'Harmattan.
- Charolles, M. (1988). La gestion des risques de confusion entre personnages dans une tâche rédactionnelle. *Pratiques*, *60*(1), 75-97.
- Charolles M., (2002), *La référence et les expressions référentielles en français*, Paris, Ophrys.
- Corblin, F. (1985). Les chaînes de référence : analyse linguistique et traitement automatique. *Intellectica. Revue de l'Association pour la Recherche Cognitive*, *1*(1), 123-143.
- Kleiber, G. (1996). Noms propres et noms communs : un problème de dénomination. *Meta: Journal Des Traducteurs*, *41*(4), 567.
- Landragin F. & C. Schnedecker (éds) (2014) *Les chaînes de référence*, Langages 185.
- Schnedecker C. (1989) La dénomination du personnage en contexte dialogué, *Pratiques* 64, 39-67.
- Schnedecker C. (1990) "Le discours rapporté a-t-il des incidences sur les chaînes de référence? Quelques observations", *Verbum* XIII/3, 1990, 165-190.
- Schnedecker C. (1997), *Nom propre et chaînes de référence*. Paris : Klincksieck.
- Schnedecker, C. (2005). Les chaînes de référence dans les portraits journalistiques : éléments de description. *Travaux De Linguistique*, *51*(2), 85.
- Schnedecker C., Glikman J. et Landragin F., éd. (2017), *Les chaînes de référence en corpus. Langue française*, n° 195.

lilpa
linguistique, langues, parole

BCL
UMR 7320, CNRS - ILF - UNS

Xie Xie !! / Merci !! / THX !!

democrat

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR
ANR