Indices lexicaux pour le repérage de structures discursives

Clémentine Adam, Cécile Fabre, Philippe Muller

« Thématiques actuelles de la recherche en TAL »
17 décembre 2012

Plan

1. La cohésion lexicale en linguistique et en TAL
2. Détection de la cohésion lexicale par analyse distributionnelle
3. Interactions entre cohésion lexicale et structure rhétorique du discours

Plan

1. La cohésion lexicale en linguistique et en TAL
2. Détection de la cohésion lexicale par analyse distributionnelle
3. Interactions entre cohésion lexicale et structure rhétorique du discours

La notion de cohésion lexicale

- La cohésion :
  - Ce qui contribue à constituer un texte comme un tout cohérent

"the means whereby elements that are structurally unrelated to one another are linked together, through the dependence of one on the other for its interpretation« (Halliday & Hasan 1976)

- La cohésion lexicale :
  - La cohésion assurée par les relations sémantiques qui s'établissent entre les mots du texte.

"the cohesive effect achieved by the continuity of lexical meaning " (Halliday & Hasan 1976)

"When people read a text, the relations between the words contribute to their understanding of it." (Morris & Hirst 2004)

Types de relations impliqués dans la cohésion lexicale

Halliday & Hasan 1976

- Réitération
  - « the repetition of a lexical item, or the occurrence of a synonym of some kind, in the context of reference; that is, where the two occurrences have the same referent. »
- Collocation
  - The use of « a word that is in some way associated with another word in the preceding text, because it is a direct repetition of it, or is in some sense synonymous with it, or tends to occur in the same lexical environment. »

Exemple

- **Relations de réitération**

Les archéologues ont mis au jour le squelette d'un mammouth. L'animal a été découvert par hasard, lors de fouilles sur un site gallo-romain. Le mastodonte est un mammouth laineux, d'après les spécialistes. Ses deux défenses sont bien conservées.

Exemple

- **Relations de collocation**

Les archéologues ont mis au jour le squelette d'un mammouth. L'animal a été découvert par hasard, lors de fouilles sur un site gallo-romain. Le mastodonte est un mammouth laineux, d'après les spécialistes. Ses deux défenses sont bien conservées.

Rôle prédominant de la cohésion lexicale

Hoey, M. (1991). *Patterns of Lexis in text*.

- Contribution la plus importante parmi les procédés cohésifs
  - “the dominant mode of creating texture”
  - “important text-forming property of lexis”
- Etude quantitative menée sur des échantillons de 7 types de textes :
  - Recensement des marques de cohésion
    - > 40% => marques de cohésion lexicale
    - 22% => référence
    - 12% => conjonction
    - 10% => ellipse
    - 4% => substitution
- Teich & Fankhauser (2005) : ≈ 50% des liens de cohésion

Beaucoup d'autres typologies proposées ensuite …

- Abandon de la contrainte de référence
- Volonté de recentrage / de clarification de la notion de collocation
  - Hoey 1991 : collocation = "a ragbag of lexical relations, many of which have no readily available name"
  - Stubbs 2001 :
    - Des sous-relations (un peu) mieux identifiées
      - ensembles (couleurs, mois...)
      - collocations de type actanciel (*repas/manger, voiture/conduire*)
      - collocations « élaboratives » : *frames, scripts*

Trois principales difficultés pour le repérage de la cohésion lexicale

- Prédominance des relations de type collocation
- Contextualisation des relations sémantiques
- Apparition de la cohésion lexicale à différentes échelles de la structure discursive

Collocations / relations non classiques

- **J. Morris et G. Hirst (2004)**

  "overwhelming use of non-classical relations"

  " This type of relationship is the most problematic (…) Such collocation relationships exist between words that tend to occur in similar lexical environments. Words tend to occur in similar lexical environments because they describe things that tend to occur in similar situations or contexts in the world. Hence, context-specific examples such as {post office, service, stamps, pay, leave} are included in the class. (…) the exact relationship between these words can be hard to classify, but there does exist a recognizable relationship." (Morris & Hirst 2004)

  - Référence à :
    - **Lakoff** : relations classiques vs relations ad-hoc = "made up on the fly for some immediate purpose"

      Ex : "things to take on a camping trip" , "what to do for entertainment on a weekend"

    - **Cruse** : relations lexicales / affinités sémantiques

- Avec une forte dépendance au genre des textes :
  - (Berzlánovich, Egg and Redeker 2008)
    - *EE = encyclopedia entry, FL = fundraising letter*

Table 4: Lexical cohesive links

Type of cohesion	*EE01*		*EE02*		*FL01*		*FL02*	
Repetition	30	21 %	35	20 %	23	29 %	39	48 %
Systematic lexical links	104	73 %	131	76 %	33	42 %	6	7 %
Collocation	8	6 %	6	4 %	23	29 %	37	45 %
TOTAL	142	100 %	172	100 %	79	100 %	82	100 %

  - Gonzalez 2010
    - Conversations téléphoniques
      - Répétition (stricte ou avec variantes morpho) : 58%
      - Cohésion associative (*frames*) 24%
      - Relations classiques : 18 %

Dimension contextuelle des relations sémantiques

- (Stubbs 2001) : “(…) the present analysis does not start from ready-made classifications which would tell us which relations are possible; we start from a text and try to establish which items are related in that particular text. Hoey formulates this in a slightly different manner:

  A pair of clauses, ‘A was x, but B was only y’, create an instantial constrastive relation between the lexical items x and y, whatever their decontextualized relation. (Hoey 1991:220)”

- Une relation bidirectionnelle :
  - “the relationship between lexical and textual relations is not unidirectional (…) the text provides the context for the creation and interpretation of lexical relations, just as the lexical relations help create the texture of the text.” (Hoey 1991)

Une relation qui opère à tous les paliers de la structure discursive

- Chaînes topicales
- Relations de discours
- Réseaux localisés
- Cohésion globale sur l'ensemble du texte

Exemple

Formes de la cohésion lexicale

- “Often, *lexical cohesion occurs not simply* between pairs of words but over a succession of a number of nearby related words spanning a topical unit of the text.” (Morris & Hirst 1991)
- Deux modèles :
  - Les chaînes lexicales
    - sur le modèle des chaînes de référence
    - Hasan 1976 : *identity chains* (coréférence) / *similarity chains*
  - Les réseaux lexicaux
    - chaque mot peut entretenir des liens avec plusieurs autres mots du texte

Exemple : chaîne vs réseau (Berlanovich 2008)

(4) EDU5[After the ***forming***$_6$ of the ***sun***$_{1,3,4}$ and the ***solar system***$_4$, our ***star***$_{3,4}$ began ***its***$_3$ long ***existence***$_5$ as a so-called ***dwarf star***$_4$.] EDU6[In the ***dwarf phase***$_{4,6}$ of ***its***$_3$ ***life***$_{5,6}$, the *energy* that the ***sun***$_{1,3,4}$ *gives off is generated* in ***its***$_3$ *core* through the *fusion* of *hydrogen* into *helium.*] EDU7[The ***sun***$_{1,3,4}$ is about five billion ***years***$_2$ *old now*] EDU8[and ***it***$_3$ still has enough *fuel* for another five billion ***years***$_2$.]

(5) EDU5[After the *forming* of the ***sun***$_{\text{repetition(sunEDU4),hyponymy(starEDU4),co-hyponymy(Proxima Centauri EDU4),co-meronymy(EarthEDU3)}}$ and the ***solar system***$_{\text{holonymy(sunEDU4),holonymy(starEDU4), holonymy (ProximaCentauriEDU4), holonymy(EarthEDU3)}}$, our ***star***$_{\text{repetition(starEDU4),co-meronymy(EarthEDU3), hyperonymy (sunEDU4),hyperonymy(Proxima CentauriEDU4)}}$ *began* its *long existence* as a so-called ***dwarf star***$_{\text{hyperonymy(sunEDU5),hyponymy(starEDU4),co-meronymy(ProximaCentauriEDU4),co-meronymy(EarthEDU3)}}$.] EDU6[In the ***dwarf phase***$_{\text{co-meronymy(formingEDU5), collocation(dwarf starEDU5)}}$ of its ***life***$_{\text{mero-nymy(formingEDU5),synonymy(existanceEDU5)}}$, the *energy* that the ***sun***$_{\text{repetition(sunEDU5),hyponymy(starEDU5), hyponymy (dwarf starEDU5),co-meronymy(EarthEDU3),meronymy(solar systemEDU5),co-meronymy(Proxima CentauriEDU4)}}$ *gives off is generated* in its *core* through the *fusion* of *hydrogen* into *helium.*] EDU7[The ***sun***$_{\text{repetition(sunEDU6),hyponymy(starEDU5),co-meronymy(EarthEDU3),co-hyponymy(ProximaCentauriEDU4),meronymy(so-lar system EDU5),hypo-nymy(dwarf starEDU5)}}$ is about five billion ***years***$_{\text{repetition(yearEDU4)co-}}$

Exemple : chaîne vs réseau (Berzlanovich *et al.* 2008)

(4) EDU5[After the ***forming***$_6$ of the ***sun***$_{1,3,4}$ and the ***solar system***$_4$, our ***star***$_{3,4}$ began ***its***$_3$ long ***existence***$_5$ as a so-called ***dwarf star***$_4$.] EDU6[In the ***dwarf phase***$_{4,6}$ of ***its***$_3$ ***life***$_{5,6}$, the *energy* that the ***sun***$_{1,3,4}$ *gives off is generated* in ***its***$_3$ *core* through the *fusion* of *hydrogen* into *helium.*] EDU7[The ***sun***$_{1,3,4}$ is about five billion ***years***$_2$ *old now*] EDU8[and ***it***$_3$ still has enough *fuel* for another five billion ***years***$_2$.]

(5) EDU5[After the *forming* of the ***sun*** repetition(sunEDU4),hyponymy(starEDU4),co-hyponymy(Proxima Centauri EDU4),co-meronymy(EarthEDU3) and the ***solar system*** holonymy(sunEDU4),holonymy(starEDU4), holonymy (ProximaCentauriEDU4), holonymy(EarthEDU3), our ***star*** repetition(starEDU4),co-meronymy(EarthEDU3), hyperonymy (sunEDU4),hyperonymy(Proxima CentauriEDU4) *began* its *long existence* as a so-called ***dwarf star*** hyperonymy(sunEDU5),hyponymy(starEDU4),co-meronymy(ProximaCentauriEDU4),co-meronymy(EarthEDU3).] EDU6[In the ***dwarf phase*** co-meronymy(formingEDU5), collocation(dwarf starEDU5) of its ***life*** mero-nymy(formingEDU5),synonymy(existanceEDU5), the *energy* that the ***sun*** repetition(sunEDU5),hyponymy(starEDU5), hyponymy (dwarf starEDU5),co-meronymy(EarthEDU3),meronymy(solar systemEDU5),co-meronymy(Proxima CentauriEDU4) *gives off is generated* in its *core* through the *fusion* of *hydrogen* into *helium.*] EDU7[The ***sun*** repetition(sunEDU6),hyponymy(starEDU5),co-meronymy(EarthEDU3),co-hyponymy(ProximaCentauriEDU4),meronymy(so-lar system EDU5),hypo-nymy(dwarf starEDU5) is about five billion ***years*** repetition(yearEDU4)co-

Annoter la cohésion lexicale

- Manuellement
  - Par l'auteur lui-même : Hoey 1991, Legallois 2004
  - Par des annotateurs : Morris et Hirst 2004, Klebanov et Shamir 2006, Cramer et al. 2006
- Automatiquement
  - En se limitant à la relation de répétition stricte (Legallois et al. 2011)

  ou
  - Par projection de ressources lexicales
    - On fait l'hypothèse que les liens sémantiques répertoriés dans les ressources sont des bonnes approximations des liens sémantiques qu'on repèrerait à la lecture

Expériences d'annotation manuelle

- Trois exemples :
  - Morris & Hirst 2004, Klebanov & Shamir 2006, Cramer et al. 2006
- Principaux enseignements :
  - Une tâche difficile :
    - Difficultés à qualifier les relations entre mots
    - Mais bon taux d'accord sur les regroupements effectués
      - 63% selon Morris & Hirst
  - Des réseaux plutôt que les chaînes :
    - impossible pour les annotateurs de linéariser l'opération
  - Variété des relations :
    - Tendance à mobiliser des relations plus riches que les relations lexicales canoniques
  - Relations entre groupes de mots plutôt qu'entre mots isolés

De l'annotation manuelle à l'annotation automatique

- "Detailed manual analyses of small samples of text (e.g., Hoey, 1991) can bring out some tendencies of how lexical cohesion is achieved; but in order to arrive at any generalizations, large amounts of texts annotated for lexical ties are needed. Manual analysis is very labor-intensive, however, and the level of interannotator agreement is typically not satisfactory. Thus, an automatic procedure is called for. Fortunately, lexical cohesion analysis is a suitable candidate for automatization: Texts systematically make use of the semantic relations between words and detecting lexical cohesive ties simply means checking the relatedness of words in a text against a thesaurus or thesaurus-like resource." (Teich & Fankhauser 2005)

  => On oublie :

  - La contextualisation des relations sémantiques
  - L'importance des relations non classiques

Expériences d'annotation automatique

- Repérage de relations de répétition + projection de thesaurus :
  - Principe : Morris & Hirst 1991 (Roget's thesaurus)
  - Implémentation : Hearst 1992

…

  - (Teich & Fankhauser 2005) :
    - Pondérations : distance dans le texte, degré de généricité dans le graphe WordNet, nature de la relation (répétition et synonymie > antonymie et hyperonymie)

Exploitation de la cohésion lexicale en Traitement Automatique des Langues

- Evaluation de la qualité d'un texte généré automatiquement
- Résumé automatique : (Barzilay & Elhadad 1997) :
  - « lexical chains are a good indicator of the central topic of at text », « picking the concepts represented by strong lexical chains gives a better indication of the central topic of a text than simply picking the most frequent words in the text"
- Découpage automatique d'un texte en fonction de ruptures thématiques
  - Text tiling (Hearst 1992)
- Repérage des fautes et des lapsus comme éléments perturbants de la cohésion lexicale (Hirst & St-Onge 1998), (Hirst & Budanisky, 2005)
  - Des unités
    - qui ne sont reliées à aucune chaîne lexicale
    - qui n'entretiennent de relation sémantique avec aucune autre unité de son contexte

    He spent his summer travelling around the word. [world]

    We all hole that you will recover swiftly. [hope]

    American Express says . . . it doesn't know what the muss is all about. [fuss]
  - Idée :
    - Générer les variantes lexicales possibles du mot
    - Choisir celle qui entretient les liens les plus forts avec le plus de mots dans le contexte.

En résumé

- Repérer la cohésion lexicale dans les textes suppose :
  - De considérer la dimension contextuelle des relations sémantiques (en discours)
  - De prendre en compte des relations sémantiques variées => relations non classiques
  - De projeter l'information lexicale sous forme de réseaux lexicaux (plutôt que de chaînes)

Plan

1. La cohésion lexicale en linguistique et en TAL
2. Détection de la cohésion lexicale par voisinage distributionnel
   1. La notion de voisinage distributionnel
   2. Projection de voisins distributionnels en texte
   3. Bilan
3. Interactions entre cohésion lexicale et structure rhétorique du discours

Voisinage distributionnel

- Principe de l'analyse distributionnelle automatique :
  - Rapprochement de mots sur la base des contextes (syntaxiques) qu'ils partagent

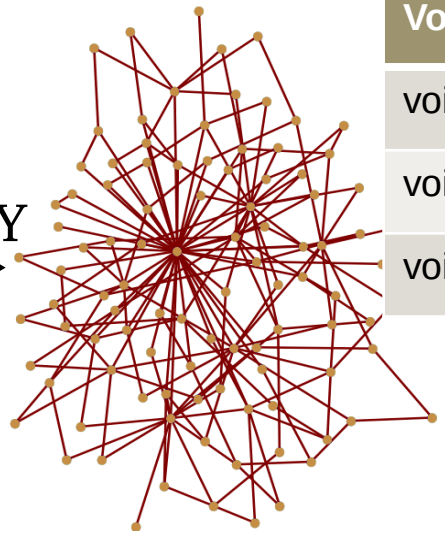
- Exemple : voiture / véhicule sont des *voisins distributionnels* car ils partagent une proportion importante de leurs contextes : percuter_suj, garer_obj, etc.

[…] lancée à 29 km/h pour que la **voiture percute** transversalement un p […]
[…] risque de suraccident est qu'un **véhicule percute** une personne (victi […]
[…] parfois le parking de l'école permet d'y **garer** une **voiture** la journée. […]
[…] procès-verbal plutôt que de **garer** leur **véhicule** loin de leur domicile […]

Construction des *Voisins de Wikipédia*

- Les *Voisins de Wikipédia*
  - Corpus :
    - Wikipédia francophone, version avril 2007
  - Chaîne de traitement :
    - SYNTEX (analyse syntaxique) – UPÉRY (analyse distributionnelle)
    - Score de similarité : Lin (1998)

SYNTEX-UPERY



Voisin *a*	Voisin *b*	Lin
voiture	véhicule	0.455
voiture	avion	0.401
voiture	moteur	0.283

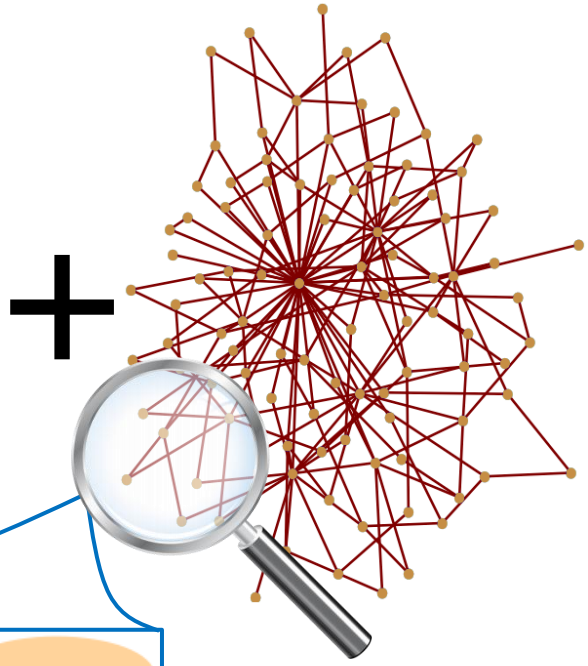
Wikipédia

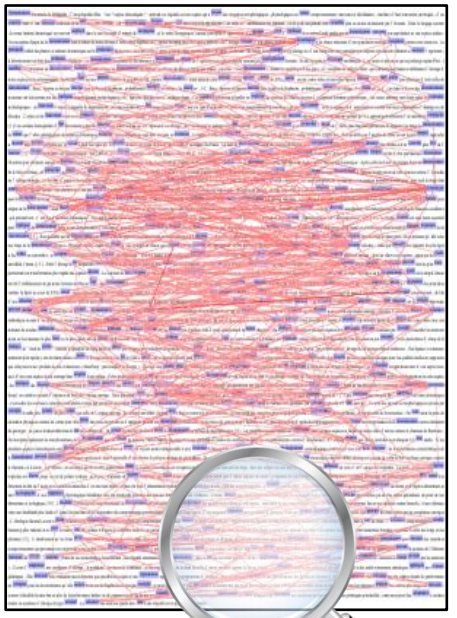
Voisins de Wikipédia

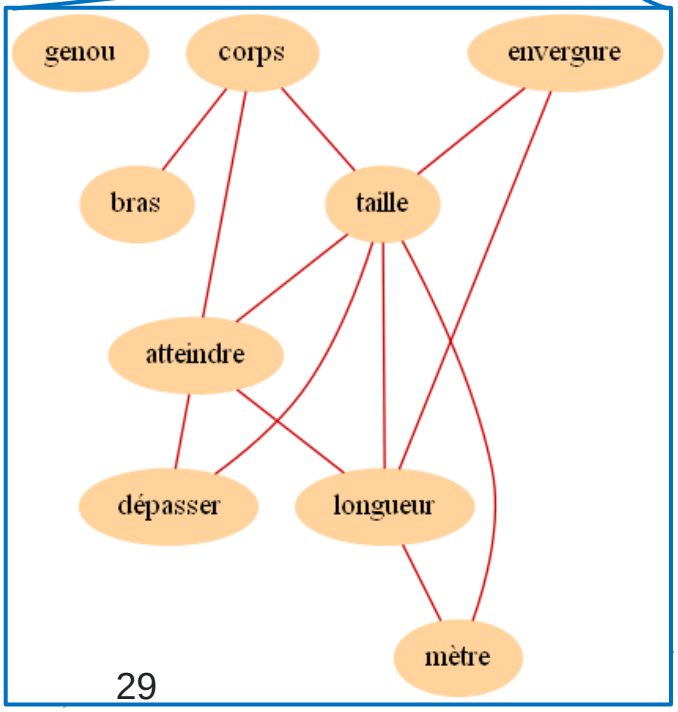
Construction des *Voisins de Wikipédia*

- L'analyse distributionnelle permet de capter une grande diversité de relations sémantiques…
  - Relations paradigmatiques classiques : synonymie, antonymie, hypéronymie, …
    - thèse_de / doctorat_de — université de Paris, médecine
    - perdre_obj / gagner_obj — pari, match
    - pomme / fruit — croquer_obj, cueillir_obj
  - Relations plus lâches : thématiques, associatives…
    - professeur_de / enseigner_obj — belles lettres, botanique
    - protester_contre / attitude_de — gouvernement, parti
  - Mais également des associations peu ou pas interprétables
    - bouteille / dé — jeter_obj, lancer_obj
    - bâtir_obj / orgue_de — cathédrale, église
    - compositeur / drapeau — 39 adj. de nationalité !

Projection des voisins en texte







Redressés , les gorilles atteignent une taille de 1,75 mètre , mais ils sont en fait un peu plus grands car ils ont les genoux fléchis . L' envergure des bras dépasse la longueur du corps et peut atteindre 2,75 mètres .

L' Albanie est un pays montagneux (70 %) , dont le point culminant s' élève à 2753 m (mont Korab) . Le reste est constitué de plaines alluviales , dont le terrain est plutôt de piètre qualité pour l' agriculture , alternativement inondé ou desséché . Les terres les plus fertiles sont situées dans le district des lacs (lac d' Ohrid , Grand Prespa et Petit Prespa) et sur certains plateaux intermédiaires entre la plaine et la montagne . La seule île notable est celle de Sazan qui fut tour à tour occupée par diverses grandes puissances européennes .

Le plus grand fleuve albanais est la Drini . Long de 282 km , elle est un des seuls à connaître un débit relativement stable tout au long de l' année . Les autres cours d' eau sont généralement presque secs durant l' été , même les rivières Semani et Vjosa qui ont pourtant une longueur de plus de 160 km .

Le climat y est méditerranéen dans les régions littorales (moyenne hivernale : 7°) , et devient plus continental dans le relief . Les précipitations sont assez élevées (1 000 à 1 500 mm annuels) , le flux d' air humide rencontrant la masse d' air continentale plus froide , surtout pendant l' hiver , qui est la saison pluvieuse .

Résumé des liens : agriculture_N/pays_N, air_N/climat_N, année_N/eau_N, climat_N/flux_N, climat_N/hiver_N, climat_N/saison_N, débit_N/longueur_N, district_N/lac_N, district_N/mont_N, district_N/plaine_N, district_N/plateau_N, eau_N/rivière_N, fleuve_N/pays_N, fleuve_N/rivière_N, flux_N/masse_N, flux_N/précipitation_N, hiver_N/relief_N, île_N/montagne_N, île_N/plateau_N, île_N/terrain_N, lac_N/mont_N, lac_N/montagne_N, lac_N/plaine_N, lac_N/plateau_N, lac_N/terrain_N, long_N/longueur_N, mont_N/montagne_N, mont_N/plaine_N, mont_N/plateau_N, mont_N/terrain_N, mont_N/terre_N, mont_N/tour_N, montagne_N/plaine_N, montagne_N/plateau_N, montagne_N/terrain_N, montagne_N/terre_N, montagne_N/tour_N, plaine_N/plateau_N, plaine_N/terrain_N, plateau_N/terrain_N, plateau_N/tour_N, région_N/rencontrer_V, reste_N/terrain_N, terrain_N/terre_N, terrain_N/tour_N, terre_N/tour_N

L' Albanie est un pays montagneux (70 %) , dont le point culminant s' élève à 2753 m (mont Korab) . Le reste est constitué de plaines alluviales , dont le terrain est plutôt de piètre qualité pour l' agriculture , alternativement inondé ou desséché . Les terres les plus fertiles sont situées dans le district des lacs (lac d' Ohrid , Grand Prespa et Petit Prespa) et sur certains plateaux intermédiaires entre la plaine et la montagne . La seule île notable est celle de Sazan qui fut tour à tour occupée par diverses grandes puissances européennes .

Le plus grand fleuve albanais est la Drini . Long de 282 km , elle est un des seuls à connaître un débit relativement stable tout au long de l' année . Les autres cours d' eau sont généralement presque secs durant l' été , même les rivières Semani et Vjosa qui ont pourtant une longueur de plus de 160 km .

Le climat y est méditerranéen dans les régions littorales (moyenne hivernale : 7°) , et devient plus continental dans le relief . Les précipitations sont assez élevées (1 000 à 1 500 mm annuels) , le flux d' air humide rencontrant la masse d' air continentale plus froide , surtout pendant l' hiver , qui est la saison pluvieuse .

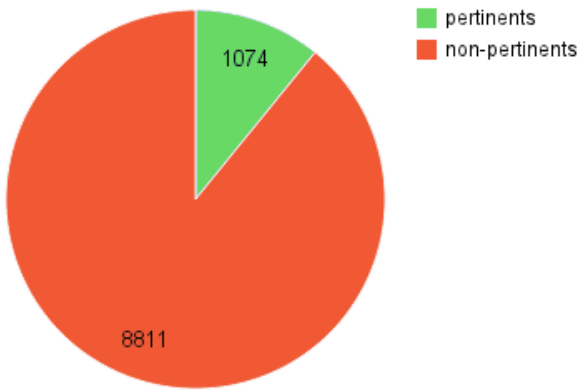
tumultueuse
trouble
chaude
courante

Résumé des liens : agriculture_N/pays_N, air_N/climat_N, année_N/eau_N, [illegible]nat_N/hiver_N, climat_N/saison_N, débit_N/longueur_N, district_N/lac_N, district_N/mont_N, distric[illegible]t_N/plateau_N, eau_N/rivière_N, fleuve_N/pays_N, fleuve_N/rivière_N, flux_N/masse_N, flux_N/pré[illegible]N/relief_N, île_N/montagne_N, île_N/plateau_N, île_N/terrain_N, lac_N/mont_N, lac_N/montagne_N, lac_N/plaine_N, lac_N/plateau_N, lac_N/terrain_N, long_N/longueur_N, mont_N/montagne_N, mont_N/plaine_N, mont_N/plateau_N, mont_N/terrain_N, mont_N/terre_N, mont[illegible]N/plaine_N, montagne_N/plateau_N, montagne_N/terrain_N, montagne_N/terr[illegible]N, plaine_N/plateau_N, plaine_N/terrain_N, plateau_N/terrain_N plateau_N/tour_N, r[illegible]este_N/terrain_N, terrain_N/terre_N, terrain_N/tour_N, terre_N/tour_N

surplomber_obj
culminer_obj
dominer_suj
abriter_suj

Détection de la cohésion lexicale par voisinage distributionnel

- Annotation de liens sélectionnés aléatoirement dans un sous-corpus de textes (sans seuil sur le score de Lin)
  - 10% des couples projetés présentent une relation sémantique pertinente selon les annotateurs

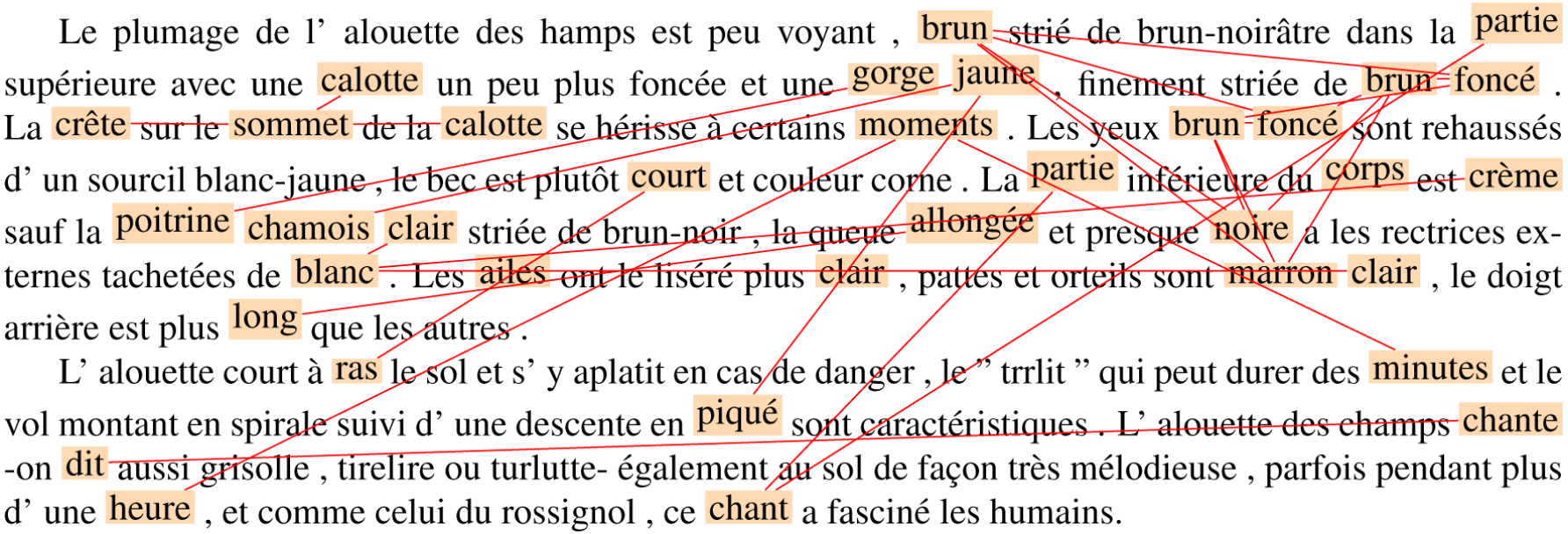


- En mettant un seuil à 0,25 sur le score de Lin (20% des voisins retenus) : 24% des couples projetés sont jugés pertinents

Détection de la cohésion lexicale par voisinage distributionnel

- Nécessité de mettre en place un filtrage en amont
  - intégrant, en plus des caractéristiques « distributionnelles » des paires de mots, la prise en compte de caractéristiques concernant le texte et la configuration des paires de voisins dans ce texte
  - très bons résultats, qui montrent l'impact de ces indices
- La détection automatique de la cohésion lexicale est un problème difficile !
  - Ce n'est pas seulement « la faute » de la ressource utilisée

Bruit avec la synonymie



Le plumage de l' alouette des hamps est peu voyant , brun strié de brun-noirâtre dans la partie supérieure avec une calotte un peu plus foncée et une gorge jaune , finement striée de brun foncé . La crête sur le sommet de la calotte se hérisse à certains moments . Les yeux brun foncé sont rehaussés d' un sourcil blanc-jaune , le bec est plutôt court et couleur corne . La partie inférieure du corps est crème sauf la poitrine chamois clair striée de brun-noir , la queue allongée et presque noire a les rectrices externes tachetées de blanc . Les ailes ont le liséré plus clair , pattes et orteils sont marron clair , le doigt arrière est plus long que les autres .

L' alouette court à ras le sol et s' y aplatit en cas de danger , le " trrlit " qui peut durer des minutes et le vol montant en spirale suivi d' une descente en piqué sont caractéristiques . L' alouette des champs chante -on dit aussi grisolle , tirelire ou turlutte- également au sol de façon très mélodieuse , parfois pendant plus d' une heure , et comme celui du rossignol , ce chant a fasciné les humains.

Résumé des liens : allongé/long, blanc/clair, blanc/crème, brun/foncé, brun/marron, brun/noir, calotte/sommet, chamois/jaune, chant/partie, chanter/dire, continuer/durer, continuer/suivre, court/ras, crête/sommet, gorge/poitrine, heure/moment, jaune/piqué, minute/moment, patte/pouvoir

Détection de la cohésion lexicale par voisinage distributionnel

- Nécessité de mettre en place un filtrage en amont
  - intégrant, en plus des caractéristiques « distributionnelles » des paires de mots, la prise en compte de caractéristiques concernant le texte et la configuration des paires de voisins dans ce texte
  - très bons résultats, qui montrent l'impact de ces indices
- La détection automatique de la cohésion lexicale est un problème difficile !
  - Ce n'est pas seulement « la faute » de la ressource utilisée
  - Il faut opter pour des approches plus fines, prenant en considération la caractérisation linguistique de la cohésion lexicale

Plan

1. La cohésion lexicale en linguistique et en TAL
2. Détection de la cohésion lexicale par analyse distributionnelle
3. Interactions entre cohésion lexicale et structure rhétorique du discours

Analyse du discours et cohésion lexicale

- La mise au jour de structures discursives est un objectif crucial
  - techniques de TAL visant à faciliter l'exploration de documents volumineux / l'accès automatisé à une information ciblée
    - Fouille de données, question-réponse, résumé automatique, etc.
- Ces structures discursives relèvent de plans variés
  - Constituants élémentaires reliés par des relations de discours
  - Segments de niveau supérieur dotés d'une fonction particulière
- Leur caractérisation / détection s'appuie sur des indices
  - Indices typographiques, marqueurs de discours (cue phrases), etc.
- Exploration du rôle de la cohésion lexicale

Au niveau de l'organisation globale des textes, la densité de la cohésion lexicale peut permettre de repérer des zones de continuité et de rupture

Description

La **carapace** sur le **dos** des tatous est formée de plaques osseuses articulées recouvertes de corne. Elles recouvrent la totalité du **dos** de l'animal, du **front** jusqu'à la **queue**, surface externe des **membres** comprise. Selon les espèces, elles forment soit des ceintures successives séparées par des replis cutanés souples, comme chez le tatou géant ; soit deux boucliers, l'un protégeant les **épaules**, l'autre les **hanches**, et séparés par des bandes d'**écailles** en nombre variable. Certaines espèces comme le tatou à trois bandes peuvent s'enrouler en boule en cas de danger

Comportement

En dehors des périodes de **reproduction**, le tatou mène une vie **solitaire**. Il cherche plutôt à éviter la présence de **congénères** venant parasiter leurs sources de nourriture. Bien qu'indépendants, les tatous ne font preuve d'aucune agressivité envers les autres **individus**. En général, les **mâles** défendent leur terrier, leurs sources de nourriture et leur **femelle** contre les autres **mâles**.

À un niveau local, certaines relations lexicales ciblées peuvent aider à l'identification d'une relation de discours

Un **véhicule** a effectué une spectaculaire **sortie** de **route**, hier vers 18h15, sur l'A36. La **voiture** circulait dans le sens Mulhouse-Montbéliard lorsqu'après être passée à hauteur du 35e RI, elle a **quitté** la **chaussée** sur sa droite.

Plan

1. La cohésion lexicale en linguistique et en TAL
2. Détection de la cohésion lexicale par analyse distributionnelle
3. Interactions entre cohésion lexicale et structure rhétorique du discours
    1. Approches rhétoriques du discours : notions
    2. Attachement des segments de discours, nature des relations de discours et cohésion lexicale
    3. Piste de recherche : marquage des relations de discours et cohésion lexicale
    4. Implémentation : intégration à un mini-parseur dédié aux relations d'élaboration et d'élaboration d'entité

Approches rhétoriques du discours

- Les approches rhétoriques du discours formalisent sa structure de manière **hiérarchique** et **ascendante** :
  - le discours est décomposé en unités minimales (typiquement la proposition)
  - … qui sont attachées *via* des relations de discours pour former des unités complexes
  - … et ainsi de suite jusqu'à obtenir une structuration de la globalité du texte

Approches rhétoriques du discours

- Exemples de relations :
  - But : le second segment présente de façon explicite le but, l'objectif pour lequel est réalisée l'action décrite dans le premier segment
    - [Les chercheurs ont fait grève]_1 [pour montrer leur mécontentement.]_2
    - goal(1,2)
  - Élaboration : le second segment fournit des détails supplémentaires sur l'éventualité décrite dans le premier segment
    - [Cette année-la vit de nombreux changements dans la vie de nos héros.]_1 [Jean epousa Adèle,]_2 [Marie s'acheta une maison à la campagne,]_3 [et Paul partit pour le Brésil.]_4
    - elaboration(1,[2-4])

- Les relations peuvent être explicitement signalées par un marqueur (ex. *pour*)

Structure rhétorique du discours

- Le projet ANNODIS dans son versant *ascendant* a visé la création d'un corpus annoté en relations de discours entre unités de discours élémentaires (UDE)
- Corpus :

	Nb de textes	Nb de mots
Wikipédia	42	17330
Autres (Est Républicain)	45	27098
Total	87	28146

- Relations de discours considérées :
  - Explication, But, Résultat, Parallèle, Contraste, Continuation, Conditionnel, Alternation, Attribution, Arrière-plan, Narration, Flashback, Encadrement, Temp (relation temporelle sous-spécifiée), Élaboration, Élaboration d'entité, Commentaire
- Différents niveaux d'annotation : annotations « naïve » et « experte »

Exemple

- [Une pluie d'étoiles]_1 [Non !]_2 [Il ne s'agit pas d'un phénomène météorologique accompagnant le solstice d'été.]_3 [Plus simplement,]_4 [les hasards du calendrier du Comité départemental d'action touristique a fait coïncider la promotion de l'Office de tourisme avec le nouveau classement de l'hôtel-restaurant Le Relais à Arc-et-Senans.]_5

  elaboration(1/[2-5])

  elaboration (2/3)

  elaboration(4/5)

  contrast ([2,3]/[4,5])

Exemple

- [Une pluie d'étoiles]_1 **[**Non !**]_2 [**Il ne s'agit pas d'un phénomène météorologique accompagnant le solstice d'été.**]_3** [Plus simplement,]_4 [les hasards du calendrier du Comité départemental d'action touristique a fait coïncider la promotion de l'Office de tourisme avec le nouveau classement de l'hôtel-restaurant Le Relais à Arc-et-Senans.]_5

  elaboration(1/[2-5])

  elaboration (2/3)

  elaboration(4/5)

  contrast ([2,3]/[4,5])

Exemple

- [Une pluie d'étoiles]_1 [Non !]_2 [Il ne s'agit pas d'un phénomène météorologique accompagnant le solstice d'été.]_3 **[**Plus simplement,**]_4 [**les hasards du calendrier du Comité départemental d'action touristique a fait coïncider la promotion de l'Office de tourisme avec le nouveau classement de l'hôtel-restaurant Le Relais à Arc-et-Senans.**]_5**

  elaboration(1/[2-5])

  elaboration (2/3)

  elaboration(4/5)

  contrast ([2,3]/[4,5])

Exemple

- [Une pluie d'étoiles]_1 **[**Non !]_2 [Il ne s'agit pas d'un phénomène météorologique accompagnant le solstice d'été.**]_3 [**Plus simplement,]_4 [les hasards du calendrier du Comité départemental d'action touristique a fait coïncider la promotion de l'Office de tourisme avec le nouveau classement de l'hôtel-restaurant Le Relais à Arc-et-Senans.**]_5**

  elaboration(1/[2-5])

  elaboration (2/3)

  elaboration(4/5)

  contrast ([2,3]/[4,5])

Exemple

- **[**Une pluie d'étoiles**]_1 [**Non !]_2 [Il ne s'agit pas d'un phénomène météorologique accompagnant le solstice d'été.]_3 [Plus simplement,]_4 [les hasards du calendrier du Comité départemental d'action touristique a fait coïncider la promotion de l'Office de tourisme avec le nouveau classement de l'hôtel-restaurant Le Relais à Arc-et-Senans.**]_5**

  elaboration(1/[2-5])

  elaboration (2/3)

  elaboration(4/5)

  contrast ([2,3]/[4,5])

Exemple 2

[Denise Jacquin s'est éteinte mercredi, à l'âge de 73 ans, à l'hôpital,]_2 [des suites d'une pénible maladie contre laquelle elle a lutté courageusement.]_3 [Denise Fachinetti est née à Mancenans au sein d'une famille de huit enfants.]_4 [Mais elle a passé toute son enfance à Geney,]_5 [où elle a gardé les vaches.]_6

explanation(2/3)

flashback(2/[4-6])

contrast(4/[5-6])

elaboration(5/6)

Exemple 2

[Denise Jacquin s'est éteinte mercredi, à l'âge de 73 ans, à l'hôpital,]_2 [des suites d'une pénible maladie contre laquelle elle a lutté courageusement.]_3 [Denise Fachinetti est née à Mancenans au sein d'une famille de huit enfants.]_4 **[**Mais elle a passé toute son enfance à Geney,**]_5 [**où elle a gardé les vaches.**]_6**

explanation(2/3)

flashback(2/[4-6])

contrast(4/[5-6])

elaboration(5/6)

Exemple 2

[Denise Jacquin s'est éteinte mercredi, à l'âge de 73 ans, à l'hôpital,]_2 [des suites d'une pénible maladie contre laquelle elle a lutté courageusement.]_3 **[**Denise Fachinetti est née à Mancenans au sein d'une famille de huit enfants.**]_4 [**Mais elle a passé toute son enfance à Geney,]_5 [où elle a gardé les vaches.**]_6**

explanation(2/3)

flashback(2/[4-6])

contrast(4/[5-6])

elaboration(5/6)

Exemple 2

[Denise Jacquin s'est éteinte mercredi, à l'âge de 73 ans, à l'hôpital,**]_2** [des suites d'une pénible maladie contre laquelle elle a lutté courageusement.]_3
[Denise Fachinetti est née à Mancenans au sein d'une famille de huit enfants.]_4 [Mais elle a passé toute son enfance à Geney,]_5 [où elle a gardé les vaches.**]_6**

explanation(2/3)
flashback(2/[4-6])
contrast(4/[5-6])
elaboration(5/6)

Exemple 2

[Denise Jacquin s'est éteinte mercredi, à l'âge de 73 ans, à l'hôpital,**]_2 [**des suites d'une pénible maladie contre laquelle elle a lutté courageusement.**]_3** [Denise Fachinetti est née à Mancenans au sein d'une famille de huit enfants.]_4 [Mais elle a passé toute son enfance à Geney,]_5 [où elle a gardé les vaches.]_6

explanation(2/3)

flashback(2/[4-6])

contrast(4/[5-6])

elaboration(5/6)

Structure rhétorique du discours

- Étude du rôle de la cohésion lexicale dans différents aspects liés à cette approche du discours :
  - Cohésion lexicale et attachement des UDE
    - Hypothèse : la cohésion lexicale est un indicateur de la cohérence du discours et suit la structure rhétorique
  - Cohésion lexicale et nature des relations de discours
    - Hypothèse : la présence de cohésion lexicale dépend du type de relation rhétorique intervenant entre segments du discours

Cohésion lexicale / Attachement des segments

[Denise Jacquin s'est éteinte mercredi, à l'âge de 73 ans, à l'hôpital,**]_2 [**des suites d'une pénible maladie contre laquelle elle a lutté courageusement.**]_3**
[Denise Fachinetti est née à Mancenans au sein d'une famille de huit enfants.]_4 [Mais elle a passé toute son enfance à Geney,]_5 [où elle a gardé les vaches.]_6

explanation(2/3)

flashback(2/[4-6])

contrast(4/[5-6])

elaboration(5/6)

Cohésion lexicale / Attachement des segments

[Denise Jacquin s'est éteinte mercredi, à l'âge de 73 ans, à l'hôpital,]_2 **[**des suites d'une pénible maladie contre laquelle elle a lutté courageusement.**]_3** **[**Denise Fachinetti est née à Mancenans au sein d'une famille de huit enfants.**]_4** [Mais elle a passé toute son enfance à Geney,]_5 [où elle a gardé les vaches.]_6

explanation(2/**3**)

flashback(2/[**4**-6])

contrast(**4**/[5-6])

elaboration(5/6)

Structure rhétorique du discours

- Étude du rôle de la cohésion lexicale dans différents aspects liés à cette approche du discours :
  - Cohésion lexicale et attachement des UDE
    - Hypothèse : la cohésion lexicale est un indicateur de la cohérence du discours et suit la structure rhétorique
  - Cohésion lexicale et nature des relations de discours
    - Hypothèse : la présence de cohésion lexicale dépend du type de relation rhétorique intervenant entre segments du discours

Cohésion lexicale / Attachement des segments

[Denise Jacquin s'est éteinte mercredi, à l'âge de 73 ans, à l'hôpital,**]_2 [**des suites d'une pénible maladie contre laquelle elle a lutté courageusement.**]_3**
[Denise Fachinetti est née à Mancenans au sein d'une famille de huit enfants.]_4 [Mais elle a passé toute son enfance à Geney,]_5 [où elle a gardé les vaches.]_6

explanation(2/3)

flashback(2/[4-6])

contrast(4/[5-6])

elaboration(5/6)

Exemple 2

[Denise Jacquin s'est éteinte mercredi, à l'âge de 73 ans, à l'hôpital,**]_2** [des suites d'une pénible maladie contre laquelle elle a lutté courageusement.]_3
[Denise Fachinetti est née à Mancenans au sein d'une famille de huit enfants.]_4 [Mais elle a passé toute son enfance à Geney,]_5 [où elle a gardé les vaches.**]_6**

explanation(2/3)
flashback(2/[4-6])
contrast(4/[5-6])
elaboration(5/6)

Méthodologie

- Comparaison de la cohésion lexicale entre :
  - UDE reliées par l'annotation
  - UDE appartenant au même texte mais n'étant pas directement reliées par l'annotation
  - UDE appartenant à des textes différents du corpus

- Pour les UDE reliées, comparaison selon la nature des relations rhétoriques les connectant

→ Comment évaluer la « force » de la cohésion lexicale entre deux segments ?

Mesurer la force de la cohésion lexicale entre deux segments

- Comment passer de la similarité entre mots à la proximité entre segments textuels ?

- Première approche : compter le nombre de liens entre les deux segments (normalisation par la taille des segments)

L' Albanie est un pays montagneux (70 %) , dont le point culminant s' élève à 2753 m (mont Korab) . Le reste est constitué de plaines alluviales , dont le terrain est plutôt de piètre qualité pour l' agriculture , alternativement inondé ou desséché . Les terres les plus fertiles sont situées dans le district des lacs (lac d' Ohrid , Grand Prespa et Petit Prespa) et sur certains plateaux intermédiaires entre la plaine et la montagne . La seule île notable est celle de Sazan qui fut tour à tour occupée par diverses grandes puissances européennes .

Le plus grand fleuve albanais est la Drini . Long de 282 km , elle est un des seuls à connaître un débit relativement stable tout au long de l' année . Les autres cours d' eau sont généralement presque secs durant l' été , même les rivières Semani et Vjosa qui ont pourtant une longueur de plus de 160 km .

Le climat y est méditerranéen dans les régions littorales (moyenne hivernale : 7°) , et devient plus continental dans le relief . Les précipitations sont assez élevées (1 000 à 1 500 mm annuels) , le flux d' air humide rencontrant la masse d' air continentale plus froide , surtout pendant l' hiver , qui est la saison pluvieuse .

Résumé des liens : agriculture_N/pays_N, air_N/climat_N, année_N/eau_N, climat_N/flux_N, climat_N/hiver_N, climat_N/saison_N, débit_N/longueur_N, district_N/lac_N, district_N/mont_N, district_N/plaine_N, district_N/plateau_N, eau_N/rivière_N, fleuve_N/pays_N, fleuve_N/rivière_N, flux_N/masse_N, flux_N/précipitation_N, hiver_N/relief_N, île_N/montagne_N, île_N/plateau_N, île_N/terrain_N, lac_N/mont_N, lac_N/montagne_N, lac_N/plaine_N, lac_N/plateau_N, lac_N/terrain_N, long_N/longueur_N, mont_N/montagne_N, mont_N/plaine_N, mont_N/plateau_N, mont_N/terrain_N, mont_N/terre_N, mont_N/tour_N, montagne_N/plaine_N, montagne_N/plateau_N, montagne_N/terrain_N, montagne_N/terre_N, montagne_N/tour_N, plaine_N/plateau_N, plaine_N/terrain_N, plateau_N/terrain_N, plateau_N/tour_N, région_N/rencontrer_V, reste_N/terrain_N, terrain_N/terre_N, terrain_N/tour_N, terre_N/tour_N

Mesurer la force de la cohésion lexicale entre deux segments

- Avec des très petits segments : trop de variabilité

[les touches noires étaient recouvertes d'ébène] [et les touches blanches d'ivoire]

- 2 liens : répétition touche et noir / blanc

[Par la suite, il s'inspire considérablement des règles établies par l'art égyptien,] [notamment en peinture et en sculpture.]

- 5 liens : art / peinture, art / sculpture, peinture / inspirer, peinture / règle, peinture / établir

Mesurer la force de la cohésion lexicale entre deux segments

- Prise en compte des valeurs de similarité lexicale (score de Lin)

[les touches noires étaient recouvertes d'ébène] [et les touches blanches d'ivoire]

- répétition touche 1 ??
- noir/blanc 0,36

Moyenne : 0,68

[Par la suite, il s'inspire considérablement des règles établies par l'art égyptien,] [notamment en peinture et en sculpture.]

- art/peinture 0,28
- art/sculpture 0,18
- peinture/inspirer 0,15
- peinture/règle 0,17
- peinture/établir 0,10

Moyenne : 0,17

Mesurer la force de la cohésion lexicale entre deux segments

- Prise en compte des valeurs de similarité lexicale (score de Lin)

[les touches noires étaient recouvertes d'ébène] [et les touches blanches d'ivoire]

- répétition touche 1 ??
- noir/blanc 0,36

Similarité maximale : 1 ?

[Par la suite, il s'inspire considérablement des règles établies par l'art égyptien,] [notamment en peinture et en sculpture.]

- art/peinture 0,28
- art/sculpture 0,18
- peinture/inspirer 0,15
- peinture/règle 0,17
- peinture/établir 0,10

Similarité maximale : 0,28

Mesurer la force de la cohésion lexicale entre deux segments

- **Score de Mihalcea *et al.* (2006)**

[les touches noires étaient recouvertes d'ébène] [et les touches blanches d'ivoire]

- répétition touche 1 ??
- noir/blanc 0,36

[Par la suite, il s'inspire considérablement des règles établies par l'art égyptien,] [notamment en peinture et en sculpture.]

- art/peinture 0,28
- art/sculpture 0,18
- peinture/inspirer 0,15
- peinture/règle 0,17
- peinture/établir 0,10

Mesurer la force de la cohésion lexicale entre deux segments

- **Score de Mihalcea *et al.* (2006)**

[les touches noires étaient recouvertes d'ébène] [et les touches blanches d'ivoire]

- répétition **touche 1 ??**
- noir/blanc 0,36

(1

[Par la suite, il s'inspire considérablement des règles établies par l'art égyptien,] [notamment en peinture et en sculpture.]

- art/peinture 0,28
- art/sculpture 0,18
- peinture/inspirer 0,15
- peinture/règle 0,17
- peinture/établir 0,10

Mesurer la force de la cohésion lexicale entre deux segments

- **Score de Mihalcea *et al.* (2006)**

[les touches noires étaient recouvertes d'ébène] [et les touches blanches d'ivoire]

- répétition touche 1 ??
- noir/**blanc** **0,36**

(1+0,36)

[Par la suite, il s'inspire considérablement des règles établies par l'art égyptien,] [notamment en peinture et en sculpture.]

- art/peinture 0,28
- art/sculpture 0,18
- peinture/inspirer 0,15
- peinture/règle 0,17
- peinture/établir 0,10

Mesurer la force de la cohésion lexicale entre deux segments

- **Score de Mihalcea *et al.* (2006)**

[les touches noires étaient recouvertes d'ébène] [et les touches blanches d'ivoire]

- répétition touche 1 ??
- noir/blanc 0,36

(1+0,36)/2=**0,68**

[Par la suite, il s'inspire considérablement des règles établies par l'art égyptien,] [notamment en peinture et en sculpture.]

- art/peinture 0,28
- art/sculpture 0,18
- peinture/inspirer 0,15
- peinture/règle 0,17
- peinture/établir 0,10

Mesurer la force de la cohésion lexicale entre deux segments

- **Score de Mihalcea *et al.* (2006)**

[les touches noires étaient recouvertes d'ébène] [et les touches blanches d'ivoire]

- répétition touche 1 ??
- noir/blanc 0,36

[Par la suite, il s'inspire considérablement des règles établies par l'art égyptien,] [notamment en peinture et en sculpture.]

- art/**peinture** **0,28**
- art/sculpture 0,18
- **peinture**/inspirer 0,15
- **peinture**/règle 0,17
- **peinture**/établir 0,10

(0,28

Mesurer la force de la cohésion lexicale entre deux segments

- **Score de Mihalcea *et al.* (2006)**

[les touches noires étaient recouvertes d'ébène] [et les touches blanches d'ivoire]

- répétition touche 1 ??
- noir/blanc 0,36

[Par la suite, il s'inspire considérablement des règles établies par l'art égyptien,] [notamment en peinture et en sculpture.]

- art/peinture 0,28
- art/**sculpture** **0,18**
- peinture/inspirer 0,15
- peinture/règle 0,17
- peinture/établir 0,10

(0,28+0,18)

Mesurer la force de la cohésion lexicale entre deux segments

- **Score de Mihalcea *et al.* (2006)**

[les touches noires étaient recouvertes d'ébène] [et les touches blanches d'ivoire]

- répétition touche 1 ??
- noir/blanc 0,36

[Par la suite, il s'inspire considérablement des règles établies par l'art égyptien,] [notamment en peinture et en sculpture.]

- art/peinture 0,28
- art/sculpture 0,18
- peinture/inspirer 0,15
- peinture/règle 0,17
- peinture/établir 0,10

(0,28+0,18)/2=**0,23**

Méthodologie

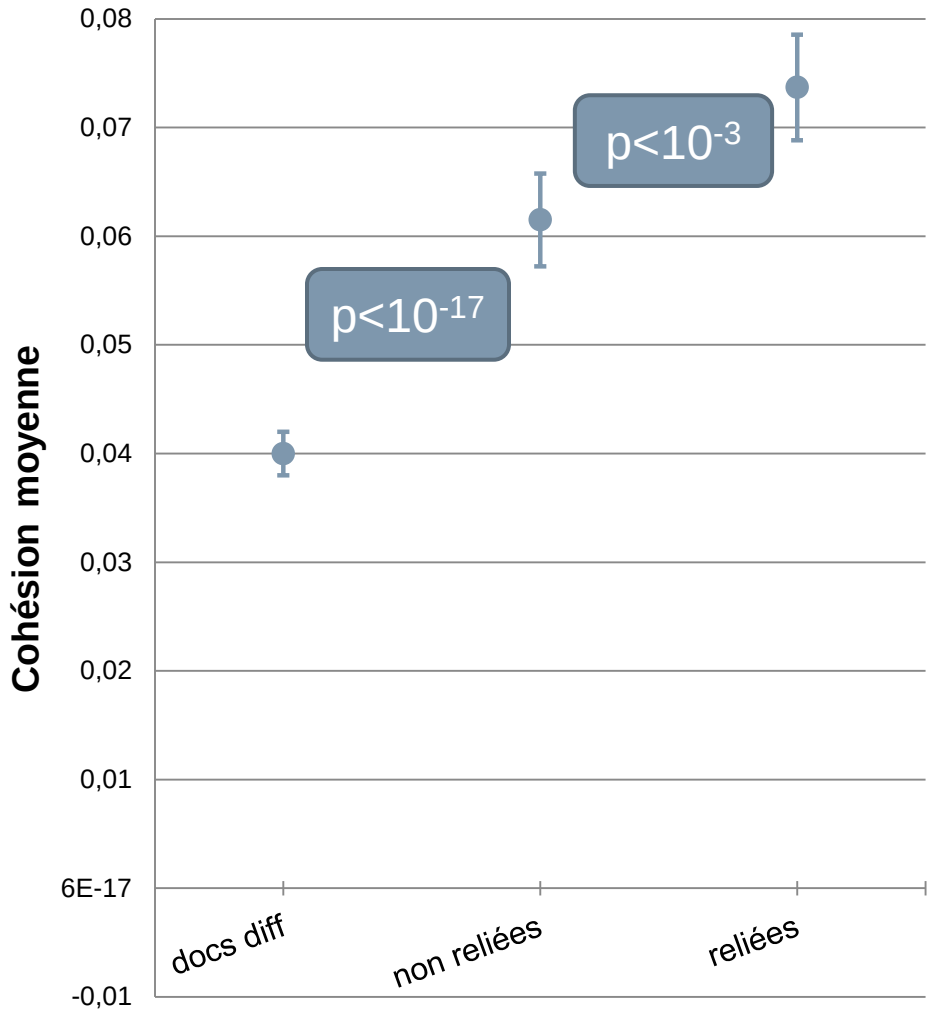
- **Comparaison de la cohésion lexicale entre :**
  - **UDE reliées par l'annotation**
  - **UDE appartenant au même texte mais n'étant pas directement reliées par l'annotation**
  - **UDE appartenant à des textes différents du corpus**

- Pour les UDE reliées, comparaison selon la nature des relations rhétoriques les connectant

Attachement des UDE

Trois groupes de paires d'UDE :
• 2000 paires d'UDE provenant de textes différents dans le corpus
• 2000 paires d'UDE provenant d'un même document mais non directement reliées
•1927 paires d'UDE reliées par l'annotation

Calcul de similarité entre UDE :
Lin + Mihalcea *et al.* (2006)



Attachement des UDE

- Autres aspects testés
  - Différence significative (p<0,001) entre UDE **adjacentes** reliées ou non :
    - UDE reliées : 96 ±12
    - UDE non reliées : 63 ±12
  - Différence importante entre :
    - UDE dans la même phrase 59 ±5
      - vs
    - UDE dans deux phrases différentes 109 ±12

    → les relations intra-phrastiques ont besoin de moins de cohésion lexicale ?

    → certaines relations de discours qui apparaissent plutôt à l'intérieur des phrases impliquent moins de cohésion lexicale ?

    → ex. Attribution (lie un acte de parole à l'agent de cette acte)

    [La direction générale de Citroën a informé ses employés]_1 [que les nouveaux contrats de travail prendront effet lundi prochain]_2

Méthodologie

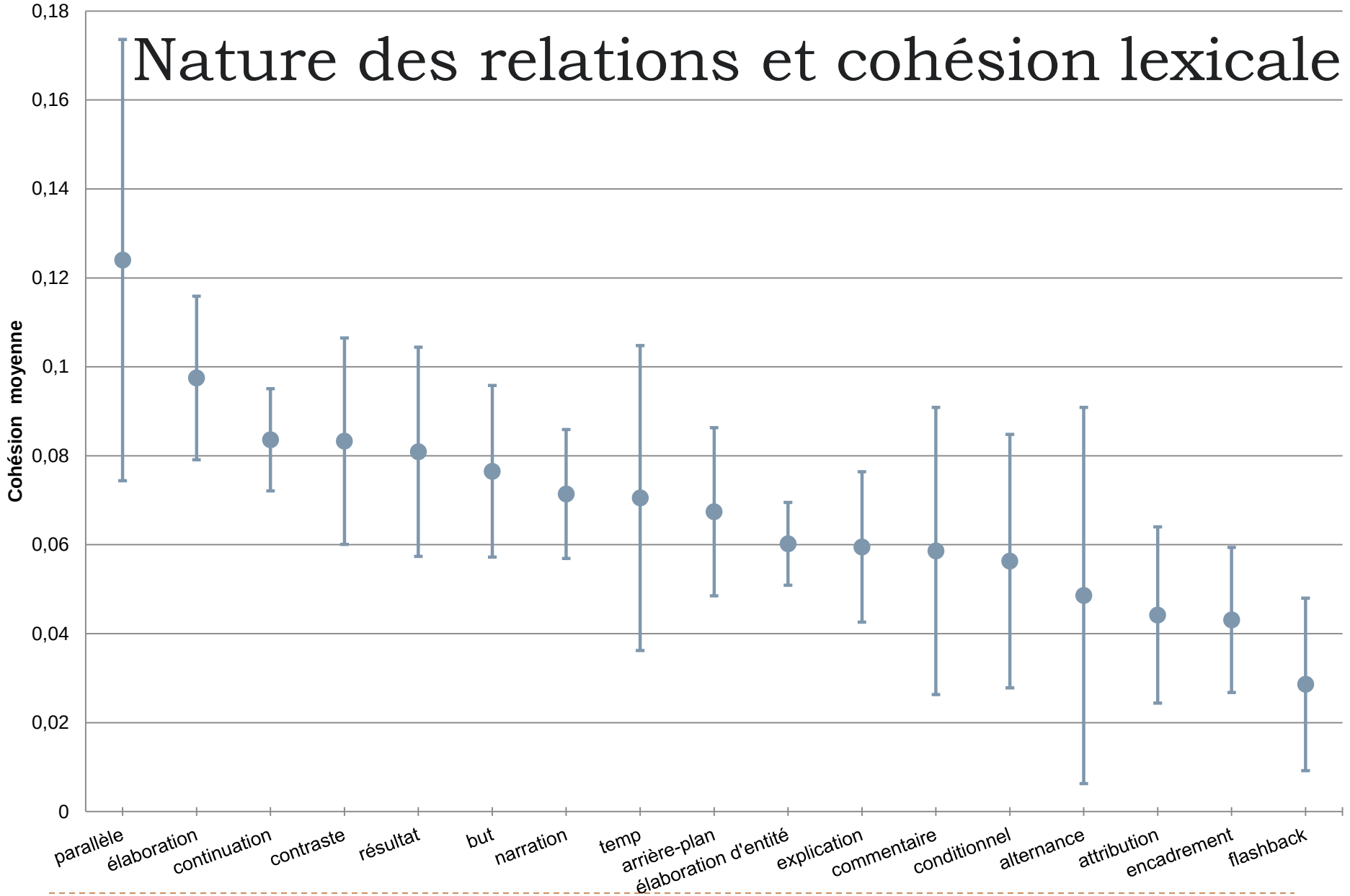
- Comparaison de la cohésion lexicale entre :
  - UDE reliées par l'annotation
  - UDE appartenant au même texte mais n'étant pas directement reliées par l'annotation
  - UDE appartenant à des textes différents du corpus

- **Pour les UDE reliées, comparaison selon la nature des relations rhétoriques les connectant**

Nature des relations et cohésion lexicale

Relation	total (nb)	(%)
alternation	18	0.5
attribution	75	2.2
background	155	4.6
comment	78	2.3
continuation	681	20.3
contrast	144	4.3
e-elab	527	15.7
elaboration	625	18.6
explanation	130	3.9
flashback	27	0.8
frame	211	6.3
goal	95	2.8
narration	349	10.4
parralel	59	1.8
result	163	4.9
temploc	18	0.5

Nature des relations et cohésion lexicale



Plan

1. La cohésion lexicale en linguistique et en TAL
2. Détection de la cohésion lexicale par analyse distributionnelle
3. Interactions entre cohésion lexicale et structure rhétorique du discours
    1. Approches rhétoriques du discours : notions
    2. Attachement des segments de discours, nature des relations de discours et cohésion lexicale
    3. **Piste de recherche : marquage des relations de discours et cohésion lexicale**
    4. Implémentation : intégration à un mini-parseur dédié aux relations d'élaboration et d'élaboration d'entité

Marquage des relations et cohésion lexicale

- Hypothèse : indices lexicaux davantage présents quand la relation n'est pas marquée explicitement ?

  [Ce **réchauffement** se serait déroulé en deux phases,]a
  [ces deux phases sont séparées par une période de léger **refroidissement.**]b

  contraste(a,b)

- L'interprétation de la relation de contraste repose non pas sur un marqueur (*mais*, *alors_que*, …) mais sur la relation d'antonymie entre « réchauffement » et « refroidissement » ?

Marquage des relations et cohésion lexicale

- Première vérification : projection d'un lexique de marqueurs, LexConn (Roze *et al.* 2012)
  - aussitôt que flashback,explanation
  - […]
  - en revanche contrast
  - […]
  - pour goal
  - […]
- Problème : désambiguïsation entre usages discursifs ou non

Marquage des relations et cohésion lexicale

- Pour chaque paire d'UDE
  - contient-elle un marqueur recensé par le lexique ? (exclusion des marqueurs souvent utilisés dans un sens non discursif : *ou*, *pour*…)
  - si oui, est-ce un marqueur de la relation identifiée par l'annotation ?
- seulement 55 paires d'UDE sur 1927 contiennent un marqueur qui indique potentiellement la relation qui les connecte
- pas de différence significative entre la cohésion lexicale des relations explicitement marquées ou non

Marquage des relations et cohésion lexicale

- Approche plus fine, relation par relation ?
  - Examen systématique de toutes les occurrences de certaines relations (LexConn projeté)
  - Confirmation : la relation est-elle marquée ou non ?

le roi Shoulgi construit un mur entre le Tigre et l'Euphrate,
mais celui-ci ne suffira pas à arrêter les groupes amorrites,
markers in 1:
markers in 2: mais (contrast)

Les paroles de Black Sabbath sont surtout influencées par des films d'horreur et une bonne dose de superstition.
Black Widow **au contraire**, aime citer dans ses chansons des rituels, des entités démoniaques
markers in 1:
markers in 2:

A son point **culminant**,
L'engouement et l'utilisation de BITNET **déclinèrent** avec l'avènement de TCP/IP et d'Internet
markers in 1:
markers in 2:

Bien sûr, il reste difficile de faire complétement la part de la légende et de la réalité historique.
Toutefois, ce massacre participa grandement à la réputation de cruauté sauvage et aveugle
markers in 1:
markers in 2:

l'État abrite des collecteurs cylindro-paraboliques
la plus grande centrale à tour comme Solar one puis Solar 2 ne dépasse pas 10 MW.
markers in 1:
markers in 2:

Plan

1. La cohésion lexicale en linguistique et en TAL
2. Détection de la cohésion lexicale par analyse distributionnelle
3. Interactions entre cohésion lexicale et structure rhétorique du discours
    1. Approches rhétoriques du discours : notions
    2. Attachement des segments de discours, nature des relations de discours et cohésion lexicale
    3. Piste de recherche : marquage des relations de discours et cohésion lexicale
    4. **Implémentation : intégration à un mini-parseur dédié aux relations d'élaboration et d'élaboration d'entité**

Élaboration et E-élaboration

- La relation d'**élaboration** relie deux propositions si la seconde proposition décrit un sous-état ou sous-événement de l'état ou événement décrit dans la première proposition
- La relation d'**élaboration d'entité** (e-élaboration) relie deux segments dont le second précise une propriété d'une des entités impliquées dans le premier segment

Élaboration et E-élaboration

[La Lausitz, [une région pauvre de l'est de l'Allemagne,]1 [réputée pour ses mines de charbon à ciel ouvert,]2 [a été le théâtre d'une première mondiale, mardi 9 septembre.]3 [Le groupe suédois Vattenfall a inauguré, dans la petite ville de Spremberg, une centrale électrique à charbon expérimentale]4 [qui met en œuvre toute la chaîne des techniques de captage et de stockage du carbone.]5

Continuation (1,2)

E-Elaboration (3,[1-2])

Elaboration (3,4)

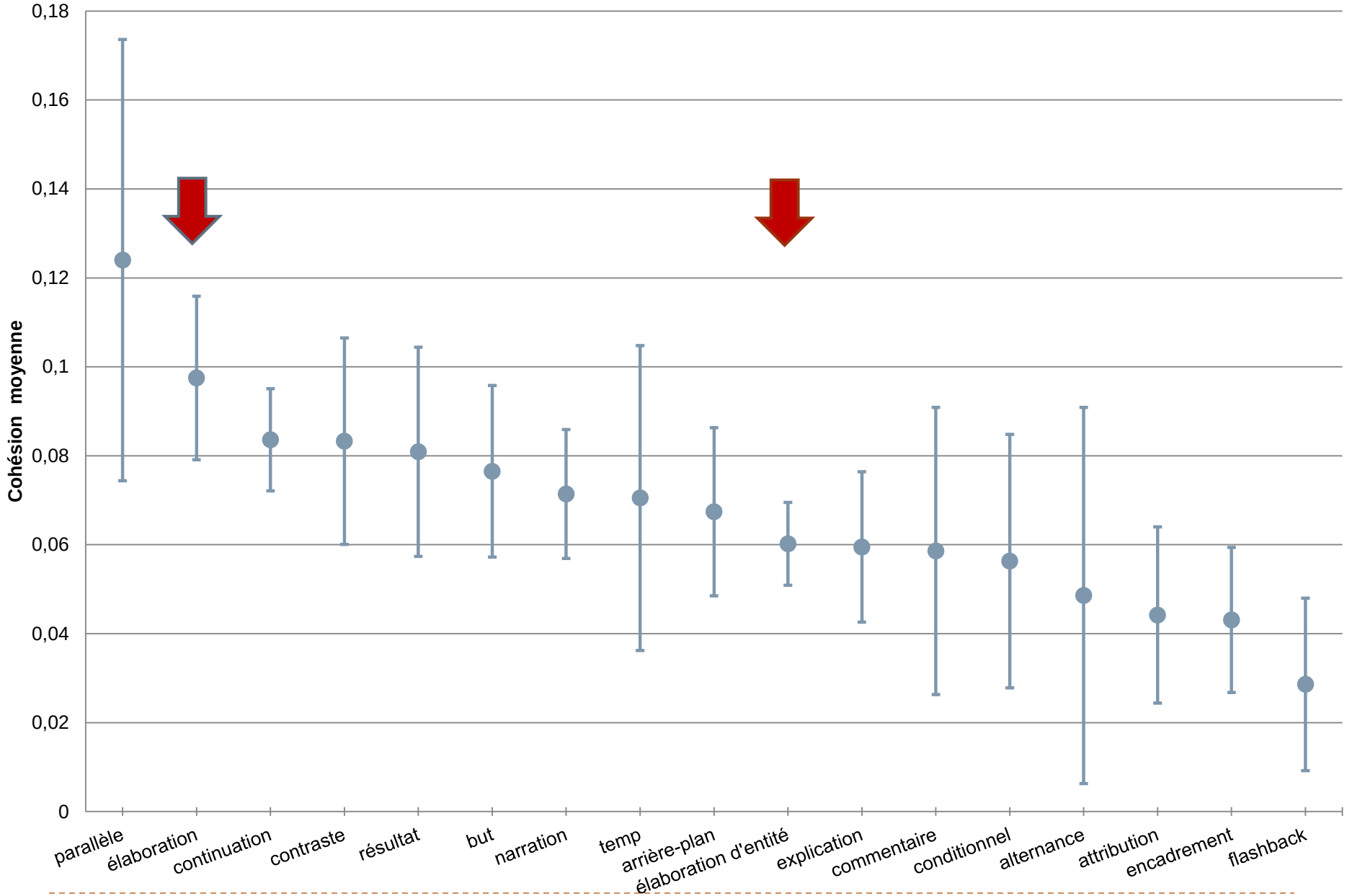
E-Elaboration (4,5)

Élaboration et E-élaboration

- Deux relations
  - très fréquentes
    - élaborations + e-élaborations : 50% de l'annotation naïve
      35% de l'annotation experte
  - très fréquemment confondues par les annotateurs naïfs…

		naïfs	
		elab	e-elab
ref	elab	302	70
	e-elab	158	216

  - … mais très différentes du point de vue de leur cohésion lexicale


0,18
0,16
0,14
0,12
0,1
0,08
0,06
0,04
0,02
0
Cohésion moyenne
parallèle
élaboration
continuation
contraste
résultat
but
narration
temp
arrière-plan
élaboration d'entité
explication
commentaire
conditionnel
alternance
attribution
encadrement
flashback

Élaboration et E-élaboration

- **[**Un soir, il faisait un temps horrible,**]16 [**les éclairs se croisaient,**]17 [**le tonnerre grondait,**]18 [**la pluie tombait à torrents.**]19**

  Elaboration(16,[17-19])

- **[**Pourquoi a-t-on abattu Paul Mariani, **[**cinquante-cinq ans,**]4 [**attaché au cabinet de M. François Doubin**]5** ?**]6**

  E-elaboration(6,[4-5])

Élaboration et E-élaboration

- Marqueurs indiqués dans le manuel d'annotation ANNODIS : *à savoir*, *c'est-à-dire*, *notamment*
  - peuvent marquer les deux relations
- Autres caractéristiques :
  - l'e-élaboration est souvent réalisée par des propositions relatives, des appositions, parenthèses... (Prévot 2009)
  - Le gérondif peut marquer différentes relations de discours dont l'élaboration (Vergez-Couret 2010)

Élaboration et E-élaboration

- Pour chaque paire de segments (Sa,Sb), implémentation des traits suivants :

Indices	Description	Valeurs
Sc	Score de cohésion lexicale	$Sc \in \mathbb{R}^+$
rel	S_b est une proposition relative	booléen
app	S_b est une apposition nominale/adjectivale	booléen
ger	S_b est un syntagme gérondif	booléen
par	S_b apparaît entre parenthèses	booléen
emb	S_b est enchâssé dans S_a	booléen
w_{Sa}	nombre de mots de S_a	$w_{S1} \in \mathbb{N}^*$
w_{Sb}	nombre de mots de S_b	$w_{S2} \in \mathbb{N}^*$
w_{tot}	$w_{Sa} + w_{Sb}$	$w_{tot} \in \mathbb{N}$
s_{Sa}	nombre de segments de S_a	$s_{S1} \in \mathbb{N}^*$
s_{Sb}	nombre de segments de S_b	$s_{S2} \in \mathbb{N}^*$
s_{tot}	$s_{Sa} + s_{Sb}$	$s_{tot} \in \mathbb{N}$

Élaboration et E-élaboration

Entrainement d'un classifieur (Weka+*RandomForest)* intégrant les scores de cohésion lexicale calculés

Les résultats (validation croisée) montrent une nette amélioration de l'annotation naïve

		naïfs	
		elab	e-elab
réf	elab	302	70
	e-elab	158	216

Exactitude : 69.4%

		naïfs + auto	
		non pert	pert.
réf	non pert.	306	66
	pert.	115	259

Exactitude : 75.7%

Mini-implémentation : Expé élab et e-élab

- Examen de l'impact des différentes catégories d'indices

Indices utilisés	Exactitude
Annotation naïve	69.4%
Score de cohésion lexicale	72.3% (+2.9%)
Indices syntaxiques	71.7% (+2.3%)
Indices structurels	69.7% (+0.3%)
All	75.7% (+6.3%)

- Rôle important de la cohésion lexicale

Conclusion

- Sur la cohésion lexicale et sa détection :
  - Réduire le fossé entre linguistique et TAL ?
    - linguistique : produire des annotations exploitables ?
    - TAL : vers des approches plus fines prenant en considération la caractérisation linguistique de la cohésion lexicale ?

Conclusion

- Sur le rôle de la cohésion lexicale dans la structuration rhétorique du discours
  - La cohésion lexicale est un indicateur de la cohérence du discours et suit la structure rhétorique
  - La présence de cohésion lexicale dépend du type de relation rhétorique intervenant entre deux segments de discours
  - Quelles interactions entre les différents modes de signalisation (marqueurs de discours / cohésion lexicale) ?