Étude des chaînes de référence dans les articles de recherche de format IMRaD

Bruno Oberle

Journée d'Étude "Référence, coréférence et structure textuelle"
(27 novembre 2017, ENS Lyon)

ANR-15-CE38-0008

Introduction: le format IMRaD

- **structure standardisée des articles de recherche en sciences expérimentales**
  - **I**ntroduction
    - présentation la recherche et son contexte
  - **M**ethods
    - description des données et des traitements
  - **R**esults
    - observation des résultats de l'expérimentation
  - **a**nd **D**iscussion
    - interprétation des résultats au regard de la littérature scientifique antérieure et réponse à la question formulée dans l'introduction

Introduction: le format IMRaD

- **variation de phénomènes linguistiques selon la section d'occurence (Swales 1990, 2004):**
  - la répartition du pronom "on" (Gjesdal-Müller 2013)
  - le domaine lexical des verbes (Reimerink 2006)
  - la distribution des citations et références (Bertin & Atanassova 2014)
  - etc. (voir Swales 1990 pour une liste plus complète)

Introduction: les chaînes de référence

- **définition: l'ensemble des expressions linguistiques qui renvoient à la même entité extra-linguistique (Corblin 1985)**
  - ex.: **[La Grotte I des Treilles]** a été découverte par L. Balsan, M.R. Galzin et J. Maillé en 1933. Louis Balsan **[y]** a effectué des fouilles... **[Ce site]** a été attribué au Chalcolithique...
- **les expressions qui composent la chaîne sont appelées des "maillons"**
- **une chaîne a trois maillons au minimum**

Introduction: les chaînes de référence

- **indicateurs**
  - le nombre de maillons
  - la distance entre les maillons
  - la portée (locale ou globale)
  - la catégorie et la fonction grammaticale des maillons
  - les patrons
  - le coefficient de stabilité
  - etc.
- **ces indicateurs varient selon le genre (Schnedecker 2014, 2017)**

Introduction: hypothèse

- **tout comme les chaînes varient en fonction du genre d’occurrence, elles varieraient en fonction de la section IMRaD dans laquelle elles apparaissent**

Le format IMRaD

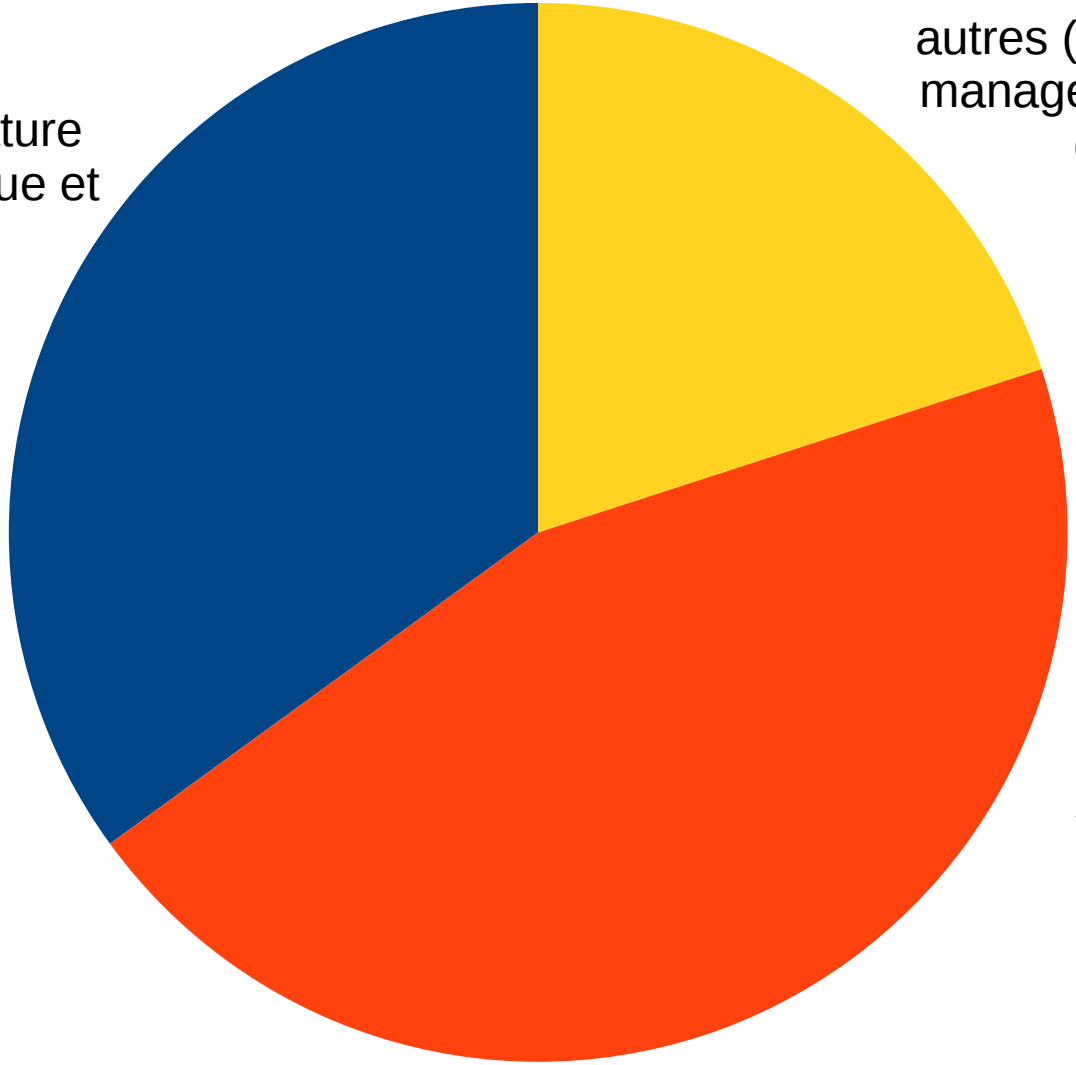
- **issu des sciences expérimentales (chimie, biologie, médecine, physique, etc.)**
- **s'étend à d'autres domaines, notamment la psychologie expérimentale (cf. *Publication Manual* de l'*American Psychological Association – APA*)**
- **standardisation de la présentation des expérimentations**
- **codification (Bazerman 1988, Swales 1990), qui permet de comparer et d'opposer les sections**

IMRaD en français sur revues.org

- **analyse de la plateforme en libre accès revues.org (Sciences Humaines et Sociales)**
- **analyse des tables des matières**

1) les 4 derniers numéros (si disponibles) de 341 revues, soit 21 689 articles
   - 255 articles IMRaD en français (soit 2% du total)
2) les 10 derniers numéros (si disponibles) des 20 revues qui avaient le plus d'articles IMRaD, soit 1 580 articles
   - 314 articles IMRaD en français (soit 20% des articles de ces 20 revues)

IMRaD en français sur revues.org



Corpus retenu

- 2 textes en sciences de l'éducation
- 2 textes en archéologie
- 1 texte en psychologie

Particularités des articles de recherche

- **des référents hétérogènes, ex.:**
  - des référents humains:
    - des individus (l'auteur, "nous")
    - des groupes (la population étudiée, les échantillons)
  - des référents abstraits (l'interdisciplinarité, le burnout)
  - des noms massifs (le carbone, l'azote)
  - des entités nommées (la Grotte I des Treilles)
  - des dates (1933)
  - etc.
- **des chaînes au comportement, à la composition, à la distribution différents**
- **première typologie (humain vs non humain): Longo & Todirascu 2014**
- **des chaînes de longueur et de portée très variables**

L’annotation: les référents saillants

- **sélection d’une vingtaine de référents saillants par texte, à partir**
  - du résumé
  - des mots-clés
  - d’une analyse lexicométrique (*via* AntConc)

L’annotation: les référents saillants

	texte	tokens	CR	maillons
éducation	T2	7 123	22	622
	T4	8 627	17	422
archéo	T1	6 008	20	292
	T3	4 318	15	186
psycho	T0	6 463	15	434
TOTAL		32 539	89	1 956

Les chaînes saillantes

- **des chaînes de portée globale mais distendues**
  - présence sur l'ensemble du texte
  - des chaînes brèves: 22 maillons en moyenne
  - distance intermaillonnaire: 330 mots
  - longueur moyenne des paragraphes: 120 mots
  - donc ***un maillon tous les trois paragraphes***
- **exemple: la chaîne "burnout":**
  - **32 maillons**
  - **présente dans 17 paragraphes:**
    - **du premier au 53e (sur 55)**
    - **11 paragraphes n'ont qu'un seul maillon de cette chaîne**
    - **5 paragraphes ont 2 à 4 maillons**
    - **1 seul paragraphe en a plus (7 maillons)**
  - **absente dans 38 paragraphes**

Les chaînes saillantes

- **une forte stabilité**
  - un coefficient de stabilité de 0.8 (sur 1)
- **peu de pronoms**
  - 13 % (vs 30 à 40 % dans des textes plus narratifs comme les portraits journalistiques ou les faits divers (Schnedecker 2005, Schnedecker & Longo 2012))

Les sections IMRaD: chaînes saillantes

- **intérêt pour la distinction entre différents types de référents**
  - hypothèse: les chaînes n’ont pas le même comportement selon le type de référent
  - ex.: les noms massifs (“l’eau, l’azote”) n’initient pas le même type de chaîne que les ensemble (“les étudiants”) ou l’auteur (“nous”)

Les sections IMRaD: chaînes saillantes

- **différents types de référents**
  - auteur (ex.: nous)
  - recherche et article (ex.: notre article)
  - entités nommées et référents définis (ex.: la Grotte des Treilles)
  - ensembles (ex.: les étudiants)
  - massifs (ex.: l’azote)
  - référents génériques (ex.: le vétérinaire rural)
  - noms abstraits (ex.: l’interdisciplinarité)
  - noms prédicatifs (ex.: la modération)
  - ensembles “flous” (sans limites clairement définies: “les vestiges archéologiques”)
  - “variables liées” (référence spécifique mais non particulière: “un capitaine d’équipe”)

Les sections IMRaD: chaînes saillantes: l'auteur

- **un individu ou un petit groupe d'individus bien définis**
- **s'inscrit dans des recherches sur l'expression de l'auteur scientifique**
- **la chaîne de l'auteur**
  - seulement dans 2 textes (de façon significative)
  - presque exclusivement des pronoms ("nous") et des déterminants ("notre, nos")
  - presque exclusivement sujet
- **une répartition inégale**
  - 36 % des maillons de la chaîne dans l'introduction:
    - nous proposons ici de contribuer...
    - le cadre dans lequel nous nous situons...
    - à notre connaissance...
  - 6 % des maillons dans la méthodologie
- **des résultats en désaccord avec d'autres recherches (Regent 1980, Heslot 1983, Müller Gjesdal 2013)**

Les sections IMRaD: chaînes saillantes: les ensembles

- **les ensembles: des référents incontournables dans les études statistiques**
  - la population étudiée
  - l'échantillon (les participants à l'expérimentation)
- **une répartition inégale**
  - 56 % des maillons dans les résultats et la discussion
  - plus forte proportion d'ensembles (par rapport aux autres types de référents) dans la méthodologie: un tiers des maillons de cette section sont des ensembles
    - description des groupes de participants et leurs tâches
      - **Les étudiants** appartiennent à 1 049 établissements différents
      - une communauté d'**étudiants** de 99 pays différents
      - **Ils** devaient d'abord sélectionner **les cinq items**

Les sections IMRaD: chaînes saillantes: les ensembles

- **une évolution des fonctions: méthodologie vs. conclusion**

	méthodologie	conclusion
comp. du nom	43%	69%
comp. d'objet	29%	10%

  - dans la méthodologie, des CDN et des CO
    - on donne des tâches aux participants (CO)
      - Il était demander **aux étudiants**
      - il **leur** a été proposer de hiérarchiser les évocations produites
    - les compléments du noms permettent de situer les participants:
      - le niveau **des étudiants**
      - les informations sociodémographiques **des étudiants**
  - dans la conclusion, on décrit la population → seulement des CDN
    - les spécificités de **la population d'élèves étudiée**
    - le comportement **des étudiants**

Les sections IMRaD: chaînes saillantes: les noms abstraits

- les référents abstraits
  - ex.: l'interdisciplinarité, l'entreprenariat, l'innovation, le burnout
- concentration dans l'introduction
  - la moitié de tous les maillons de l'introduction renvoient à des concepts abstraits
  - l'introduction présente les concepts étudiés, qui sont généralement abstraits
- composition:
  - des GN définis (69%) et sans déterminant (19%: "facteurs **de stress**")
  - des compléments du noms (49%): "le modèle **du burnout**", "les dimensions **du burnout**", "le goût des étudiants pour **l'interdisciplinarité**"
  - un coefficient de stabilité élevé: 0.9 → peu d'anaphores infidèles

Des chaînes saillantes aux chaînes de paragraphe

- **les chaînes saillantes sont distendues (1 maillon tous les 3 paragraphes)**
- **d'autres chaînes sont moins saillantes au niveau du texte mais jouent un rôle important au niveau du paragraphe**
- **le paragaphe**
  - unité sémantique (Bessonnat 1988)
  - influence sur les relations anaphoriques et les chaînes de référence (Schnedecker 1997, 2014, Ariel 1990, Huang 2000)

L’annotation des chaînes de paragraphe

- **“chaîne de paragraphe”**
  - une chaîne qui a trois maillons ou plus dans les limites d’un paragraphe
  - annotation systématique (pas de présélection)
- **un traitement différent des “chaînes de texte” et des “chaînes de paragraphe”**

L’annotation des chaînes de paragraphe

	texte	tokens	CR	maillons
éducation	T2	7 123	82	396
	T4	8 627	78	343
archéo	T1	6 008	35	147
	T3	4 318	9	28
psycho	T0	6 463	--	--
TOTAL		32 539	204	914

Les chaînes de paragraphe

- **des chaînes brèves et resserrées**
  - 204 chaînes
  - longueur: 4.5 maillons
- **62 % des paragraphes n’ont pas de chaîne**
  - segmentation importante des textes
  - concentration des chaînes dans quelques paragraphes

Les chaînes de paragraphe

- **17 % de pronoms (vs 12 % pour les chaînes saillantes)**
- **des reprises plus diverses**
  - coefficient de stabilité: 0.74 (vs 0.80 pour les chaînes saillantes)
- **ex.:**
  - **[La Grotte I des Treilles]** a été découverte par L. Balsan, M.R. Galzin et J. Maillé en 1933. Louis Balsan **[y]** a effectué des fouilles... **[Ce site]** a été attribué au Chalcolithique...
- **distance intermaillonnaire: 40 tokens**

Les chaînes “éphémères”

- **un tiers des chaînes de paragraphe**
- **chaînes de trois ou quatre maillons, dont la distance intermaillonnaire est de moins de 20 tokens, souvent dans une même proposition**
- **généralement un nom suivi de pronoms ou de déterminants, souvent coordonnés**
- **ex.:**
  - **[Tous les arguments] [qui]** viennent d’être mentionnés **et [qui]** auraient été susceptibles d'expliquer l’aversion des étudiants pour l'interdisciplinarité
  - Il n'a pas été possible, pour des raisons de disponibilité, d'inclure des étudiants préparant **[le concours C]**, alors que **[sa]** spécificité **et [son]** mode de recrutement pourraient avoir un impact sur les représentations des étudiants

Chaînes liées et chaînes uniques

- **un référent peut initier**
  - une seule chaîne dans un seul paragraphe → “chaîne unique”
  - des chaînes dans plusieurs paragraphes → “chaînes liées”
- **105 référents initient une ou plusieurs chaînes**
  - 79 (soit 75%) dans un seul paragraphe
    → 79 chaînes (soit 39% des chaînes)
  - 26 (soit 25%) dans plusieurs paragraphes
    → 132 chaînes (soit 61% des chaînes)

Chaînes liées et chaînes uniques

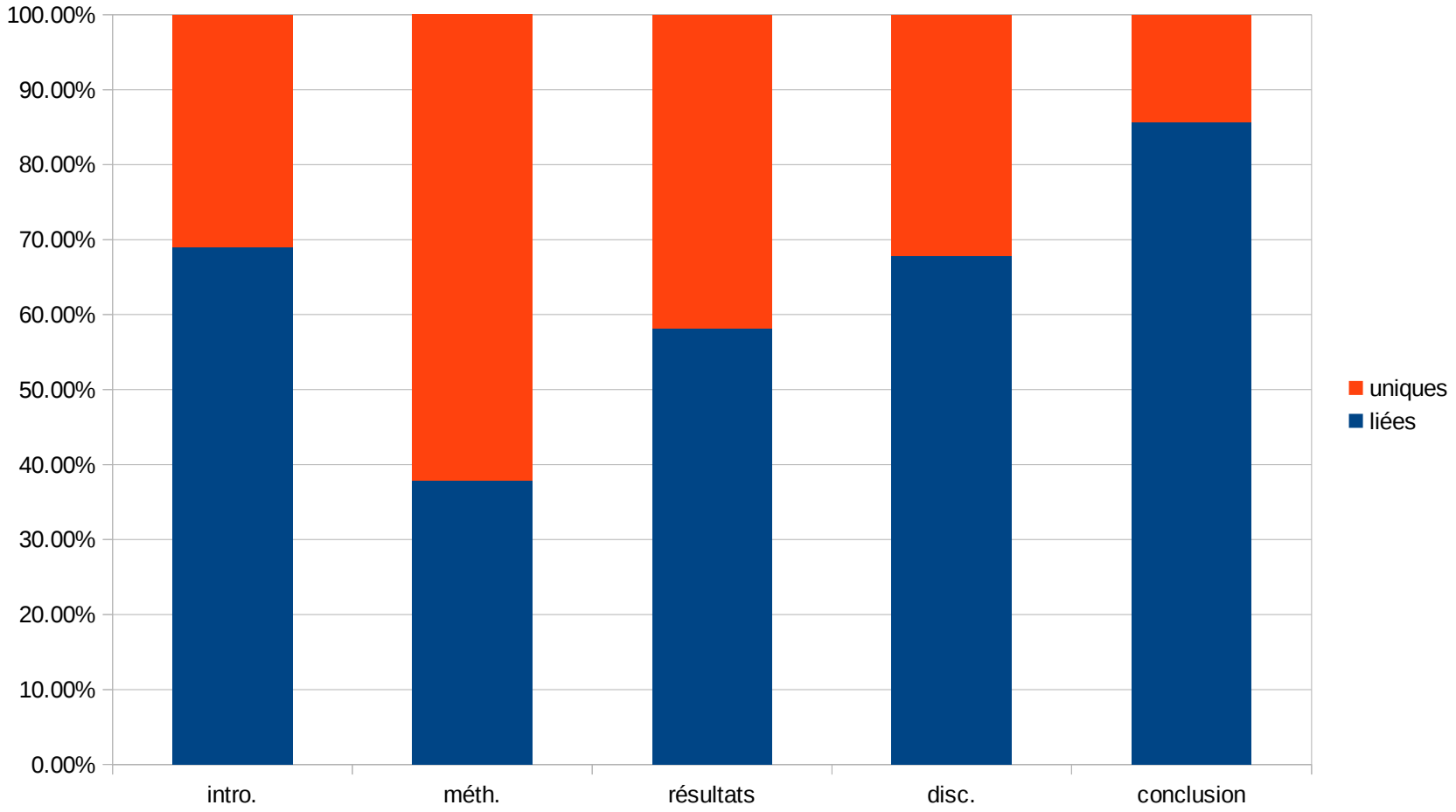
- **chaînes liées**
  - ouvertes sur le reste du texte
  - liées aux paragraphes précédents et suivants
  - support transitoire de l’information
  - participent à la cohérence et à la cohésion du texte
- **chaînes uniques**
  - leur cycle de vie est entièrement compris dans un paragraphe
  - repliées sur elles-mêmes

Chaînes uniques

- **plus de pronoms**
  - 22% contre 15% pour les chaînes liées
  - surtout des relatifs → expansions
- **d’où le patron**
  - GN défini > relatif > GN défini/démonstratif
  - généralement sujet
  - note: la relative de ce patron est déterminative, puisque la chaîne commence par un défini
  - ex.: C'est le cas notamment des appareils monospire comme **[le MS2D de Bartington] [qui]** a été utilisé pour la plupart des études sur le projet Canal Seine-Nord Europe. **[Cet appareil]** permet une mesure point par point de la susceptibilité magnétique volumique.

Les sections IMRaD: chaînes de paragraphe

- **chaînes liées vs. chaînes uniques**

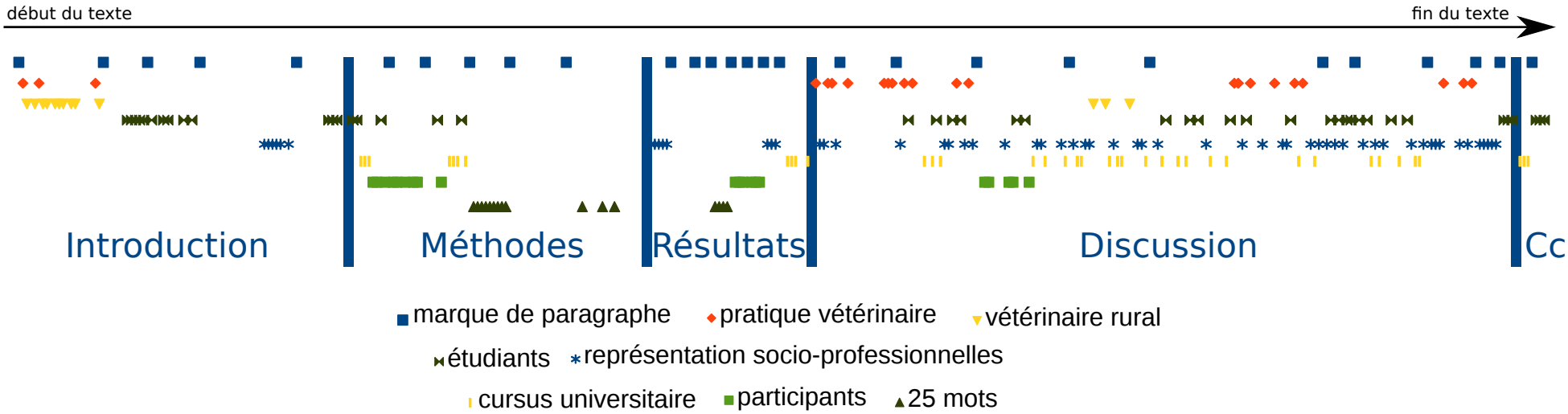


Les sections IMRaD: chaînes de paragraphe

- **dans la méthodologie:**
  - près de deux tiers de chaînes uniques: les référents ne dépassent pas les limites d’un paragraphe
  - succession de brèves descriptions d'outils et de procédures
  - ex.:
    - Dans un premier temps, un recueil du corpus sémantique de la représentation a été réalisé auprès de **[38 étudiants]** à l'aide de la méthode du réseau d'association (De Rosa, 2003). Il a été demandé **[aux participants]** de noter spontanément au plus **[dix mots ou courtes expressions] [qu]**'**[ils]** associaient à l'expression stimulus: "activité vétérinaire rurale". **[Ils]** devaient aussi numéroter en chiffres arabes, au fur et à mesure de **[leur]** apparition, **[ces mots ou expressions**...

Les sections IMRaD: chaînes de paragraphes

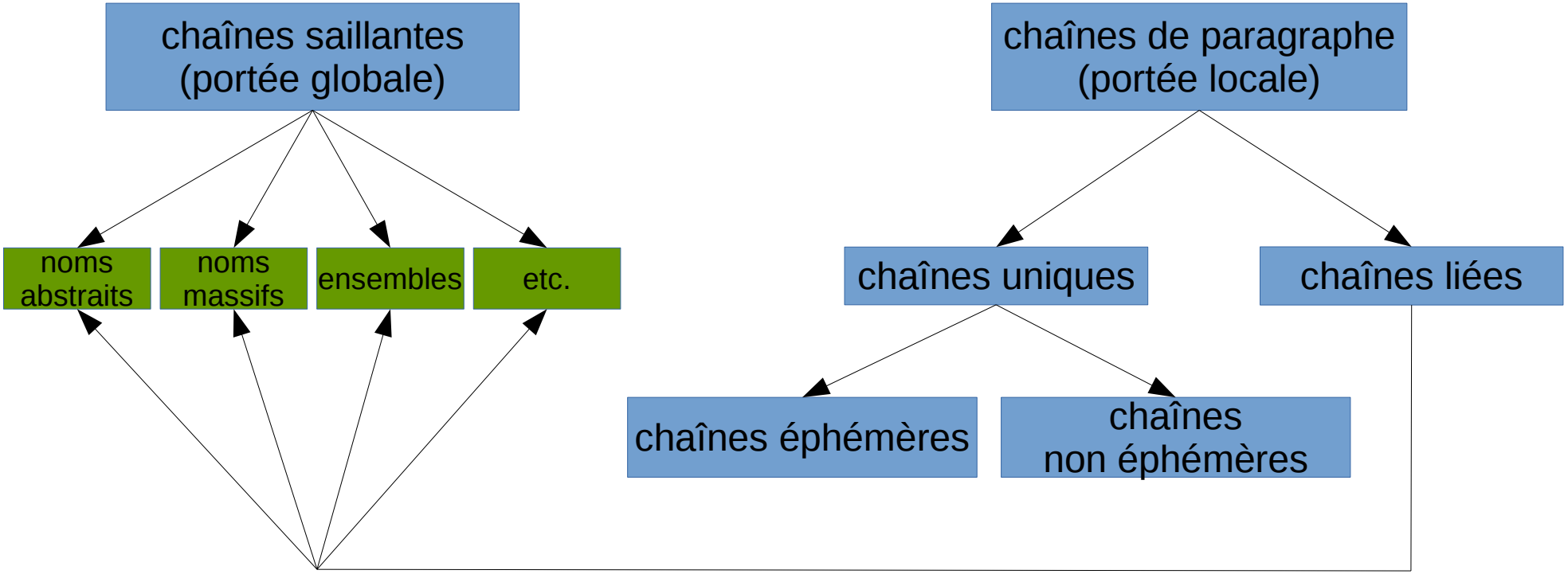
- **chaînes liées vs. chaînes uniques**



Conclusion: bilan

- **le format IMRaD**
  - dans le paysage SHS français
  - caractéristiques propres des chaînes de référence dans chaque section (lien avec la fonction rhétorique de chaque section)
- **réflexion sur une typologie des chaînes de référence**

Conclusion: bilan



Conclusion: limites et perspectives

- **le format IMRAD**
  - des pistes à confirmer ou infirmer avec un corpus plus large
  - mise en relation avec d'autres phénomènes linguistiques (lexique, type de verbes)
  - étude plus fine des sections (les "mouvements" de Swales 2004)
  - différence entre les disciplines
- **typologie des chaînes:**
  - une approche plus systématique
  - reprendre la catégorisation des référents

Bibliographie

- Ariel M. (1990). Accessing noun-phrase antecedents. London ; New York : Routledge.
- Bazerman C. (1988). Shaping written knowledge : the genre and activity of the experimental article in science. Madison : University of Wisconsin Press.
- Bertin M. & Atanassova I. (2014). A study of lexical distribution in citation contexts through the IMRaD standard. In Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval co-located with 36th European Conference on Information Retrieval (ECIR 2014), p. 5–12.
- Bessonnat D. (1988). Le découpage en paragraphes et ses fonctions. Pratiques, 57, 81–105.
- Corblin F. (1985a). Les chaînes de référence : analyse linguistique et traitement automatique. Intellectica, 1 (1), 123–143.
- Heslot J. (1983). Récit et commentaire dans un article scientifique. Documentation et recherche en linguistique allemande contemporaine, 29, 133–54.
- Huang Y. (2000). Anaphora : a cross-linguistic approach. Oxford ; New York : Oxford University Press.
- Longo L. et Todirascu A. (2013). Une étude de corpus pour la détection automatique de thèmes. Revue électronique Texte et corpus, 4, 143–155.
- Müller-Gjesdal A. (2013). The Influence of Genre Constraints on Author Representation in Me- dical Research Articles. The French Indefinite Pronoun On in IMRAD Research Articles. Dis- cours. Revue de linguistique, psycholinguistique et informatique, 12.
- Régent O. (1980). Approche comparative des discours de spécialité pour l’entraînement à l’an- glais écrit. Mélanges Pédagogiques du CRAPEL.
- Reimerink A. (2006). The use of verbs in research articles : Corpus analysis for scientific writing and translation. New Voices in Translation Studies, 2, 9–27.
- Schnedecker C. (1997). Nom propre et chaînes de référence. Metz : Université de Metz
- Schnedecker C. (2014). Chaînes de référence et variations selon le genre. Langages, 195 (3), 23–42.
- Schnedecker C. (2017). Les Chaines de références: une configuation d'indices pour distinguer et identifier les genres textuels
- Swales J. M. (1990). Genre analysis : English in academic and research settings. Cambridge, New York : Cambridge University Press.
- Swales J. M. (2004). Research genres : explorations and applications. Cambridge, New York : Cambridge University Press.